A Mathematical Introduction to Signals and Systems

Volume V. System Theory

Andrew D. Lewis

This version: 2022/03/07

Preface for series

The subject of signals and systems, particularly linear systems, is by now an entrenched part of the curriculum in many engineering disciplines, particularly electrical engineering. Furthermore, the offshoots of signals and systems theory—e.g., control theory, signal processing, and communications theory—are themselves well-developed and equally basic to many engineering disciplines. As many a student will agree, the subject of signals and systems is one with a reliance on tools from many areas of mathematics. However, much of this mathematics is not revealed to undergraduates, and necessarily so. Indeed, a complete accounting of what is involved in signals and systems theory would take one, at times quite deeply, into the fields of linear algebra (and to a lesser extent, algebra in general), real and complex analysis, measure and probability theory, and functional analysis. Indeed, in signals and systems theory, many of these topics are woven together in surprising and often spectacular ways. The existing texts on signals and systems theory, and there is a true abundance of them, all share the virtue of presenting the material in such a way that it is comprehensible with the bare minimum background.

Should I bother reading these volumes?

This virtue comes at a cost, as it must, and the reader must decide whether this cost is worth paying. Let us consider a concrete example of this, so that the reader can get an idea of the sorts of matters the volumes in this text are intended to wrestle with. Consider the function of time

$$f(t) = \begin{cases} \mathrm{e}^{-t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

In the text (Example IV-6.1.3–2) we shall show that, were one to represent this function in the frequency domain with frequency represented by ν, we would get

$$\hat{f}(\nu) = \int_{\mathbb{R}} f(t)\mathrm{e}^{-2\mathrm{i}\pi\nu t}\,\mathrm{d}t = \frac{1}{1 + 2\mathrm{i}\pi\nu}.$$

The idea, as discussed in Chapter IV-2, is that $\hat{f}(\nu)$ gives a representation of the "amount" of the signal present at the frequency ν. Now, it is desirable to be able to reconstruct f from $\hat{f}$, and we shall see in Section IV-6.2 that this is done via the formula

$$f(t)\text{“}=\text{”} \int_{\mathbb{R}} \hat{f}(\nu)\mathrm{e}^{2\mathrm{i}\pi\nu t}\,\mathrm{d}\nu. \tag{FT}$$

The easiest way to do the integral is, of course, using a symbolic manipulation program. I just tried this with MATHEMATICA®, and I was told it could not do the computation. Indeed, the integral *does not converge*! Nonetheless, in many tables of

Fourier transforms (that is what the preceding computations are about), we are told that the integral in (FT) does indeed produce $f(t)$. Are the tables wrong? Well, no. But they are only correct when one understands exactly what the right-hand side of (FT) means. What it means is that the integral converges, *in* $\mathsf{L}^2(\mathbb{R};\mathbb{C})$ to f. Let us say some things about the story behind this that are of a general nature, and apply to many ideas in signal and system theory, and indeed to applied mathematics as a whole.

1. The story—it is the story of the L^2-Fourier transform—is not completely trivial. It requires *some* delving into functional analysis at least, and some background in integration theory, if one wishes to understand that "L" stands for "Lebesgue," as in "Lebesgue integration." At its most simple-minded level, the theory is certainly understandable by many undergraduates. Also, at its most simple-minded level, it raises more questions than it answers.
2. The story, even at the most simple-minded level alluded to above, takes some time to deliver. The full story takes *a lot* of time to deliver.
3. It is not necessary to fully understand the story, perhaps even the most simple-minded version of it, to be a user of the technology that results.
4. By understanding the story well, one is led to new ideas, otherwise completely hidden, that are practically useful. In control theory, quadratic regulator theory, and in signal processing, the Kalman filter, are examples of this.
5. The full story of the L^2-Fourier transform, and the issues stemming from it, directly or otherwise, is beautiful.

The nature of the points above, as they relate to this series, are as follows. Points 1 and 2 indicate why the story cannot be told to all undergraduates, or even most graduate students. Point 3 indicates why it is okay that the story not be told to everyone. Point 4 indicates why it is important that the story be told to someone. Point 5 should be thought of as a sort of benchmark as to whether the reader should bother with understanding what is in this series. Here is how to apply it. If one reads the assertion that this is a beautiful story, and their reaction is, "Okay, but there better be a payoff," or, "So what?" or, "Beautiful to who?" then perhaps they should steer clear of this series. If they read the assertion that this is a beautiful story, and respond with, "Really? Tell me more," then I hope they enjoy these books. They were written for such readers. Of course, most readers' reactions will fall somewhere in between the above extremes. Such readers will have to sort out for themselves whether the volumes in this series lie on the right side, for them, of being worth reading. For these readers I will say that this series is *heavily* biased towards readers who react in an unreservedly positive manner to the assertions of intrinsic beauty.

For readers skeptical of assertions of the usefulness of mathematics, an interesting pair of articles concerning this is [Wigner 1960] and [Hamming 1980].

What is the best way of getting through this material?

Now that a reader has decided to go through with understanding what is in these volumes, they are confronted with actually doing so: a possibly nontrivial matter, depending on their starting point. Let us break down our advice according to the background of the reader.

I look at the tables of contents, and very little seems familiar. Clearly if nothing seems familiar at all, then a reader should not bother reading on until they have acquired an at least passing familiarity with some of the topics in the book. This can be done by obtaining an undergraduate degree in electrical engineering (or similar), or pure or applied mathematics.

If a reader already possess an undergraduate degree in mathematics or engineering, then certainly some of the following topics will appear to be familiar: linear algebra, differential equations, some transform analysis, Fourier series, system theory, real and/or complex analysis. However, it is possible that they have not been taught in a manner that is sufficiently broad or deep to quickly penetrate the texts in this series. That is to say, relatively inexperienced readers will find they have some work to do, even to get into topics with which they have some familiarity. The best way to proceed in these cases depends, to some extent, on the nature of one's background.

I am familiar with some or all of the applied topics, but not with the mathematics. For readers with an engineering background, even at the graduate level, the depth with which topics are covered in these books is perhaps a little daunting. The best approach for such readers is to select the applied topic they wish to learn more about, and then use the text as a guide. When a new topic is initiated, it is clearly stated what parts of the book the reader is expected to be familiar with. The reader with a more applied background will find that they will not be able to get far without having to unravel the mathematical background almost to the beginning. Indeed, readers with a typical applied background will normally be lacking a good background in linear algebra and real analysis. Therefore, they will need to invest a good deal of effort acquiring some quite basic background. At this time, they will quickly be able to ascertain whether it is worth proceeding with reading the books in this series.

I am familiar with some or all of the mathematics, but not with the applied topics. Readers with an undergraduate degree in mathematics will fall into this camp, and probably also some readers with a graduate education in engineering, depending on their discipline. They may want to skim the relevant background material, just to see what they know and what they don't know, and then proceed directly to the applied topics of interest.

I am familiar with most of the contents. For these readers, the series is one of reference books.

Comments on organisation

In the current practise of teaching areas of science and engineering connected with mathematics, there is much emphasis on "just in time" delivery of mathematical ideas and techniques. Certainly I have employed this idea myself in the classroom, without thinking much about it, and so apparently I think it a good thing. However, the merits of the "just in time" approach in written work are, in my opinion, debatable. The most glaring difficulty is that the same mathematical ideas can be "just in time" for multiple non-mathematical topics. This can even happen in a single one semester course. For example—to stick to something germane to this series—are differential equations "just in time" for general system theory? for modelling? for feedback control theory? The answer is, "For all of them," of course. However, were one to choose one of these topics for a "just in time" written delivery of the material, the presentation would immediately become awkward, especially in the case where that topic were one that an instructor did not wish to cover in class.

Another drawback to a "just in time" approach in written work is that, when combined with the corresponding approach in the classroom, a connection, perhaps unsuitably strong, is drawn between an area of mathematics and an area of application of mathematics. Given that one of the strengths of mathematics is to facilitate the connecting of seemingly disparate topics, inside and outside of mathematics proper, this is perhaps an overly simplifying way of delivering mathematical material. In the "just simple enough, but not too simple" spectrum, we fall on the side of "not too simple."

For these reasons and others, the material in this series is generally organised according to its mathematical structure. That is to say, mathematical topics are treated independently and thoroughly, reflecting the fact that they have life independent of any specific area of application. We do not, however, slavishly follow the Bourbaki[1] ideals of logical structure. That is to say, we do allow ourselves the occasional forward reference when convenient. However, we are certainly careful to maintain the standards of deductive logic that currently pervade the subject of "mainstream" mathematics. We also do not slavishly follow the Bourbaki dictum of starting with the most general ideas, and proceeding to the more specific. While there is something to be said for this, we feel that for the subject and intended readership of this series, such an approach would be unnecessarily off-putting.

Andrew D. Lewis Kingston, ON, Canada

[1]Bourbaki refers to "Nicolas Bourbaki," a pseudonym given (by themselves) to a group of French mathematicians who, beginning in mid-1930's, undertook to rewrite the subject of mathematics. Their dictums include presenting material in a completely logical order, where no concept is referred to before being defined, and starting developments from the most general, and proceeding to the more specific. The original members include Henri Cartan, André Weil, Jean Delsarte, Jean Dieudonné, and Claude Chevalley, and the group later counted such mathematicians as Roger Godement, Jean-Pierre Serre, Laurent Schwartz, Emile Borel, and Alexander Grothendieck among its members. They have produced eight books on fundamental subjects of mathematics.

Preface for Volume 5

This final volume in this series is the second of the two core volumes in terms of our principal theme of a mathematical theory of signals and systems. In this volume we develop some aspects of system theory. As we explain in the text, system theory is an enormous topic and one that simply cannot be presented in a comprehensive way, never mind within the confines of one piece of work. Our focus, therefore, is on a certain sort of system, namely one described by differential or difference equations and with a finite-dimensional state space. We give special attention to linear systems, although we do thoroughly develop some aspect of the theory outside the traditional linear confines. We do this in order to develop some aspect of system theory in a sufficiently general way that we can discuss linear systems in a useful context.

We begin this volume in a manner similar to how we began our volume on signal theory: by giving some extensive motivation in Chapter 1. We do this by developing a large number of examples that show how differential and difference equations arise in many applications, coming from engineering, science, economics, and social science. We also consider examples that illustrate the sort of problems that system theory is designed to deal with.

After the presentation of these motivational examples, we present in Chapter 2 a framework for "general system theory." We do this so that we can discuss, separately from the specific settings we deal with later in the volume, the characteristics of systems that one might encounter. In this way, when we encounter these in the subsequent specific settings, the questions about structure are real questions and not just properties. We also use the development of a general setting for system theory to consider again some system theoretic problems, now in a more precise way than we were able to do in Chapter 1 in the context of examples.

The models we present in Chapter 1, and the models we will study in later chapters, come in the form of differential and difference equations. The former arise in continuous-time settings while the latter arise in discrete-time settings. We carry forward the project initiated in Volume 4 of a parallel development of continuous- and discrete-time cases, now in the context of system theory. This begins in earnest in Chapter 3 where we discuss differential and difference equations in generality. Here we go to great lengths to precisely say what differential and difference equations *are*, and not just give specific examples of these and them talk about how to solve them; this is the usual way in which these objects are considered. We develop differential and difference equations in a setting whose generality exceeds what we will encounter in later chapters. We do this for the usual reasons, namely that the generality adds context. Specifically, we develop a framework where partial differential and partial difference equations are considered, although we do not do anything with these later. One device that we carefully develop (and which is often not carefully developed) is that of the flow for an ordinary differential or difference equation.

In the next two chapters, 4 and 5, we consider classes of ordinary differential and difference equations in detail. The focus in both chapters is on linear equations, and as such the material we develop here is part of the undergraduate education for all engineer students and most physical science students. Our presentation of this material is different from what most such students will encounter. The differences arise in the by now expected ways: we do not focus so much on the computational facets of these topics, instead concentrating on structure and on the presentation of general results that will be useful to us later. Students who have had a normal undergraduate course in differential equations will recognise some of the main topics we cover, but will find the details we present rather different. Also, we develop difference equations alongside differential equations as equals. A presentation of difference equations at the level of generality we give seems difficult to find in the existing literature. We do not get deeply into the topic of dynamical systems—one direction one can go after learning about differential and difference equations—we do introduce some qualitative methods for studying these equations, as these qualitative methods are more insightful than the usual presentation of mechanical computational procedures. While we do not spend as much effort on computation techniques as is often seen in a course on differential equations, we do recognise that computation is essential. With this in mind, we do present sections on the use of computer tools for numerical solution of differential and difference equations.

The beating heart of this volume is Chapter 6 where we introduce eight classes of systems, coming in three pairs according to continuous-time/discrete-time, (state space)/(input/output), and linear/(not linear). A central focus of our development is the focus of continuity property of systems. This is not an emphasis one typically sees in presentations of system theory, either at the introductory or advanced level. It is here that we make careful and systematic use of our topologies for signals presented in Sections IV-1.3 and IV-1.2, and which rely on the machinery to which we devoted much of Volume 3. We also see that convolution, presented in detail in Chapter IV-4, features prominently for linear systems.

The next two chapters, 7 and 8, deal specifically with topics connected to linear system theory: the transfer function and the frequency response. These connect, respectively, the Laplace transforms and the Fourier transforms to the theory of linear systems, and provide a set of tools that are very useful in fields where linear system theory is applied, such as control theory and signal processing.

finish

Andrew D. Lewis Kingston, ON, Canada

Table of Contents

This version: 2022/03/07

Chapter 1

Motivation for system theory

Throughout this volume we will use examples of systems from engineering, physics, biology, economics, etc., to illustrate concepts from system theory. In this chapter we present a large number of examples to show the breadth of problems that can fit under the umbrella of the general theory we present. This will, we hope, make it possible for readers with various backgrounds (and with sufficient mathematical background) to read this volume and relate to the tools that are introduced. We shall not provide a comprehensive overview of any of the subject areas we rough upon, since to do this would be the subject of a treatise devoted solely to modelling, and it is not our intention to provide this. A reader needing a more in-depth treatment of the modelling for any of the areas we present can refer to the literature we review in Section 1.3.[1]

We breakdown our presentation into two parts. In Section 1.1 we illustrate how differential and difference equation models arise in various sorts of models. In this volume, we shall restrict our attention to specific sorts of differential and difference equations, namely ordinary differential and difference equations. In Section 1.1 we consider some models that do not fall into this category, in order to illustrate that there are many interesting models that do not fall exactly into the framework we develop. In Section 1.2, equipped with differential and difference equation models, we present some ideas in system theory that arise in these models.

Do I need to read this chapter? A reader who feels like they need something motivational to cling to as we go through the detailed mathematical presentation will definitely need to read this chapter. Other readers will find it merely enjoyable. •

Contents

[1] In some areas, a truly mathematical presentation of the modelling background would be a significant contribution, in and of itself, and we hope that this is undertaken at some point in all cases.

Section 1.1

How do differential and difference equations arise in mathematical modelling?

We shall almost exclusively study models that arise from differential and difference equations. It is possible to approach the subject of differential and difference equations from a purely mathematical point of view. And, indeed, even if one is interested in only applying the theory of differential and difference equations in specific areas, a good knowledge of this mathematical subject is necessary. However, a primary reason for the importance of differential and difference equations in mathematics is that they arise so naturally and broadly in areas of application, ranging from engineering, physics, economics, and biology, to name a few. Indeed, it may not be inaccurate to say that differential and difference equations provide the most important (but definitely not the only) conduit from developments in mathematics to applications. In this section, we illustrate this with an array of examples.

Caveat We mainly shall not be precise in this section with things like whether functions are continuous, differentiable, etc. In the remainder of the text we shall be more careful about these things. •

1.1.1 Mass-spring-damper systems

Let us start by considering a single mass connected to the ground by a spring and a damper, as in Figure 1.1. The mass has mass m, the spring is a linear spring with

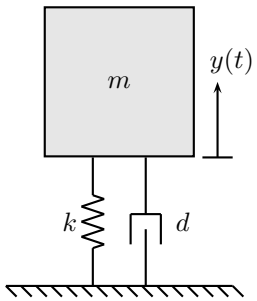


Figure 1.1 A simplified model of a car suspension

a restoring force proportional to the change in length from its equilibrium—i.e., the spring force is $-k\Delta$, $k \geq 0$, where Δ is the change in length— and the damper is also linear with a restoring force proportional of the velocity at which the damper is contracted–i.e., the damper force is $-d\dot{\Delta}$, $d \geq 0$, where "$\dot{}$" means "derivative with respect to time." This may be thought of as a simple model for a car suspension.

We shall derive an equation that governs the vertical motion of the mass as a function of time. We let $y(t)$ be the vertical displacement of the mass, with the assumption that $y = 0$ corresponds to the undeflected position of the spring. We suppose that we have a gravitational force acting "downwards" in the diagram and with a gravitational constant a_g. One then performs a force balance, setting vertical forces equal to the mass times the acceleration:

$$-d\dot{y}(t) - ky(t) - ma_g = m\ddot{y}(t) \iff m\ddot{y}(t) + d\dot{y}(t) + ky(t) = -ma_g. \tag{1.1}$$

Note that this is an equation with single independent variable t (time) and single dependent variable y (vertical displacement). Moreover, the equation us *not* an algebraic equation for y as a function of t, since derivatives of y with respect to t arise.

During the course of this volume, we shall learn how to exactly solve a differential equation like this. But before we do so, let us see if we can, based on our common sense, deduce what sort of behaviour a system like this should exhibit. First let's determine the equilibrium of the system, since it is *not* when $y = 0$, because of the gravitational force. Indeed, as equilibrium the mass should not be in motion and so we ought to have $\dot{y} = 0$ and $\ddot{y} = 0$. In this case, $y = -\frac{ma_g}{k}$. Now let's think about what happens when $d = 0$. What we expect here is that the mass will oscillate in the vertical direction around the equilibrium. Moreover, we may expect that as k becomes relatively larger, the frequency of oscillations will increase. Now, adding the damping constant $d > 0$, perhaps our intuition is not quite so reliable a means of deducing what is going on here. But what happens is this: the damper dissipates energy. This causes the oscillations to decay to zero as $t \to \infty$. Moreover, if d gets relatively large, it actually happens that the oscillations do not occur, and the mass just moves towards its equilibrium. These are things we will investigate systematically.

Next let us complicate matters a little, and consider two interconnected masses as in Figure 1.2. In this case, to simplify things we interconnect the masses only

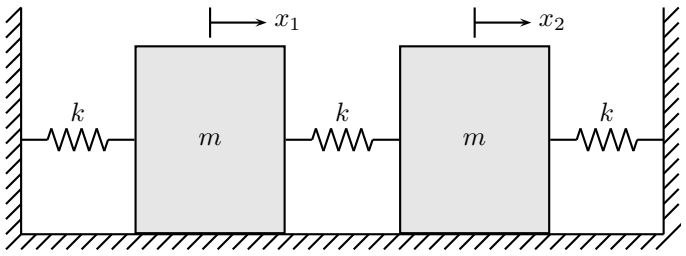


Figure 1.2 Interconnected masses

with springs. As in the figure, we let x_1 and x_2 denote the positions of the masses, assuming that all springs are uncompressed with $x_1 = x_2 = 0$. In this case, the force

balance equations for the two masses give the equations

$$\begin{aligned} -kx_1(t) - k(x_1(t) - x_2(t)) = m\ddot{x}_1(t), &\iff m\ddot{x}_1(t) + 2kx_1(t) - x_2(t) = 0,\\ -kx_2(t) - k(x_2(t) - x_1(t)) = m\ddot{x}_2(t), &\iff m\ddot{x}_2(t) + 2kx_2(t) - x_1(t) = 0. \end{aligned}$$

Let us express this using matrix/vector notation:

$$m\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \ddot{x}_1(t) \\ \ddot{x}_2(t) \end{bmatrix} + k\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

If we introduce the notation

$$\boldsymbol{M} = m\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{K} = k\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad \boldsymbol{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix},$$

then we can further write this as

$$\boldsymbol{M}\ddot{\boldsymbol{x}}(t) + \boldsymbol{K}\boldsymbol{x}(t) = \boldsymbol{0}. \tag{1.2}$$

Note that this is an equation with single independent variable t (time) and two dependent variables x_1 and x_2, or equivalently a vector dependent variable $(x_1, x_2) \in \mathbb{R}^2$ (horizontal displacements). As was the case with the single mass, the key point is that the equation involves derivatives of the dependent variables with respect to the independent variable.

In the text, we will see how to analyse such equations as this. Let us say a few words about the most interesting features of how this system behaves. There are two interesting classes of behaviours, one occurring when $x_1(t) = x_2(t)$ (the masses move together) and one occurring when $x_1(t) = -x_2(t)$ (the masses move exactly opposite one another). These "modes" of the system are important, as we shall see that every solution is a linear combination of these two. This has to do with fundamental properties of systems of this general type.

1.1.2 The motion of a simple pendulum

Let us consider the motion of a pendulum as depicted in Figure 1.3. We suppose that we have a mass m attached to a rod of length ℓ whose mass we consider to be negligible compared to m. We have a gravitational force with gravitational constant a_g that acts downward in the figure. Summing moments about the pivot point gives

$$-ma_g\ell\sin\theta(t) = m\ell^2\ddot{\theta}(t) \iff \ddot{\theta}(t) + \frac{a_g}{\ell}\sin\theta(t) = 0. \tag{1.3}$$

This is an equation in a single independent variable t (time) and a single dependent variable θ (pendulum angle), and again is an equation in derivatives of the dependent variable with respect to the independent variable.

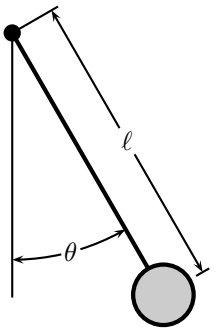


Figure 1.3 A simple pendulum

We shall *not* learn how to solve this equation in this text, although a "closed-form solution" is possible with a suitably flexible notion of "closed-form." However, problems such as this one call into question the value of having a closed-form solution. What is, perhaps, a more useful way to understand the behaviour of a simple pendulum is to try some sort of approximation. We shall make an approximation near the two equilibria of the pendulum, corresponding to $\theta = 0$ (the "down" equilibrium) and $\theta = \pi$ (the "up" equilibrium). To make the approximation, we note that, for ϕ near zero,

$$\begin{aligned}&\sin\phi \approx \phi,\\ &\sin(\pi + \phi) = \sin\pi\cos\phi + \cos\pi\sin\phi \approx -\phi.\end{aligned}$$

Therefore, the equation governing the behaviour of the simple pendulum are approximated near $\theta = 0$ (say $\theta = 0 + \phi$) by

$$\ddot{\phi}(t) + \frac{a_g}{\ell}\phi(t) = 0.$$

We shall see during the course of our studies that a general solution to these equations takes the form

$$\phi(t) = \phi(0)\cos(\omega\phi(t)) + \frac{\dot{\phi}(0)}{\omega}\sin(\omega\phi(t)),$$

where $\omega = \sqrt{a_g/\ell}$. Thus, if the approximation is valid, this suggests that the motion of the simple pendulum, for small angles, consists of periodic motions with frequency ω. It turns out that this behaviour is indeed close to that of the genuine pendulum equations. To be precise, the motion of the pendulum for small angles is indeed periodic, and as the angle gets smaller, the frequency approaches ω. However, the motion is *not* sinusoidal. Moreover, the period gets larger for larger amplitude motions.

A very large amplitude motion would be when θ starts at π. If we take $\theta = \pi + \phi$ then the governing equation is approximately

$$\ddot{\phi}(t) - \frac{a_g}{\ell}\phi(t) = 0.$$

We shall see that a general solution to these equations takes the form

$$\phi(t) = \phi(0)\cosh(\omega\phi(t)) + \frac{\dot{\phi}(0)}{\omega}\sinh(\omega\phi(t)), \tag{1.4}$$

where $\omega = \sqrt{a_g/\ell}$. (Here cosh and sinh are the hyperbolic cosine and sine functions, defined by

$$\cosh(x) = \tfrac{1}{2}(e^x + e^{-x}), \quad \sinh(x) = \tfrac{1}{2}(e^x - e^{-x}).)$$

For most values of $\dot{\phi}(0)$ and $\dot{\phi}(0)$, the solutions of this equation diverge to ∞ as $t \to \infty$. Of course, as ϕ gets large, this approximation becomes unreliable. Nonetheless, the behaviour observed for small times agrees with what we think the dynamics ought to be: since the "up" equilibrium is unstable, trajectories generally move away from this equilibrium. Note, however, that there are a small number of the solutions (1.4) that do not diverge to ∞, but approach $\phi = 0$ as $t \to \infty$, namely those for which $\phi(0) = -\frac{\dot{\phi}(0)}{\omega}$. In terms of the physics of the pendulum, these solutions correspond to the motions of the pendulum where the pendulum swings with just enough energy to approach the upright equilibrium as $t \to \infty$.

1.1.3 Bessel's equation

We shall not motivate here precisely how the equation we consider in this section arises in practice. We shall be content with the following description: If one tries to solve the potential equation (1.19) in two-dimensions and in polar coordinates, then one arrives at the equation

$$r^2\frac{\partial^2 y}{\partial r^2} + r\frac{\partial y}{\partial r} + (r^2 - \alpha^2)y = 0, \tag{1.5}$$

for $\alpha \in \mathbb{R}$ (actually, in the particular case of the potential equation, α is a nonnegative integer). This equation, for example, describes the radial displacement in a drum when it has been struck. The equation is known as ***Bessel's equation***.

We note that Bessel's equation has one independent variable r, one dependent variable y, and is an equation in the derivatives of the dependent variable with respect to the independent variable.

1.1.4 RLC circuits

Next let us consider differential equations such as arise in circuits comprised of ideal resistors, inductors, and capacitors. Let us define these terms. We will use "E," "I," and "q" to denote voltage, current, and charge, respectively.

1. A ***resistor*** is a device across which the voltage drop is proportional to the current through the device. The constant of proportionality is the ***resistance*** R: $E = RI$.
2. An ***inductor*** is a device across which the voltage drop is proportional to the time rate of change of current through the device. The constant of proportionality is the ***inductance*** L: $E = L\frac{\mathrm{d}I}{\mathrm{d}t}$.

3. A ***capacitor*** is a device across which the voltage drop is proportional to the charge in the device. The constant of proportionality is the $\frac{1}{C}$ with C being the ***capacitance***: $E = \frac{1}{C}q$.

The three devices are typically given the symbols as in Figure 1.4. The physical

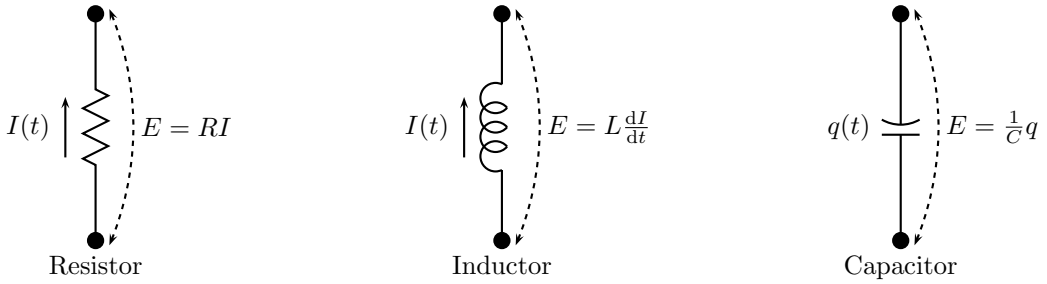


Figure 1.4 Electrical devices

laws governing the behaviour of ideal circuits are:

1. the current I is related to the charge q by $I = \frac{dq}{dt}$;
2. ***Kirchhoff's voltage law*** states that the sum of voltage drops around a closed loop must be zero;
3. ***Kirchhoff's current law*** states that the sum of the currents entering a node must be zero.

Given a collection of such devices arranged in some way—i.e., a "circuit"—along with voltage and/or current sources, we can imagine that governing equations for the behaviour of the circuit can be deduced. In Figure 1.5 we have a particularly

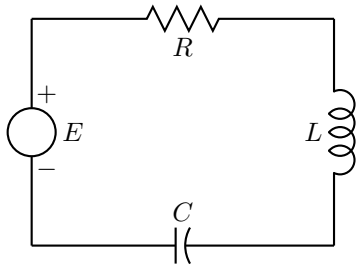


Figure 1.5 A series *RLC* circuit

simple configuration. The voltage drop around the circuit must be zero which gives the governing equations

$$E(t) = RI(t) + L\dot{I}(t) + \tfrac{1}{C}q(t) \qquad \Longrightarrow \qquad L\ddot{q}(t) + R\dot{q}(t) + \tfrac{1}{C}q(t) = E(t)$$

where $E(t)$ is an external voltage source. This may also be written as a current equation by merely differentiating:

$$L\ddot{I}(t) + R\dot{I}(t) + \tfrac{1}{C}I(t) = \dot{E}(t). \tag{1.6}$$

In either case, we have an equation in a single independent variable (time) and a single dependent variable (charge or current). The equations involve, of course, derivatives of the dependent variable with respect to the dependent variable.

We comment here on similarity with the equation (1.6) with the equation (1.1) describing the motion of a damped mass/spring system are worth remarking upon. The capacitor plays the rôle of a spring (stores energy), the resistor plays the rôle of a damper (dissipates energy), and the inductor plays the rôle of a mass (it energy is obtained from "motion" in the circuit). This gives rise to an important "electro-mechanical analogy" in the modelling of physical systems.

1.1.5 Tank systems

Here we consider two tanks with fluid in a configuration shown in Figure 1.6. Here are the variables and parameters:

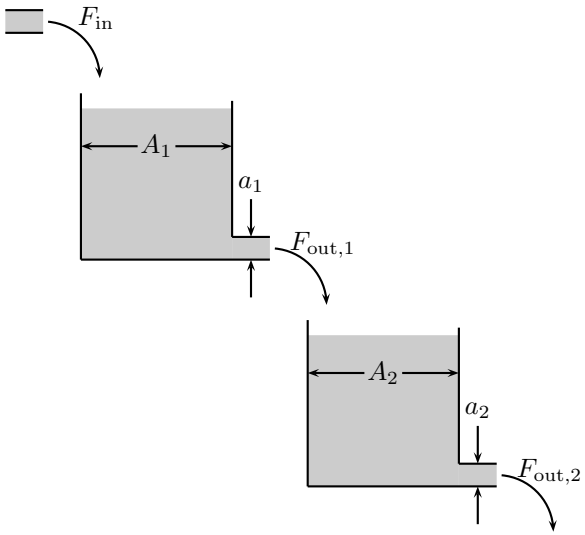


Figure 1.6 Mass balance in coupled tank flow

F_{in}	volume flow into tank 1
$F_{\text{out,j}}$	volume flow out of tank j, $j \in \{1, 2\}$
A_j	cross-sectional area of tank j, $j \in \{1, 2\}$
a_j	cross-sectional area of orifice j, $j \in \{1, 2\}$
h_j	height of water in tank j, $j \in \{1, 2\}$

Let us state the rules we shall use to deduce the behaviour of the system, assuming that the fluid is "incompressible" so the mass of a given volume of fluid will be constant:

1. according to ***Bernoulli's Law***, the velocity of the fluid exiting a small orifice at the bottom of a tank with level h is $\sqrt{2a_g h}$, where a_g is the acceleration due to gravity;
2. the volume of rate of fluid flow passing through an orifice with constant cross-sectional area A with velocity v (assumed to be constant across the cross-section) is Av;
3. the rate of change of volume in a tank with constant cross-sectional area A and fluid height h is $A\frac{\mathrm{d}h}{\mathrm{d}t}$.

We can thus form the balance equations for each tank by setting the rate of change of volume in the tank equal to the volume flow in minus the volume flow out:

$$\begin{aligned} A_1\dot{h}_1(t) &= F_{\text{in}}(t) - F_{\text{out},1} = F_{\text{in}}(t) - \sqrt{2a_1h_1(t)}, \\ A_2\dot{h}_2(t) &= F_{\text{out},1}(t) - F_{\text{out},2} = \sqrt{2a_1h_1(t)} - \sqrt{2a_2h_2(t)}. \end{aligned} \tag{1.7}$$

The equations governing the behaviour of the system have one independent variable t (time) and two dependent variables h_1 and h_2, or a single vector variable $(h_1, h_2) \in \mathbb{R}^2$ (the heights of fluid in the tanks). As with all of our examples, the equations involve the derivatives of the dependent variables with respect to the independent variable.

1.1.6 Population models

An important area of application of differential equations is in biological sciences, in areas such as epidemiology and population dynamics. We shall consider here two simple models of population dynamics as an illustration.

First let us consider a population that we model as a scalar variable $p \in \mathbb{R}$. First we consider a situation where the rate of population growth is proportional to p for small values of p, but then diminishes as we approach some "limiting population size, p_0, representing the fact that there may be limited resources. This can be represented by a model like

$$\dot{p}(t) = kp(t)\left(1 - \frac{p(t)}{p_0}\right). \tag{1.8}$$

This is often referred to as the ***logistical model*** of population dynamics. This is an equation with a single independent variable t (time) and a single dependent variable p (population).

While we will not explicitly examine this equation in this text, the reader may relatively easily verify the following behaviour, under the natural assumption that $k > 0$.

1. There is an equilibrium at $p = 0$ that is not stable. That is, for small positive populations, the rate of population change is positive.
2. There is an equilibrium at $p = p_0$ that is stable. That is, for populations less than the limiting population p_0, the rate of population change is positive.

Let us now consider two populations a and b, with a representing the population of a prey species and b representing the population of a predator species. The following assumptions are made:

1. prey population increases exponentially in the absence of predation;
2. predators die off exponentially in the absence of predation;
3. predator growth and prey death due to predation is proportional to the rate of predation;
4. the rate of predation is proportional to the encounters between predators and prey, and encounters themselves are proportional to the populations.

Putting all of this together, the behaviour of the prey population a can be modelled by

$$\dot{a}(t) = \alpha a(t) - \beta a(t)b(t)$$

and the behaviour of the predator population can be modelled by

$$\dot{b}(t) = \delta a(t)b(t) - \gamma b(t).$$

We should combine these equations:

$$\begin{aligned} \dot{a}(t) &= \alpha a(t) - \beta a(t)b(t), \\ \dot{b}(t) &= \delta a(t)b(t) - \gamma b(t). \end{aligned} \tag{1.9}$$

These equations have a single independent variable t (time) and two dependent variables a and b, or equivalently a single vector variable $(a, b) \in \mathbb{R}^2$. This model is called the ***Lotka–Volterra predator-prey model***.

We shall not in this text undertake a detailed analysis of this equation. However, a motivated reader can easily find many sources where this model is discussed in great depth and detail.

1.1.7 Economics models

Another area where differential equations are useful is in social sciences, and especially economics. We consider an example of this, known as the ***Rapoport production and exchange model***.

The setup is this. Individuals A and B produce goods that we measure by scalar variables $a, b \in \mathbb{R}$. The individuals A and B trade, each trying to maximise their "happiness," typically referred to as "utility."[2] We denote by p the proportion of

[2]In philosophy, the notion of "utility" as a measure of general happiness dates, in its most explicit form, to Thomas Hobbes (1588–1679) and John Locke (1632–1704). While early versions of utilitarianism were based in religion, John Stuart Mill (1806–1873) developed a powerful secular utilitarian ethic, which itself led to the secular philosophy of Jeremy Bentham (1748–1832).

goods produced and retained, and by q the proportion of goods produced and traded: thus $p + q = 1$. The assumptions made by Rapoport are these:

1. people are lazy, so the act of production is a loss of utility;
2. people are gauche, so possessing something produced is a gain in utility;
3. the loss of utility due to the agonies of production are proportional to the amount produced;
4. while there is no cap in a person's desire to acquire crap, the utility they gain from acquiring crap diminishes, the more crap they have;
5. the rate at which A or B makes product a and b is proportional to the rate at which utility increases with respect to a and b.

With all this as backdrop, let us introduce something meaningful. First of all, let us give the utility functions for A and B:

$$U_A(a, b) = \log(1 + pa + qb) - r_A a, \quad U_B(a, b) = \log(1 + qa + pb) - r_B b.$$

If one examines these expressions, one can see that they capture in form and shape the characteristics of individuals A and B described above. Of course, many other forms are also viable candidates.

Now, according to condition 5, the equations that govern the amounts a and b are:

$$\begin{aligned} \dot{a}(t) &= c_A \left(\frac{p}{1 + pa(t) + qb(t)} - r_A a(t) \right), \\ \dot{b}(t) &= c_B \left(\frac{p}{1 + pa(t) + qb(t)} - r_B b(t) \right). \end{aligned} \tag{1.10}$$

These equations have a single independent variable t (time), and two dependent variables a and b, or equivalently one vector variable $(a, b) \in \mathbb{R}^2$ (production). The equation is one that involves the derivatives of the dependent variables with respect to the independent variable.

An indepth analysis of these equations is not something we will undertake here.

1.1.8 Euler–Lagrange equations

We consider here the following problem. Suppose we are given $y_1, y_2 \in \mathbb{R}$ and $x_1, x_2 \in \mathbb{R}$ with $x_1 < x_2$. Denote by

$$\begin{aligned} &\Gamma(y_1, y_2, x_1, x_2) \\ &= \{\gamma \colon [x_1, x_2] \to \mathbb{R} \mid \gamma \text{ is twice continuously differentiable } \gamma(x_1) = y_1,\ \gamma(x_2) = y_2\} \end{aligned}$$

the set of all twice continuously differentiable functions with value y_1 at the left endpoint and y_2 at the right endpoint, as in Figure 1.7. Suppose that we have a function $L \colon [x_1, x_2] \times \mathbb{R}^2 \to \mathbb{R}$ that we call the ***Lagrangian***. Associated to this Lagrangian and a function $\gamma \in \Gamma(y_1, y_2, x_1, x_2)$ we have an associated ***cost***

$$C_L(\gamma) = \int_{x_1}^{x_2} L(x, \gamma(x), \gamma'(x)) \, \mathrm{d}x.$$

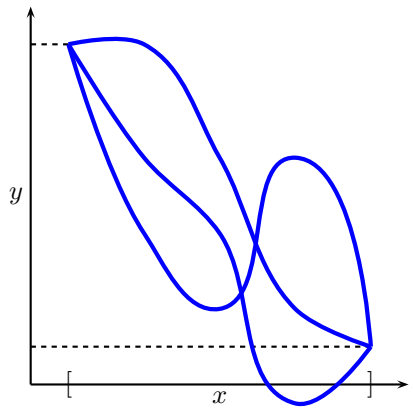


Figure 1.7 Candidate curves in an optimisation problem

The objective is to find γ that minimises $C_L(\gamma)$. That is, we seek $\gamma_* \in \Gamma(y_1, y_2, x_1, x_2)$ such that

$$C_L(\gamma_*) \leq C_L(\gamma), \qquad \gamma \in \Gamma(y_1, y_2, x_1, x_2).$$

Such a function γ_* is a ***minimiser*** for the Lagrangian L. One can show, without much difficulty, but using methods from the calculus of variations that are a little far afield for us at the moment, that if γ_* is given by $\gamma_*(x) = y(x)$ is a minimiser for L, then it necessarily satisfies the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial y'}\right) - \frac{\partial L}{\partial y} = 0,$$

which are the ***Euler–Lagrange equations*** for this problem. We give the equations in their traditional form, although this form is genuinely confusing. Let us be a little more explicit about what the equations mean. By an application of the Chain Rule, the Euler–Lagrange equations can be written as

$$\frac{\partial^2 L}{\partial y' \partial y'} y''(x) + \frac{\partial^2 L}{\partial y' \partial y} y'(x) - \frac{\partial L}{\partial y} = 0. \tag{1.11}$$

Note that this is an equation in the single independent variable x and the single dependent variable y. Again, it is an equation involving derivatives of the dependent variable with respect to the independent variable. However, this equation has, in general, an important difference with some of the other equations we have seen. To illustrate this, let us consider the Lagrangians $L(x, y, y') = y'$. In this case

$$\frac{\partial^2 L}{\partial y' \partial y'} y''(x) + \frac{\partial^2 L}{\partial y' \partial y} y'(x) - \frac{\partial L}{\partial y}$$

is identically zero: a circumstance unlike the equations we have encountered before.

The Euler–Lagrange equations are important equations in physics and optimisation, but to study them in any depth is not something we will be able to undertake in this text.

1.1.9 Maxwell's equations

Maxwell's equations are famously important equations governing the behaviour of electromagnetic phenomenon. Let us introduce the physical variables of Maxwell's equations:

$\boldsymbol{E}$	electric field
$\boldsymbol{B}$	magnetic field
$\boldsymbol{J}$	current density
ρ	charge density

The first three of these quantities are vector fields on the physical space $\mathbb{R}^3$. Thus we should think of each of these physical quantities as defining a direction in $\mathbb{R}^3$ and a length at each point in $\mathbb{R}^3$, i.e., an arrow. The charge density ρ is a scalar-valued function on $\mathbb{R}^3$. Let us say a word or two about how we should interpret these quantities. First of all, the charge density ρ is relatively easy to understand: it prescribes the density of charge provided by subatomic particles per unit volume as we move through physical space. The electric field indicates how charge moves through space; at each point (x_1, x_2, x_3) in space, it moves in the direction of $\boldsymbol{E}(x_1, x_2, x_3)$. Thus $\boldsymbol{E}(x_1, x_2, x_3)$ can be thought of as a "force" acting on a charge at the point (x_1, x_2, x_3). The magnetic field $\boldsymbol{B}$[3] acts for magnetic field lines rather like the electric field acts from the flow of charge: it indicates the direction of magnetic force applied to a moving charge. The current density $\boldsymbol{J}$ gives the current, as a vector quantity, rather in the manner of a fluid flow.

There are also some physical constants in the equations of electromagnetism. These are the following:

ϵ_0	permittivity of free space
μ_0	permeability of free space

These constants are proportionality constants, rather in the manner of the acceleration due to gravity, which we have been denoting by a_g.

With this preparation, we shall produce ***Maxwell's equations*** which indicate

[3]There is another quantity $\boldsymbol{H}$ that also represents the magnetic field, and is proportional to $\boldsymbol{B}$ in a vacuum, but has a more complicated relationship within a magnetic material. Very often $\boldsymbol{H}$ is referred to as the magnetic field, and $\boldsymbol{B}$ is called something different. But often the name "magnetic field" is applied to $\boldsymbol{B}$

how these quantities interact with one another:

$$\begin{aligned} \epsilon_0 \nabla \cdot \boldsymbol{E} &= \rho, \\ \nabla \cdot \boldsymbol{B} &= \boldsymbol{0}, \\ \nabla \times \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t}, \\ \nabla \times \boldsymbol{B} &= \mu_0 \boldsymbol{J} + \mu_0 \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}. \end{aligned} \tag{1.12}$$

Let us first describe the mathematical symbols "$\nabla\cdot$" and "$\nabla\times$" that you will learn about in a course on vector calculus. The operator $\nabla\cdot$ is the ***divergence*** and acts on a vector field $\boldsymbol{X} = (X_1, X_2, X_3)$, giving a function according to the definition

$$\nabla \cdot \boldsymbol{X} = \frac{\partial X_1}{\partial x_1} + \frac{\partial X_2}{\partial x_2} + \frac{\partial X_3}{\partial x_3}.$$

The precise meaning of the divergence of a vector field requires a few ways of thinking about things that are not part of ones makeup prior to a course like this, but basically vanishing divergence corresponds to "volume preserving." The operator $\nabla\times$ is ***curl*** and again acts on a vector field $\boldsymbol{X} = (X_1, X_2, X_3)$ giving another vector field according to the definition

$$\nabla \times \boldsymbol{X} = \left(\frac{\partial X_2}{\partial x_3} - \frac{\partial X_3}{\partial x_2}, \frac{\partial X_1}{\partial x_3} - \frac{\partial X_3}{\partial x_1}, \frac{\partial X_2}{\partial x_1} - \frac{\partial X_1}{\partial x_2} \right).$$

As with divergence, a really good understanding of curl of a bit beyond us at this point. Let us say two things: (1) $\nabla \times \boldsymbol{X}$ measures the "rotationality" of a vector field $\boldsymbol{X}$, so its vanishing somehow means it is not rotational; (2) if $\nabla \times \boldsymbol{X} = \boldsymbol{0}$, then there exists a function f such that $\boldsymbol{X} = \nabla f$, with ∇f being the ***gradient*** of f:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3} \right).$$

What we can now see is that there are four independent variables (x_1, x_2, x_3, t) in Maxwell's equations, representing spacetime, and $3 + 3 + 3 + 1 = 10$ dependent variables $\boldsymbol{E}$, $\boldsymbol{B}$, $\boldsymbol{J}$, and ρ. The equations involve the partial derivatives of the dependent variables with respect to the independent variable.

Now we can say a few words about the meaning of Maxwell's equations. The first equation, called ***Gauss's law for electricity***, says that the "expansiveness" of the electric field is proportional to the charge density. The second equation, called ***Gauss's law for magnetism***, says that the expansiveness of the magnetic field is zero. The third equation, called ***Faraday's law of induction***, tells us that a time-varying magnetic field gives rise to an electric field. Finally, the fourth equation, called ***Ampére's law***, says that both a time-varying electric field and a current density field give rise to magnetic field.

Of course, any systematic investigation of Maxwell's equations is not something we can undertake here, and indeed in complete generality is not possible, by any reasonable meaning of "systematic investigation."

1.1.10 The Navier–Stokes equations

The Navier–Stokes equations deal with the motion of a Newtonian, viscous, and compressible fluid. This means (1) there are viscous, i.e., friction, effects that are accounted for, (2) the viscous stresses arise as a consequence of temporal deformation of the fluid, (3) and the mass of fluid in a given volume is allowed to vary. The motion of the fluid we represent by a mapping $\boldsymbol{\phi}\colon \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$, so that $\boldsymbol{\phi}(t, \boldsymbol{x})$ indicates where the fluid particle at $\boldsymbol{x} \in \mathbb{R}^3$ at time 0 resides at time t. We shall abbreviate $\boldsymbol{\phi}_t\colon \mathbb{R}^3 \to \mathbb{R}^3$ the mapping $\boldsymbol{\phi}_x(\boldsymbol{x}) = \boldsymbol{\phi}(t, \boldsymbol{x})$. We shall not deal directly with this mapping $\boldsymbol{\phi}$, but rather with its associated velocity field, by which we mean the mapping $\boldsymbol{u}\colon \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$\boldsymbol{u}(t, \boldsymbol{\phi}_t(\boldsymbol{x})) = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\phi}_t(\boldsymbol{x}).$$

Thus $\boldsymbol{u}(t, \boldsymbol{x})$ is the velocity of the fluid particle initially at position $\boldsymbol{x}$ at time t. In Figure 1.8 we illustrate how one can think of the velocity field by depicting the

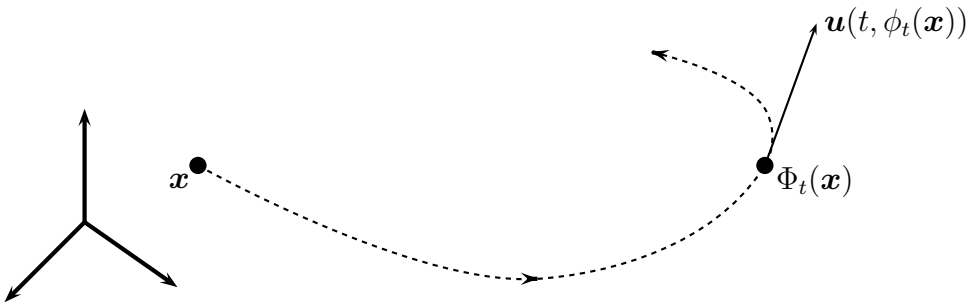


Figure 1.8 The velocity field for a fluid motion

trajectory followed by a single particle, along with the velocity of that particle at time t.

The Navier–Stokes equations are equations for the velocity field $\boldsymbol{u}$. The first part of these equations is the ***continuity equation***, which represents the law of conservation of mass:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0. \tag{1.13}$$

Here ρ is a scalar-valued function on $\mathbb{R}^3$ giving the mass density of the fluid as a function on physical space. The operator "$\nabla\cdot$" is the divergence which we encountered in our discussion of Maxwell's equations above. Note that when ρ

is constant—which corresponds to incompressible flow—the continuity equation reads

$$\nabla \cdot \boldsymbol{u} = 0,$$

meaning that the velocity field preserves volume. Along with the mass conservation equation, we have a force/momentum balance equation that we will not provide any details for:

$$\rho\left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u}\right) = -\nabla p + \nabla \cdot (\mu(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T) - \tfrac{2}{3}\mu(\nabla \cdot \boldsymbol{u})\boldsymbol{I}) + \boldsymbol{f}. \tag{1.14}$$

These are the ***Navier–Stokes equations***.

Let us first define all of the mathematical components of this equation, at least so one can imagine writing these equations down in explicit form. The term $\nabla \boldsymbol{u}$ is the ***gradient*** or ***Jacobian*** of the velocity field, which is a 3×3-matrix:

$$\nabla \boldsymbol{u} = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \frac{\partial u_1}{\partial x_3} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} & \frac{\partial u_2}{\partial x_3} \\ \frac{\partial u_3}{\partial x_1} & \frac{\partial u_3}{\partial x_2} & \frac{\partial u_3}{\partial x_3} \end{bmatrix}.$$

The second term in the Navier–Stokes equations is the vector obtained by multiplying this matrix on the left by the vector $\boldsymbol{u}$. The variable p is the ***pressure field*** which is a scalar function, and ∇p represents the gradient of the pressure field, i.e., the vector field $\operatorname{grad} p = (\frac{\partial p}{\partial x_1}, \frac{\partial p}{\partial x_2}, \frac{\partial p}{\partial x_3})$. The variable μ is the ***viscosity***, and represents the internal forces in the fluid due to friction causes when creating strain gradients. Of course, $\boldsymbol{I}$ is the 3×3 identity matrix. Note that the second term on the right-hand side has the form $\nabla \cdot \boldsymbol{M}$ for a matrix function $\boldsymbol{M}$. This is a vector field, called the ***divergence*** of M. It is given explicitly by

$$\nabla \cdot \boldsymbol{M} = \left(\sum_{j=1}^{3} \frac{\partial M_{1j}}{\partial x_j}, \sum_{j=1}^{3} \frac{\partial M_{2j}}{\partial x_j}, \sum_{j=1}^{3} \frac{\partial M_{3j}}{\partial x_j}\right).$$

Finally, $\boldsymbol{f}$ are ***body forces***, e.g., gravitational effects.

The Navier–Stokes equations have four independent variables (x_1, x_2, x_3, t) and five dependent variables, ρ, p, and (u_1, u_2, u_3). It is, of course, an equation in the derivatives of the dependent variables with respect to the independent variables.

1.1.11 Heat flow due to temperature gradients

Our next modelling task is that of heat flow in a homogeneous medium. Let us specify the physical assumptions we make.

1. For simplicity we work with a one-dimensional medium, i.e., a rod.

2. We assume a homogeneous medium, i.e., its characteristics are constant as we move throughout. We assume the rod to have a constant cross-sectional area A.
3. Thermal energy is given by $Q = c\rho Vu$, where ρ is the mass density, V is the volume, u is temperature, and c is the specific heat of the medium. We assume ρ and c to be constant throughout the material.
4. We assume that rate of heat transfer from one region to another through a slice of the rod is proportional to the temperature gradient:

$$q = -K\frac{\partial u}{\partial x},$$

   where q is the heat flow per unit area and x measures the distance along the rod. This is ***Fourier's law***.
5. Thermal energy is conserved in each chunk of the rod.

Let us use these assumptions to derive an equation governing the temperature distribution in a rod. Consider a chunk of the rod as shown in Figure 1.9. In the

Figure 1.9 A chunk of rod used in the derivation of the heat equation

figure, the rod chunk is shown at a fixed time. The quantity $q(a)$ denotes the rate of heat flow at the position $x = a$ on the rod, and $q(b)$ denotes the rate of heat flow at the position $x = b$ on the rod. In terms of the quantities in Figure 1.9, Fourier's law reads

$$q(a) = -K\frac{\partial u}{\partial x}\Big|_a, \quad q(b) = K\frac{\partial u}{\partial x}\Big|_b$$

for some constant $c > 0$. The signs result from the fact that heat will flow in a direction opposite the temperature gradient. If we assume that no heat escapes from the upper and lower boundaries of the rod, then the net change in heat in the rod chunk in a time Δt will be

$$\Delta Q = KA\Delta t\left(\frac{\partial u}{\partial x}\Big|_b - \frac{\partial u}{\partial x}\Big|_a\right), \tag{1.15}$$

With the assumptions we have made, the net change in heat in the chunk over a time Δt is given by

$$\Delta Q = c\rho A(b-a)\Delta t \frac{\partial u}{\partial t}, \tag{1.16}$$

where $\frac{\partial u}{\partial t}$ is the average of the time rate of change of temperature throughout the chunk and ρ is the mass density of the material. By making $(b-a)$ and Δt sufficiently small, one may ensure that $\frac{\partial u}{\partial t}$ does not vary much through the chunk. Equating (1.15) and (1.16) we get

$$c\rho A(b-a)\Delta t \frac{\partial u}{\partial t} = KA\Delta t \left(\frac{\partial u}{\partial x}\Big|_b - \frac{\partial u}{\partial x}\Big|_a \right)$$

Now, dividing by $\mu\Delta t(b-a)$ and taking the limit as $b-a$ goes to zero we get the ***heat equation***:

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2}, \tag{1.17}$$

where $k = \frac{K}{c\rho} > 0$ is the ***diffusion constant***.

The heat equation has two independent variables x and t and a single dependent variable u. It is an equation in the derivatives of the dependent variable with respect to the independent variables. A multidimensional (in space) analogue of the heat equation is imaginable, and takes the form

$$\frac{\partial u}{\partial t} = k\left(\frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} \right).$$

The operator in the right-hand side is of independent interest, and is known as the ***Laplacian*** of u and given by

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}.$$

With this bit of notation, the heat equation can be written as

$$\frac{\partial u}{\partial t} = k\Delta u.$$

We shall subsequently look at the heat equation in some detail, and shall say some things about the behaviour of its solutions at that time.

1.1.12 Waves in a taut string

Next we consider the small transverse vibrations of a taut string when it is plucked, e.g., a guitar string. To derive the equations governing these transverse vibrations, we use simple force balance on a short segment of the string. In Figure 1.10 we depict a little segment of a string with its transverse displacement

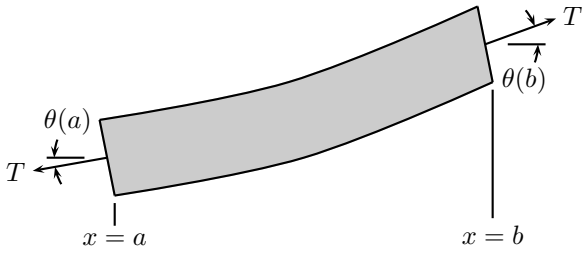


Figure 1.10 A segment of string used in the derivation of the wave equation

denoted u. It is assumed that the tension T in the string is independent of x and t. This is acceptable for small string deflections. The vertical component of the force on the string is given by

$$F_y = -T\sin(\theta(a)) + T\sin(\theta(b)).$$

Let us manipulate this until it looks like something we want. We denote the vertical deflection of the string by u. We then have

$$\tan\theta(a) = \frac{\partial u}{\partial x}\Big|_a, \quad \tan\theta(b) = \frac{\partial u}{\partial x}\Big|_b.$$

Now recall that for small angles θ we have $\sin\theta \approx \tan\theta$. This then gives

$$F_y \approx T\left(\frac{\partial u}{\partial x}\Big|_b - \frac{\partial u}{\partial x}\Big|_a\right).$$

Now the mass of the segment of string is $\rho(b-a)$ with ρ the length mass density of the string, which we assume to be constant. The vertical acceleration is then $\frac{\partial^2 u}{\partial t^2}$, which we suppose to be constant in the segment. By making the length of the segment sufficiently small, this becomes closer to being true. An application of force balance now gives

$$\rho(b-a)\frac{\partial^2 u}{\partial t^2} \approx T\left(\frac{\partial u}{\partial x}\Big|_b - \frac{\partial u}{\partial x}\Big|_a\right).$$

Dividing by $\rho(b-a)$ and letting $b-a$ go to zero, we have the ***wave equation***:

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, \tag{1.18}$$

where $c = \sqrt{\frac{T}{\rho}} > 0$ is the ***wave speed*** for the problem.

There are two independent variables x and t for the wave equation, and a single dependent variable u. The equation itself is one involving derivatives of the

dependent variable with respect to the independent variables. As with the heat equation, a multidimensional (in space) analogue of the heat equation is possible, and takes the form

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} \right).$$

The operator in the right-hand side is the Laplacian which we saw with the heat equation:

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}.$$

The wave equation can be thus written as

$$\frac{\partial^2 u}{\partial t^2} = k \Delta u.$$

In the text we shall examine the wave equation in a little detail, and say some things about the behaviour of its solutions.

1.1.13 The potential equation in electromagnetism and fluid mechanics

In this section we shall see how the Laplacian, introduced in our discussion of the wave equation, arises in special cases of Maxwell's and Navier–Stokes' equations.

We first consider Maxwell's equations of electromagnetism. We make a few assumptions about the physics that will allow us to simplify the complicated Maxwell's equations.

1. We assume we are in steady-state, so the dependent variable do not depend on time.
2. We assume that the electric field $\boldsymbol{E}$ is a potential field. This means that there exists a function V, called the ***electric potential***, such that $\boldsymbol{E} = \nabla V = (\frac{\partial V}{\partial x_1}, \frac{\partial V}{\partial x_2}, \frac{\partial V}{\partial x_3})$.
3. We assume that we are in free space so the charge density is zero.

The equations for the potential function are determined by Gauss's law:

$$\nabla \cdot \boldsymbol{E} = 0 \quad \Longrightarrow \quad \nabla \cdot \nabla V = 0.$$

A direct computation gives

$$\nabla \cdot \nabla V = \Delta V = \frac{\partial^2 V}{\partial x_1^2} + \frac{\partial^2 V}{\partial x_2^2} + \frac{\partial^2 V}{\partial x_3^2}. \tag{1.19}$$

This is the ***potential equation*** in $\mathbb{R}^3$.

Next we turn to a special case of the Navier–Stokes equations, making the following physical assumptions.

1. The flow in inviscid, so the viscosity μ vanishes.

2. The flow in incompressible, so the divergence of the fluid velocity vanishes.
3. We assume the fluid velocity is derived from a velocity potential: $\boldsymbol{u} = -\nabla\phi$.
4. We suppose that body forces are potential forces, i.e., $\boldsymbol{f} = -\nabla V$, e.g., gravitational forces.

In this case, the assumptions of incompressibility and the existence of a velocity potential give the following form of the equation of continuity:

$$\nabla \cdot \boldsymbol{u} = 0 \quad \Longrightarrow \quad \Delta\phi = 0.$$

Let us investigate the impact of this, along with the other physical assumptions, in describing properties of the fluid flow. First of all, a direct computation gives

$$\boldsymbol{u} \cdot \nabla\boldsymbol{u} = (\nabla \times \boldsymbol{u}) \times \boldsymbol{u} + \operatorname{grad}(\tfrac{1}{2}\boldsymbol{u} \cdot \boldsymbol{u}),$$

where $\boldsymbol{a} \times \boldsymbol{b}$ denotes the vector cross-product and $\boldsymbol{a} \cdot \boldsymbol{b}$ denotes the Euclidean inner product of $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^3$. Since $\boldsymbol{u} = -\nabla\phi$, we calculate that $\nabla \times \boldsymbol{u} = \boldsymbol{0}$, and so the Navier—Stokes equations read

$$\nabla\left(\frac{\partial\phi}{\partial t} + \frac{1}{2}(\boldsymbol{u} \cdot \boldsymbol{u}) + \frac{p}{\rho} + V\right) = 0.$$

This implies that

$$\frac{\partial\phi}{\partial t} + \frac{1}{2}(\boldsymbol{u} \cdot \boldsymbol{u}) + \frac{p}{\rho} + V$$

depends only on t. This is known as ***Bernoulli's principle***.

Let us indicate another way in which the Laplacian arises in fluid flow problems, in this case with planar flow problems, i.e., that $u_3 = 0$. We assume that the fluid velocity $(u_1, u_2, 0)$ has the special form

$$u_1 = \frac{\partial\psi}{\partial x_2}, \quad u_2 = -\frac{\partial\psi}{\partial x_1}$$

for a function ψ of (x_1, x_2) called the ***stream function***. Note that the resulting fluid velocity automatically satisfies the incompressible continuity equation:

$$\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} = \frac{\partial^2\psi}{\partial x_1\partial x_2} - \frac{\partial^2\psi}{\partial x_2\partial x_1} = 0.$$

If we additionally require that $\Delta\psi = 0$, then $\nabla \times \boldsymbol{u} = \boldsymbol{0}$. In this case, we recall from vector calculus that $\boldsymbol{u} = -\operatorname{grad}\phi$, i.e., the flow is a potential flow.

1.1.14 Einstein's field equations

In Einstein's theory of general relativity, a ***spacetime*** is a four-dimensional "differentiable manifold." This means that around every point in spacetime there is a parameterisation by $\mathbb{R}^4$. To keep things simple (and still representative), we just assume that our spacetime is equal to $\mathbb{R}^4$. There are two physical objects defined on spacetime, and Einstein's field equations relate these. The first is the ***stress-energy tensor*** $\boldsymbol{T}$ which is a symmetric 4×4 matrix function. This encodes the properties of spacetime like mass and electromagnetic fields. The other object defined on spacetime of interest is the ***metric tensor*** $\boldsymbol{g}$, which is another symmetric 4×4 matrix function, this one having the property that it has one negative and three positive eigenvalues. Physically, $\boldsymbol{g}$ determines the gravitational properties of spacetime, as well as the space and time structure.

We definitely will not derive Einstein's field equations, but will simply produce them. First of all, we denote the coordinates for spacetime by (x^1, x^2, x^3, x^4); the use of superscripts as indices is traditional in general relativity. The components of the matrices $\boldsymbol{T}$ and $\boldsymbol{g}$ we denote by T^{jk} and g_{jk}, $j, k \in \{1, 2, 3, 4\}$. First we define the ***Christoffel symbols*** associated with $\boldsymbol{g}$:

$$\gamma^j_{kl} = \frac{1}{2} \sum_{m=1}^{4} g^{jm} \left(\frac{\partial g_{mk}}{\partial x^l} + \frac{\partial g_{ml}}{\partial x^k} - \frac{\partial g_{kl}}{\partial x^m} \right),$$

where g^{jk}, $j, k \in \{1, 2, 3, 4\}$, are the components of $\boldsymbol{g}^{-1}$. Next, the ***curvature tensor*** is then defined by

$$R^j_{klm} = \frac{\partial \Gamma^j_{lm}}{\partial x^k} - \frac{\partial \Gamma^j_{km}}{\partial x^l} + \Gamma^j_{km} \Gamma^m_{lm} - \Gamma^j_{lm} \Gamma^m_{km},$$

the ***Ricci tensor*** is the 4×4-symmetric matrix function **Ric** defined by

$$\mathrm{Ric}_{jk} = \sum_{l=1}^{4} R^l_{ljk}, \qquad j, k \in \{1, 2, 3, 4\},$$

and the ***scalar curvature*** is function defined by

$$\rho = \sum_{j,k=1}^{4} g^{jk} \, \mathrm{Ric}_{jk} \, .$$

Finally, we define the contravariant form of the stress-energy tensor, which is the symmetric 4×4-matrix function $\overline{\boldsymbol{T}}$ with components

$$\overline{T}_{jk} = \sum_{l,m=1}^{4} g_{jl} g_{km} T^{lm}, \qquad j, k \in \{1, 2, 3, 4\}.$$

With all of this data, we can now write the ***Einstein field equations***:

$$\mathbf{Ric} - \tfrac{1}{2}\rho \boldsymbol{g} + \Lambda \boldsymbol{g} = \frac{8\pi G}{c^4}\overline{\boldsymbol{T}}, \tag{1.20}$$

where Λ is the ***cosmological constant***, G is the ***gravitational constant***, and c is the speed of light in a vacuum.

There are four independent variables in Einstein's field equations, the coordinates (x^1, x^2, x^3, x^4) for spacetime. There are nominally ten dependent variables (the sixteen components of $\boldsymbol{g}$ taking into account symmetry). The equations are complicated equations in the derivatives of dependent variables with respect to the independent variables.

Of course, we will not say anything about the nature of the solutions to Einstein's field equations. This is the subject of deep work by many smart people.

1.1.15 The Schrödinger equation

In quantum mechanics, the Schrödinger equation governs the behaviour of a function known as the ***wave function***. The wave function encodes the state of a quantum system in the form of a "probability amplitude." These are typically complex-valued as they come equipped with, not just an amplitude, but a phase. This phase allows for the wave part of the particle/wave duality seen in the behaviour of subatomic particles. We shall not delve into the quantum mechanical machinations required to understand where the equation comes from, but shall merely produce the Schrödinger equation for the wave function ψ of a single particle moving in $\mathbb{R}^3$ in an electric field with electric potential function V:

$$\mathrm{i}\hbar\frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2\mu}\Delta\psi + V\psi, \tag{1.21}$$

where $\mathrm{i} = \sqrt{-1}$, $\hbar$ is Planck's constant, and μ is the effective mass of the particle.

Note that the Schrödinger equation is an equation with four independent variables, (x_1, x_2, x_3) and t, and a single complex-valued dependent variable ψ, or equivalently, regarding a complex number as determined by its real and imaginary parts, two real dependent variables. Of course, the equation is one involving the derivatives of the dependent variable with respect to the independent variables.

1.1.16 The Black–Scholes equation

The model we arrive at in this section is widely used in options trading, and has garnered a Nobel Prize in Economics for its developers. It is also true that the widespread misuse of this model, and models like it, combined with greed and governments divesting themselves of regulatory responsibilities, has led to the ruination of millions of lives. So mathematics *can* make a difference in peoples lives!

The equation we give provides the price V of an option as a function of stock price S and time t. It also has the following parameters:

r risk-free compound interest rate
σ standard deviation of stock's returns

We shall not describe the "derivation" of the model, but simply state the ***Black–Scholes equation***:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0.$$

For this equation, there are two independent variables (t, S) and a single independent variable V. The equation involves derivatives of the independent variable with respect to the dependent variables.

Now you can go off into a room and run Black–Scholes simulations, and make yourself rich!

1.1.17 Fibonacci numbers and rabbits

A completely unrealistic model for a rabbit population assumes that (1) rabbits appear in opposite sex pairs, (2) a pair of opposite sex rabbits gives birth to a pair of opposite sex rabbits at the end of every time period, and (3) rabbits never die. If f_n is the number of opposite sex pairs at the end of the nth time period, then we have

$$f_{n+1} = f_n + f_{n-1}, \qquad n \in \mathbb{Z}_{>0}.$$

If we specify $f_0 = 0$ and $f_1 = 1$, then the sequence $0, 1, 1, 2, 3, 5, 8, \ldots$ is the ***Fibonacci sequence***.

1.1.18 Bank balance model

Suppose that one has a bank account that earns annual interest $\alpha \in \mathbb{R}_{>0}$. At the end of each year, the account owner withdraws an amount w that might depend on time (perhaps it increases each year) and on the account balance. Let us denote the "time" here by $n \in \mathbb{Z}_{\geq 0}$, with $n = 0$ being the opening of the account with a balance x_0. Subsequence balances are then computed by

$$x_{n+1} = (1 + \alpha)x_n - w(n, x_n), \qquad n \in \mathbb{Z}_{\geq 0}. \tag{1.22}$$

A question that might arise is then whether the interest rate α, the withdrawal rule w, and the initial balance x_0 leads to growth or depletion of the account.

1.1.19 Keynesian national income model

Consider the following variables:

Y national income

C consumer expenditure
I private investment in equipment
G government expenditure.

If "time" is a counter $n \in \mathbb{Z}_{\geq 0}$ indicating a regular time interval, say a year, then we can measure these quantities at the end of each time interval and then we have

$$Y_n = C_n + I_n + G_n.$$

Now these quantities are interrelated, and these interrelations give the behaviour of the model. For example, maybe consumer expenditure is proportional to the previous national income,

$$C_{n+1} = \alpha Y_n.$$

Maybe capital investment is proportional to the increase of consumer expenditure from one year to another,

$$I_{n+1} = \beta(C_{n+1} - C_n).$$

Finally, we may assume that government expenditure is constant, normalised to be $G_n = 1$. Assembling this all together gives

$$Y_{n+2} - \alpha(1 + \beta)Y_{n+1} - Y_n = 1. \tag{1.23}$$

One can wonder whether, in such a model, the national income grows or shrinks.

1.1.20 A discrete model for heat flow

Suppose that we have a rod of infinite length whose temperature we measure at time intervals Δ_{time} and spatial intervals Δ_{space}. Thus we record temperatures

$$u(m\Delta_{\text{time}}, n\Delta_{\text{space}}), \qquad m \in \mathbb{Z}_{\geq 0},\ n \in \mathbb{Z}.$$

Heat will flow from position $(n-1)\Delta_{\text{space}}$ to position $n\Delta_{\text{space}}$ and from position $n\Delta_{\text{space}}$ to position $(n+1)\Delta_{\text{space}}$. The discrete analogue to Fourier's Law gives the temperature increase at position $n\Delta_{\text{space}}$ from time $m\Delta_{\text{time}}$ to time $(m+1)\Delta_{\text{time}}$ as

$$\begin{aligned}
&u((m+1)\Delta_{\text{time}}, n\Delta_{\text{space}}) - u(m\Delta_{\text{time}}, n\Delta_{\text{space}}) \\
&\qquad = k\Big(u(m\Delta_{\text{time}}, (n-1)\Delta_{\text{space}}) - u(m\Delta_{\text{time}}, n\Delta_{\text{space}})\Big) \\
&\qquad\quad - k\Big(u(m\Delta_{\text{time}}, n\Delta_{\text{space}}) - u(m\Delta_{\text{time}}, (n-1)\Delta_{\text{space}})\Big) \\
&\qquad = k\Big(u(m\Delta_{\text{time}}, (n-1)\Delta_{\text{space}}) - 2u(m\Delta_{\text{time}}, n\Delta_{\text{space}}) + u(m\Delta_{\text{time}}, (n-1)\Delta_{\text{space}})\Big),
\end{aligned}$$

for a constant $k \in \mathbb{R}_{>0}$.

1.1.21 Summary

In this section we have presented myriad illustrations of how equations involving various numbers of independent and dependent variables, along with derivatives or iterates of these, may arise in applications. The subject of this volume is how to solve some such equations, and how to look for the essential attributes of equations such as these and of system models described by these equations. These are the subjects of "differential equations" and "difference equations." These are subjects that are impossible to comprehend fully in any sort of generality, which is not unreasonable since differential and difference equations describe physical phenomenon that we do not expect to be able to understand fully. Thus the subject of differential and difference equations is a combination of looking deeply at certain special cases (particularly linear equations) and working hard to determine characteristic behaviour of general classes of systems.

1.1.22 Notes

[Brown 2007, page 68]

Exercises

1.1.1 Think of, or Google, three models (not included in the text) where differential or difference equations arise in practice. In each case, do the following:

(a) indicate the independent and dependent variables;

(b) give some meaning to these variable in terms of the particular application;

(c) provide a tiny bit of background about where the equations come from.

Section 1.2

System thinking

In the preceding section we saw a multitude of examples from a variety of areas. These examples as presented fall into the general area of "dynamical systems," in the sense that we have models and we want to understand the behaviour of the models. In this section we shall reconsider some of these examples from a systems point of view, in order to illustrate some of the questions that can arise in the general subject of "system theory." In system theory, one typically has the important additional concepts of "input" which allows one to affect the behaviour of the model. This is an important way in which system theory differs from dynamical system theory.

The discussion here will be highly abbreviated, both in terms of theory and in terms of the details of the particular applications. Some of the theoretical problems are considered in detail subsequently in this volume.

1.2.1 Mass-spring-damper systems

Let us first think about the simplified car suspension model depicted in Figure 1.1. We can suppose that the mass is subject to a vertical force F. In this case, one is interested in the response of the system; namely, given values for the physical constants m, k, and d, how does the system respond? Specific questions might include: what is the amplitude of the oscillations for a given input? how long does it take for oscillations to damp out? what is the maximum acceleration experienced by the mass? These sorts of questions are related to the behaviour of the system that transfers the input F to the response y. In terms of design, one is interested in selecting parameters k and d so that the input/response behaviour satisfies certain criterion. This exact system is examined in some detail in Example 4.3.20.

For the coupled masses depicted in Figure 1.2, one might consider various inputs to the system. For example, one might consider applying a force F_1 applied to mass m_1, a force F_2 applied to mass m_2, or an application of both forces. One might also consider the situation where $F_2 = \alpha F_1$ for a constant α. Questions that might arise are: can one select the inputs so as to have the response of the system behave in a certain way? are some combinations of inputs more effective than others at achieving this objective? Here we see questions of a somewhat different character than for the car suspension model. Namely, we are interested in designing the inputs so as to achieve a desired response.

1.2.2 RLC circuits

The RLC circuits discussed in Section 1.1.4 have associated with them a variety of system theoretic questions that are electrical analogues of the mechanical

questions for the mass-spring-damper problems from the preceding section. Here the response is perhaps the current in the circuit or the charge in a capacitor, and the input in a voltage or current source at some point in the circuit. One can then consider questions of how the physical constants, i.e., values for resistance, capacitance, or inductance, affect the response of the system for a given input. One can also think about how to design inputs to achieve desired behaviour of the response.

1.2.3 Tank systems

For the tank system depicted in Figure 1.6, one can ask questions similar to those for mass-spring-damper systems and RLC circuits. For example, one can regard the input flow F_{in} as being fixed, and then examine how the physical constants A_1, A_2, a_1, and a_2 affect the fluid levels in each of the tanks. Alternatively, one can think about designing the input flow F_{in} in such a way that the fluid levels in the tanks behave as desired.

1.2.4 Population models

The most important features of the population models of Section 1.1.6 are the equilibria for the systems and their stability. Equilibria are to be thought of as states where there is a balance between the various factors that affect a population. The stability of an equilibrium reflects whether the balance is maintained when there are small changes in the populations.

1.2.5 Euler–Lagrange equations

In Section 1.1.8 we considered a sort of problem that is important in system theory, that of optimisation. While it is important, it is not a subject to which we will devote any substantial attention in this volume. We give a brief overview of how to formulate optimality problems as so-called "goal-seeking systems" in Section 2.3.1.

1.2.6 Heat flow due to temperature gradients

In Section 1.1.11 we considered the temperature distribution in a rod. Here there are system theoretic problems that are of the "analysis" type and of the "design" type, as we have seen above. A typical "analysis" problem is a description of the temperature distribution as time gets large. A typical design problem might be the specification of boundary and/or initial conditions to give a desired temperature distribution.

We note that the particular model for the heat equation (and the wave and potential equations) are not of the sort to which we will devote attention in this volume. This is because these are instances of so-called "infinite-dimensional systems." They are infinite-dimensional because the temperature distribution (or the string displacement, or the charge distribution) is, at a given instant of time, a

function and so will typically live in some infinite-dimensional space of functions, such as considered in Sections IV-1.2 and IV-1.3.

1.2.7 The Black–Scholes equation

The Black–Scholes equation is actually used in the financial tool to assist in making decisions regarding stocks. A simplified description of this sort of activity would be that one considers how parameter changes in the Black–Scholes model, based on empirical observations, leads to decision-making strategies.

1.2.8 Bank balance model

The simple bank balance model of (1.22) can be thought of with the withdrawal strategy w as being fixed, or as being an input to be designed. In the former situation, one wishes to examine the effects of the withdrawal strategy on the long-term balance in the account, whereas, in the latter situation, one wishes to determine a strategy that leads to a desired behaviour as the bank balance, e.g., maximising it at retirement.

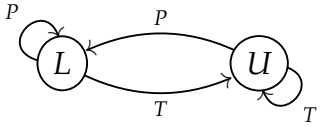
1.2.9 Keynesian national income model

For the simplified national income model presented in (1.23), one can again think of the parameters of the model as being fixed or as being designed. Also, in the derivation of this model, various assumptions were made about the model, and these may be changed in order to examine their effects on the national income.

1.2.10 A token-operated turnstile

All of the examples presented in Section 1.1 are modelled by differential or difference equations. However, it is not the case that all natural examples of systems are modelled in this way. To illustrate this, we consider an example of a system that is modelled by a deterministic finite state automaton, which we consider in a general setting in Example 2.2.11–2.

Consider a token-operator turnstile that has an arm that blocks access, opening when a token is inserted in the machine. Such a machine has two states, "locked" (L) and "unlocked" (U). The turnstile has two possible inputs, "insert token" (T) and "push to open" (P). The default state is L. An input of T will change the state to U, while an input of P will not change the state from L. If the state is U and an input of P is given, the state changes to L. An input of T while in state U will not change the state. This simple process is summarised by the diagram



1.2.11 Image transmission

Suppose that we wish to transmit an image over a communication channel. A general schematic for this is shown in Figure 1.11, and we see that, abstractly, one

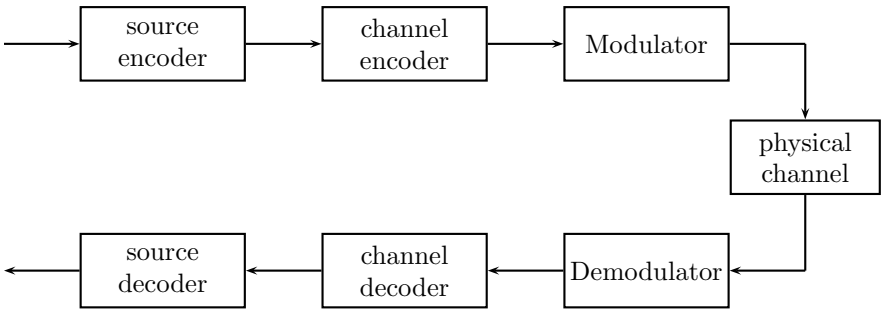


Figure 1.11 A communication channel

transmits a signal through various stages in the transmission process. If one has an image, then one must convert this into a format that can be converted into a transmittable signal that can be decoded by the receiver. There are many ways to convert an image file into a transmittable signal, so let us just illustrate this abstractly in a special case. Consider the greyscale image shown in Figure 1.12.[4] This image is 256×256 pixels, and each pixel is assigned a number in the set

Figure 1.12 A greyscale image

$\{0, 1, \ldots, 255\}$, with 0 corresponding to white and 255 to black. Thus the image can be thought of as a map $f\colon \{0, 1, \ldots, 255\}^2 \to \{0, 1, \ldots, 255\}$. One way to think of this is as a $\mathbb{R}$-valued signal defined on a two-dimensional discrete domain.

[4] Image downloaded from Waterloo BragZone, `http://links.uwaterloo.ca/bragzone.base.html`.

Exercises

1.2.1 Think of, or Google, three instances (not included in the text) of instances where "systems thinking" of the sort illustrated in this section arise. In each case, do the following:

(a) identify inputs and outputs;

(b) indicate whether there is an "analysis" or a "design" problem;

(c) provide some context about why the system theoretic problems are interesting and/or useful.

Section 1.3

Notes

1.3.1 Mechanics

1.3.2 Fluid mechanics

[Newton 1687]

Chapter 2

General classes of systems and their properties

We shall, for the most part, focus our attention on specific sorts of systems, typically systems described by differential and difference equations. However, it is sometimes useful to give a lower resolution view of the subject, since by doing so one can separate out specifically system-theoretic concepts from other concepts particular to the certain classes of systems. Thus, in this chapter we present a "general" theory of systems that will include as special cases all of the systems we shall subsequently consider in detail.

The objectives of the presentation is to define in a general setting important system-theoretic notions such as input, output, state, linearity, causality, time dependence, interconnection, etc. The generality will allow us to represent these important concepts in a framework that is free from the baggage of structure that is not required for these concepts to make sense. We shall also carefully describe the additional assumptions that must be placed on a general system framework to arrive at the special classes of systems to which we shall devote the greatest attention: linear systems.

Do I need to read this chapter? For a reader whose interest is in the standard theory of linear systems—and the development of such systems is the primary objective of this volume—this chapter might be regarded as optional. However, we believe that a general and abstract framework that gives *an* (not *the*) answer to the question, "What is a system?" is useful and interesting. •

Contents

Section 2.1

Abstract formulations of systems

In this section we introduce the basic definition of what we shall mean by a "general system," and flesh out some consequences of this definition. We shall introduce as special cases of this definition a few examples that we shall only explore in detail in subsequence chapters.

2.1.1 General systems

Our most general notion of system is the following.

2.1.1 Definition (General system) A ***general system*** is a pair $\Sigma = (\mathscr{V}, \mathscr{B})$ where

(i) $\mathscr{V} = (V_i)_{i \in I}$ is a family of sets (for $i \in I$, the set V_i is an ***object*** of Σ) and

(ii) $\mathscr{B} \subseteq \prod_{i \in I} V_i$ (the ***behaviours*** of the system). •

The way in which one should think about a general system is this. Each of the sets V_i, $i \in I$, represents some component of the system. A behaviour, by definition, is a mapping $\beta \colon I \to \cup_{i \in I} V_i$ with the property that $\beta(i) \in V_i$, $i \in I$ (Definition I-1.6.7). Given a behaviour β, $\beta(i)$ represents the way in which the component V_i contributes to that behaviour.

We shall not really work at length with the preceding very general notion of a system. However, to illustrate how it works, it is worth looking at a few examples. These examples are chosen for their simplicity, and for their ability to represent the general definitions.

2.1.2 Examples (General systems)

1. Let us consider a particle of mass m falling under the influence of a gravitational force determined by the gravitational acceleration a_g. The objects in the system are the mass $m \in \mathbb{R}_{>0}$, the gravitational acceleration $a_g \in \mathbb{R}_{>0}$, and the twice continuously differentiable path $\xi \colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ followed by the particle for nonnegative times. Thus

$$\mathscr{V} = \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times C^2(\mathbb{R}_{\geq 0}; \mathbb{R}).$$

A behaviour is a selection $(m, a_g, \xi) \in \mathscr{V}$ of an element of each of the three objects of the system. Of course, a behaviour is not arbitrary, but involves a relationship between the objects of the system; this is what makes the concepts of a system have content. And this is where the specifics of the system enter into the description, i.e., this is where the modelling considerations of Chapter 1 enters the picture. The physics, i.e., force balance, mandates that

$$\mathscr{B} = \{(m, a_g, \xi) \in \mathscr{V} \mid m\ddot{\xi}(t) = -ma_g\},$$

if gravity acts in the opposite direction of increasing ξ. Let us make a few observations about this set of behaviours:

(a) The behaviour is actually independent of the object m. This is fine.

(b) The behaviours place restrictions on $\xi \in C^2(\mathbb{R}_{\geq 0}; \mathbb{R})$. Thus the object $C^2(\mathbb{R}_{\geq 0}; \mathbb{R})$ is far larger than is needed to capture the behaviours of the system. For example, we could replace this object with $C^\infty(\mathbb{R}_{\geq 0}; \mathbb{R})$, the infinitely differentiable mappings from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}$. But even this is larger than is required. Sometimes the precise specification of the objects is important. For example, one could also make the set of behaviours empty by over-prescribing the set of functions from $C^2(\mathbb{R}_{\geq 0}; \mathbb{R})$ to which ξ belongs. For example, if we require that $\xi(t) = \mathrm{e}^t$, there will be no behaviours that satisfy the physical model.

(c) We have prescribed the time domain $\mathbb{R}_{\geq 0}$ on which paths for the particle are defined. One could change this time interval, or one could allow paths with varying time intervals.

(d) Similarly, the codomain $\mathbb{R}$ for the paths of the particle is too large. Thus the system description contains behaviours one will never observe in practice.

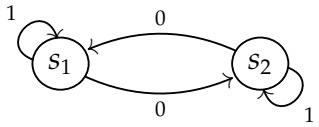
(e) For each $(m, a_g) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, there are many $\xi \in C^2(\mathbb{R}_{\geq 0}; \mathbb{R})$ for which $(m, a_g, \xi) \in \mathscr{B}$. Again, this is to be expected, and is something we shall subsequently deal with in a systematic way.

The objective with this sort system is to understand the behaviour of the behaviours. In this case this is easily done, but generally this might be difficult. This sort of system described by a linear ordinary differential equation is one we shall study in detail in the sequel.

2. Let $Q = \{s_1, s_2\}$, let $\Lambda = \{0, 1\}$, and define $\delta \colon Q \times \Lambda \to Q$ according to the following table:

	0	1
s_1	s_2	s_1
s_2	s_1	s_2

We call the elements of Q ***states*** and the elements of Λ ***letters*** (thus Λ is the ***alphabet***). Note that $\delta(s, 1) = s$ and $\delta(s, 0) \neq s$; thus a 0 changes the state and a 1 leaves the state unchanged. Thus, alongside the table, we can represent δ by the diagram



(2.1)

Define W to be the set of finite sequences of 0's and 1's. We thus write an element $\boldsymbol{w} \in W$ as $\boldsymbol{w} = w_1 w_2 \cdots w_k$ for some $k \in \mathbb{Z}_{>0}$, and where $w_j \in A$ for $j \in \{1, \ldots, k\}$. An element of W is called a ***word***. Given a word $\boldsymbol{w} = w_1 w_2 \cdots w_k \in W$, there

exists $q_0, q_1, \ldots, q_k \in Q$ such that, recursively,

$$q_{j+1} = \delta(q_j, w_{j+1}), \qquad j \in \{0, 1, \ldots, k-1\}.$$

We call q_k the ***terminal state*** of the word $\boldsymbol{w}$. Note that, if $q_0 = s_1$, then q_j, $j \in \{0, 1, \ldots, k\}$, measures the evenness (when $q_j = s_1$) or oddness (when $q_j = s_2$) of the number of zeros in $w_1 w_2 \cdots w_j$, $j \in \{1, \ldots, k\}$.
The system is then $\mathscr{V} = W \times Q$ and $\mathscr{B} \subseteq W \times Q$ is the set of pairs $(\boldsymbol{w}, q)$ where $\boldsymbol{w}$ is a word and q is the terminal state of $\boldsymbol{w}$.
This system (including the final condition that $q_k = s_1$) is an example of what is known as a deterministic finite state automaton. Turing machines in the theory of computation are variations of deterministic finite state automatons. One can think of this as a system that accepts certain inputs; in the case described above, it would accept strings of 0's and 1's with an even number of 0's. While interesting, we shall not work with these systems in detail in this volume.

3. Let us consider a pair of simple digital logic systems.
Let $n \in \mathbb{Z}_{>0}$. Suppose we have a device that has 2^n input channels, labelled $\{i_0, i_1 \ldots, i_{2^n-1}\}$, and each channel receives an input from the set $\{0, 1\}$. The device also has n output channels, labelled $\{o_0, o_1, \ldots, o_{n-1}\}$, that each return an output from the set $\{0, 1\}$. A ***one hot*** input to the 2^n input channels means that at most one of the 2^n channels is a 1, and all others are 0. A ***binary encoder*** takes a one hot input and returns n outputs according to the rule that, if the input channel i_k receives the 1, then the output $o_j \in \{0, 1\}$, $j \in \{0, 1, \ldots, o_{n-1}\}$ satisfies

$$k = \sum_{j=0}^{n-1} o_j 2^j;$$

that is to say, $o_0 o_1 \cdots o_{n-1}$ is the binary representation of k. If $i_0 = i_1 = \cdots = i_{2^n-1} = 0$, then there is no output. We can regard a binary encoder as a system by

$$\mathscr{B}_{\text{enc}} = \left\{ (\boldsymbol{i}, \boldsymbol{o}) \in \{0,1\}^{\{0,1,\ldots,2^n-1\}} \times \{0,1\}^{\{0,1,\ldots,n-1\}} \;\middle|\; \|\boldsymbol{i}\| = 1, \sum_{j=0}^{n-1} \boldsymbol{o}(j) 2^j = k, \; \boldsymbol{i}(k) = 1 \right\}.$$

A ***binary decoder*** undoes this operation. Thus it has n input channels $\{I_0, I_1, \ldots, I_n\}$ and 2^n output channels $\{O_0, O_1, \ldots, O_{2^n-1}\}$, with each input channel receiving an element from $\{0, 1\}$ and each output channel receiving an element from $\{0, 1\}$. The rule for producing an output is that, if

$$\sum_{j=0}^{n-1} I_j 2^j = k \in \{0, 1, \ldots, 2^n - 1\},$$

output O_k is set to 1 and all other outputs are zero. Thus the output is one hot. The system in this case is

$$\mathscr{B}_{\text{dec}} = \left\{ (\boldsymbol{I}, \boldsymbol{O}) \in \{0,1\}^{0,1,\dots,n-1\}} \times \{0,1\}^{\{0,1,\dots,2^n-1\}} \;\middle|\; \|\boldsymbol{O}\| = 1,\ \boldsymbol{O}\left(\sum_{j=0}^{n-1} \boldsymbol{I}(j) 2^j \right) = 1 \right\}$$

These can each be clearly regarded as inverses of one another. One way one can use these is as a means of converting input from a keyboard with 2^n keys input into a sequence of n 0's and 1's. Crucial to this is that keyboard input is naturally one hot since only one key at a time can be depressed.

4. Suppose we perform a set of experiments where we control quantities $c_1, \dots, c_k$ and measure quantities $m_1, \dots, m_r$. Each of these quantities we suppose to take values in $\mathbb{R}_{>0}$. Thus we have $\mathscr{V} = (\mathbb{R}_{>0})^k \times (\mathbb{R}_{>0})^r$. The experiment gives us a system $\mathscr{B} \subseteq \mathscr{V}$ comprised of the data from each measurement. •

2.1.2 General input/output systems

A special class of system is that which involves an explicit identification of what one calls "inputs" and what one calls "outputs." The definition is the following.

2.1.3 Definition (General input/output system) A ***general input/output system*** is a triple $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$, where

(i) $\mathscr{U}$ is a set (the set of ***inputs***),

(ii) $\mathscr{Y}$ is a set (the set of ***outputs***), and

(iii) $\mathscr{B} \subseteq \mathscr{U} \times \mathscr{Y}$ (the ***behaviours*** of the system).

The set

$$\text{dom}(\Sigma) = \{\mu \in \mathscr{U} \mid \text{there exists } \eta \in \mathscr{Y} \text{ with } (\mu, \eta) \in \mathscr{B}\}$$

is the ***domain*** of Σ and the set

$$\text{rng}(\Sigma) = \{\eta \in \mathscr{Y} \mid \text{there exists } \mu \in \mathscr{U} \text{ with } (\mu, \eta) \in \mathscr{B}\}$$

is the ***range*** of Σ. Given $\mu \in \mathscr{U}$ we denote

$$\mathscr{B}(\mu) = \{\eta \in \mathscr{Y} \mid (\mu, \eta) \in \mathscr{B}\}.$$ •

In a very abstract sense, this notion of "input/output system" is easy to understand: given a pair $(\mu, \eta) \in \mathscr{B}$, one should think of $\mu \in \mathscr{U}$ as being data that is input to the system and $\eta \in \mathscr{Y}$ as being a possible outcome. Some readers may find it insightful to recall from Definition I-1.2.1 that $\mathscr{B} \subseteq \mathscr{U} \times \mathscr{Y}$ is what we have called a "relation" from $\mathscr{U}$ to $\mathscr{Y}$. Note that, generally, there may be more than one output η for a given input μ. We shall deal with this in a systematic way in the sequel. However, sometimes one does have a well-defined mapping from inputs to outputs, and the following definition captures this, keeping in mind that a mapping between sets is a particular example of a relation (Definition I-1.3.1).

2.1.4 Definition (Functional input/output system) A general input/output system $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$ is a ***functional input/output system*** if there exists $F_\Sigma \colon \mathscr{U} \to \mathscr{Y}$ such that $\mathscr{B} = \operatorname{graph}(F_\Sigma)$. •

Let us consider some general input/output systems.

2.1.5 Examples (General input/output systems)

1. For a general system $\Sigma = (\mathscr{V}, \mathscr{B})$ with $\mathscr{V} = (V_i)_{i \in I}$, it is sometimes natural to partition the index set I as $I = I_i \cup I_o$ with $I_i \cap I_o$, and then take $\mathscr{U} = \prod_{i \in I_i} V_i$ and $\mathscr{Y} = \prod_{i \in I_o} V_i$.
2. For the particle in a gravitational field of Example 2.1.2–1, we take the inputs to be the mass m of the particle and the gravitational acceleration a_g. Thus we take $\mathscr{U} = \mathbb{R}_{>0} \times \mathbb{R}_{>0}$. As output we take $\mathscr{Y} = C^2(\mathbb{R}_{\geq 0}; \mathbb{R})$. In this case we see that there are many possible outputs corresponding to a given output. Thus this is not a functional input/output system.
3. We can consider the deterministic finite state automaton of Example 2.1.2–2 as a general input/output system. To do so, we take $\mathscr{U} = W$ and $\mathscr{Y} = Q$. Thus inputs are words formed of 0's and 1's and outputs are the terminal states for the corresponding word. We note that this is a functional input/output system since the input uniquely determines the output.
4. We consider the binary encoder and decoder of Example 2.1.2–3. These have the natural structure of general input/output systems by taking

$$\mathscr{U}_{\text{enc}} = \{0,1\}^{\{0,1,\dots,2^n-1\}}, \quad \mathscr{Y}_{\text{enc}} = \{0,1\}^{\{0,1,\dots,n\}}$$

and

$$\mathscr{U}_{\text{dec}} = \{0,1\}^{\{0,1,\dots,n\}}, \quad \mathscr{Y}_{\text{dec}} = \{0,1\}^{\{0,1,\dots,2^n-1\}}.$$

5. For the experimental data system of Example 2.1.2–4, it is natural to take $\mathscr{U}$ to be the set $(\mathbb{R}_{>0})^k$ where the controlled variables reside, while we take $\mathscr{Y} = (\mathbb{R}_{>0})^r$, the set where the measured variables reside. In this case, there is a single output for any input, corresponding to the fact that setting the controls yields a unique set of measurements; otherwise, it is not a very good experiment. Thus this is a functional input/output system. •

2.1.3 States for general input/output systems

We have observed that, for a general input/output system, there is the possibility of having multiple outputs for a single output. In this section we explore a way of parameterising this lack of uniqueness of the input→output process. The following definition captures this.

2.1.6 Definition (Response function, state object) For a general input/output system $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$, let a set X_Σ and a map $\rho_\Sigma \colon X_\Sigma \times \mathscr{U} \to \mathscr{Y}$ be given such that

$$\mathscr{B} = \{(\mu, \eta) \in \mathscr{U} \times \mathscr{Y} \mid \text{there exists } x \in X_\Sigma \text{ such that } \rho_\Sigma(x, \mu) = \eta\}.$$

Then ρ_Σ is a ***response function*** and X_Σ is a ***state object***. •

Let us verify the existence of response functions.

2.1.7 Proposition (General input/output systems have response functions) *If $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$ is a general input/output system for which $\mathrm{dom}(\Sigma) = \mathscr{U}$, then there exists a set X_Σ and a map $\rho_\Sigma \colon X_\Sigma \times \mathscr{U} \to \mathscr{Y}$ such that ρ_Σ is a response function with state object X_Σ.*

Proof Let

$$X_\Sigma = \left\{ f \in \mathscr{Y}^{\mathscr{U}} \;\middle|\; \mathrm{graph}(f) \subseteq \mathscr{B} \right\}$$

and define $\rho_\Sigma \colon X_\Sigma \times \mathscr{U} \to \mathscr{Y}$ by $\rho_\Sigma(f, \mu) = f(\mu)$. Because $\mathrm{dom}(\Sigma) = \mathscr{U}$, $X_\Sigma \neq \emptyset$.

Now let $(\mu_0, \eta_0) \in \mathscr{B}$. We claim that there exists $f_{(\mu_0,\eta_0)} \in X_\Sigma$ such that $f_{(\mu_0,\eta_0)}(\mu_0) = \eta_0$. Indeed, let $f \in X_\Sigma$ and then define $\hat{f} \colon \mathscr{U} \to \mathscr{Y}$ by

$$\hat{f}(\mu) = \begin{cases} f(\mu), & \mu \neq \mu_0, \\ \eta_0, & \mu = \mu_0. \end{cases}$$

It is clear that $\hat{f} \in X_\Sigma$ and that $\hat{f}(\mu_0) = \eta_0$, hence our claim holds. Now note that

$$\rho_\Sigma(f_{(\mu_0,\eta_0)}, \mu_0) = f_{(\mu_0,\eta_0)}(\mu_0) = \eta_0.$$

This shows that

$$\mathscr{B} \subseteq \{(\mu, \eta) \in \mathscr{U} \times \mathscr{Y} \mid \text{there exists } x \in X_\Sigma \text{ such that } \rho_\Sigma(x, \mu) = \eta\}.$$

For the converse, suppose that $(\mu_0, \eta_0) \in \mathscr{U} \times \mathscr{Y}$ is such that $\eta_0 = \rho_\Sigma(f, \mu_0)$ for some $f \in X_\Sigma$. Then $\eta_0 = f(\mu_0)$, whence $(\mu_0, \eta_0) \in \mathscr{B}$ since $\mathrm{graph}(f) \subseteq \mathscr{B}$. ■

The preceding result is interesting abstractly. However, it is not very useful in practice since the construction of the response function in the proof will not generally have useful properties. Something of some concern to us will be the matter of the existence of useful state objects and corresponding response functions.

It is insightful to consider response functions in terms of the examples we have presented.

2.1.8 Examples (Response functions, state objects)

1. Let us work with the mass in a gravitational field from Example 2.1.5–2. Here the inputs were the parameters $(m, a_g) \in \mathscr{U} = \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and the outputs were paths for the particle $\xi \in C^2(\mathbb{R}_{\geq 0}; \mathbb{R})$. If one remembers[1] that a solution of a sufficiently well-behaved ordinary differential equation exists and is uniquely determined by its initial condition, then one sees that we can take $X_\Sigma = \mathbb{R}^2$ and define

$$\rho_\Sigma((x_0, v_0), (m, a_g)) = \xi_{(x_0,v_0)},$$

[1] This is a subject we treat in detail in Section 3.2.1.

where $\xi_{(x_0,v_0)}\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ is the unique solution of the initial value problem

$$m\ddot{\xi}(t) = -ma_g, \qquad \xi(0) = x_0,\ \dot{\xi}(0) = v_0. \tag{2.2}$$

We see that $((m, a_g), \xi) \in \mathscr{U} \times \mathscr{Y}$ if and only if there exists $(x_0, v_0) \in X_\Sigma$ such that $\rho_\Sigma((x_0, v_0), (m, a_g)) = \xi$; this is just the statement of the existence and uniqueness theorem for ordinary differential equations, along with the fact that all solutions to the initial value problem (2.2) exist on $\mathbb{R}_{\geq 0}$.

2. Next we work with the Example 2.1.5–3, the deterministic finite state automaton. While we called elements of Q "states," it will not be until we talk systematically about systems with time that we shall be able to recognise this property as a "state space." In our current framework, since an input uniquely determines the behaviour, there is an "obvious" state object and response function, as the reader can show in Exercise 2.1.1.
3. The binary encoder and decoder of Example 2.1.2–4, as functional input/output systems, possess an "obvious" state object and response function.
4. The experimental measurement Example 2.1.5–5, as a functional input/output system, possesses an "obvious" state object and response function. •

2.1.4 Complex general input/output systems

The terminology "complex system" is one that can be used as some sort of fashion statement. Here we shall use this terminology to describe systems that are interconnected in some way. The idea is that one has a generalised input/output system, but that the relationship between the input and the output is the result of an interconnection of subsystems. While such systems are, at a low resolution, just general input/output systems, the purpose of the complex systems point of view is to understand how the behaviours of the subsystems contribute to the behaviour of the full system, and the rôle of the specific interconnections in this process.

Let us define, in our general setting, what we mean by a complex system.

2.1.9 Definition (Complex general input/output system) A ***complex general input/output system*** is a pair $\Sigma = (\mathscr{S}, \mathscr{B})$ where

(i) $\mathscr{S} = ((\mathscr{U}_i, \mathscr{Y}_i, \mathscr{B}_i))_{i\in I}$ is a family of general input/output systems (the ***subsystems*** of Σ) and

(ii) $\mathscr{B} \subseteq \prod_{i\in I} \mathscr{B}_i$ is a general system with objects $(\mathscr{B}_i)_{i\in I}$. •

Let us unravel this rather featureless definition. We recall from Definition I-1.6.7 that an element of $\mathscr{B}$ is a mapping $\beta\colon I \to \cup_{i\in I}\mathscr{B}_i$ for which $\beta(i) \in \mathscr{B}_i$. Let us denote an element of $\prod_{i\in I}\mathscr{U}_i$ by $\mu\colon I \to \cup_{i\in I}\mathscr{U}_i$ with $\mu(i) \in \mathscr{U}_i$. Similarly, elements of $\prod_{i\in I}\mathscr{Y}_i$ are written $\eta\colon I \to \cup_{i\in I}\mathscr{Y}_i$ with $\eta(i) \in \mathscr{Y}_i$. Also as in Definition I-1.6.7, we have the projections $\mathrm{pr}_j\colon \prod_{i\in I}\mathscr{B}_i \to \mathscr{B}_j$, $j \in I$, and similarly for $\prod_{i\in I}\mathscr{U}_i$ and $\prod_{i\in I}\mathscr{Y}_i$. This then allows us to state the following property of complex general input/output systems.

2.1.10 Proposition (Complex general input/output systems are general input/output systems) *A complex general input/output system $\Sigma = (\mathscr{S}, \mathscr{B})$ with $\mathscr{S} = ((\mathscr{U}_i, \mathscr{Y}_i, \mathscr{B}_i))_{i\in I}$ induces a general input/output system $\Sigma' = (\mathscr{U}, \mathscr{Y}, \mathscr{B}')$ satisfying*

(i) $\mathscr{U} = \prod_{i\in I} \mathscr{U}_i$ *and* $\mathscr{Y} = \prod_{i\in I} \mathscr{Y}_i$,

(ii) $\mathscr{B}' = \{(\mu, \eta) \in \mathscr{U} \times \mathscr{Y} \mid \exists\, \beta \in \mathscr{B}$ *such that* $\mathrm{pr}_i(\beta) = (\mathrm{pr}_i(\mu), \mathrm{pr}_i(\eta), i \in I\}$, *and*

(iii) $\mathscr{B} = \{\phi \in \prod_{i\in I}(\mathscr{U}_i \times \mathscr{Y}_i) \mid \exists\, (\mu, \eta) \in \mathscr{B}'$ *such that* $\mathrm{pr}_i(\phi) = (\pi_i(\mu), \pi_i(\eta))\}$.

Proof We take $\mathscr{U}$, $\mathscr{Y}$, and $\mathscr{B}'$ as in parts (i) and (ii), and show that condition (iii) holds. Note that $\mathscr{B}_i \subseteq \mathscr{U}_i \times \mathscr{Y}_i$. Let $\beta \in \mathscr{B}$. Thus $\beta(i) = (\mu_\beta(i), \eta_\beta(i))$ for some mappings

$$\mu_\beta \colon I \to \cup_{i\in I} \mathscr{U}_i, \quad \eta_\beta \colon I \to \cup_{i\in I} \mathscr{Y}_i$$

with $\mu_\beta(i) \in \mathscr{U}_i$ and $\eta_\beta(i) \in \mathscr{Y}_i$. That is to say, $\mu_\beta \in \prod_{i\in I} \mathscr{U}_i$ and $\eta_\beta \in \prod_{i\in I} \mathscr{Y}_i$. Thus $(\mu_\beta, \eta_\beta) \in \mathscr{U} \times \mathscr{Y}$. This shows that

$$\mathscr{B} \subseteq \{\phi \in \prod_{i\in I}(\mathscr{U}_i \times \mathscr{Y}_i) \mid \exists\, (\mu, \eta) \in \mathscr{B}' \text{ such that } \mathrm{pr}_i(\phi) = (\pi_i(\mu), \pi_i(\eta))\}.$$

The opposite inclusion is simply the definition of $\mathscr{B}'$. ■

While this does not completely clarify the situation, it does provide us with some interpretation of a complex general input/output system as a subset of behaviours from all possible combinations of inputs and outputs. However, this is all best illustrated via concrete examples.

2.1.11 Examples (Complex general input/output systems)

1. Consider two general input/output systems $\Sigma_i = (\mathscr{U}_i, \mathscr{Y}_i, \mathscr{B}_i)$, $i \in \{1, 2\}$, and we suppose that we additionally have a set $\mathscr{C}$ and mappings

$$\pi_1 \colon \mathscr{Y}_1 \to \mathscr{C}, \quad \pi_2 \colon \mathscr{U}_2 \to \mathscr{C}.$$

We define the ***serial interconnection*** of Σ_1 and Σ_2 to be the general input/output system

$$\Sigma_2 \circ \Sigma_1 \subseteq (\mathscr{U}_1 \times \mathscr{U}_2, \mathscr{Y}_1 \times \mathscr{Y}_2, \mathscr{B}_2 \circ \mathscr{B}_1)$$

defined by

$$\mathscr{B}_2 \circ \mathscr{B}_1 = \{((\mu_1, \mu_2), (\eta_1, \eta_2)) \in (\mathscr{U}_1 \times \mathscr{U}_2) \times (\mathscr{Y}_1 \times \mathscr{Y}_2) \mid (\mu_i, \eta_i) \in \mathscr{B}_i,\ i \in \{1, 2\},\ \pi_1(\eta_1) = \pi_2(\mu_2)\}.$$

We depict in Figure 2.1 we depict how one should think of a serial interconnection. A particular instance of this is when $\mathscr{Y}_1 = \mathscr{U}_2 = \mathscr{C}$ and $\pi_1 = \pi_2 = \mathrm{id}_{\mathscr{C}}$, in which case diagram simplifies to that shown in Figure 2.1. In this simplified case, we note that the input μ_2 and the output η_1 should not really be thought of as being part of the system inputs and outputs, since they serve the purpose of determining the character of the interconnection. Also, in the simplified case

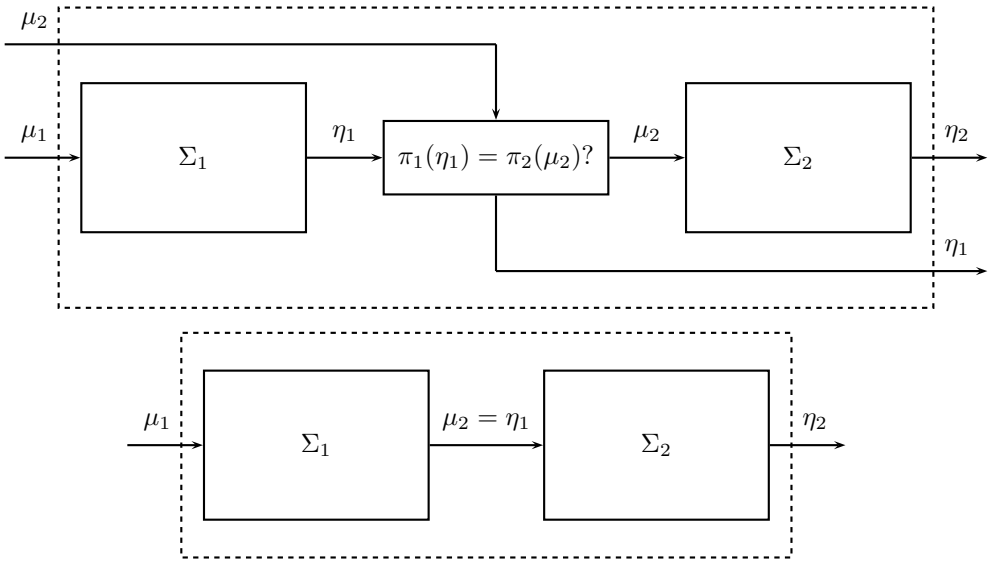


Figure 2.1 General serial interconnection (top) and simple serial interconnection (bottom)

when both Σ_1 and Σ_2 are functional input/output systems, we see one way of understanding the serial interconnection. In such a case we have

$$\eta_2 = F_{\Sigma_2}(\mu_2) = F_{\Sigma_2}(\eta_1) = F_{\Sigma_2} \circ F_{\Sigma_1}(\mu_1),$$

and so the input→output relation is a literal composition.

2. Again, consider two general input/output systems $\Sigma_i = (\mathscr{U}_i, \mathscr{Y}_i, \mathscr{B}_i)$, $i \in \{1, 2\}$. In this case we again assume that we have a set $\mathscr{C}$ and maps $\pi_i \colon \mathscr{U}_i \to \mathscr{C}$. We then define the ***parallel interconnection*** of Σ_1 and Σ_2 to be the general input/output system
$$\Sigma_1 + \Sigma_2 = (\mathscr{U}_1 \times \mathscr{U}_2, \mathscr{Y}_1 \times \mathscr{Y}_2, \mathscr{B}_1 + \mathscr{B}_2)$$
with
$$\mathscr{B}_1 + \mathscr{B}_2 = \{((\mu_1, \mu_2), (\eta_1, \eta_2)) \in (\mathscr{U}_1 \times \mathscr{U}_2) \times (\mathscr{Y}_1 \times \mathscr{Y}_2) | \\ (\mu_i, \eta_i) \in \mathscr{B}_i,\ i \in \{1, 2\},\ \pi_1(\mu_1) = \pi_2(\mu_2)\}.$$
In Figure 2.2 we depict how one can think of the parallel interconnection. The situation simplifies if we take $\mathscr{U}_1 = \mathscr{U}_2 = \mathscr{C}$ and $\pi_1 = \pi_2 = \mathrm{id}_{\mathscr{C}}$. In this case we show in Figure 2.2 the manner in which the interconnection simplifies. To understand the "+" notation in the case of functional input/output systems, we make the following "computation," assuming that all symbols make sense:
$$\eta_1 + \eta_2 = F_{\Sigma_1}(\mu_1) + F_{\Sigma_2}(\mu_2) = (F_{\Sigma_1} + F_{\Sigma_2})(\mu),$$
where $\mu = \mu_1 = \mu_2$. •

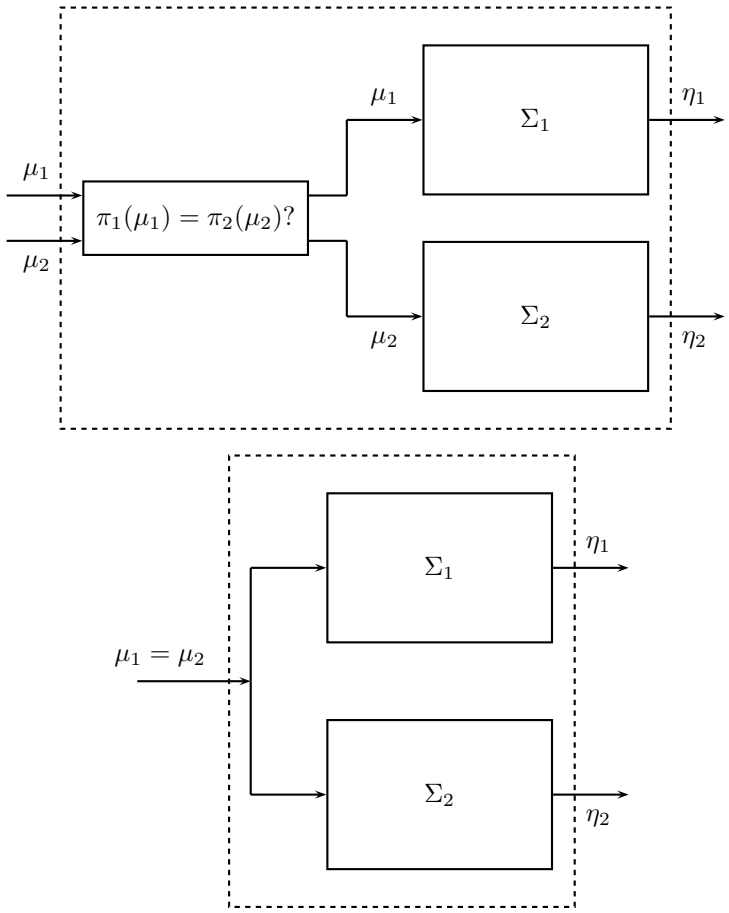


Figure 2.2 General parallel interconnection (top) and simple parallel interconnection (bottom)

2.1.5 Linear general input/output systems

We now introduce, in our general setting, the notion of a linear system. Linear systems comprise the balance of systems in which we shall be interested, so this section marks, in some way, the beginning of the content of the volume.

We start with the definition.

2.1.12 Definition (Linear general input/output system) A ***linear general input/output system*** is a triple $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$, where

(i) $\mathscr{U}$ and $\mathscr{Y}$ are vector spaces over some field F and

(ii) $\mathscr{B}$ is a subspace of $\mathscr{U} \oplus \mathscr{Y}$. •

Linear systems admit particularly illustrative response functions.

2.1.13 Proposition (Linear general input/output systems admit linear response functions) *Let* F *be a field, let* $\mathscr{U}$ *and* $\mathscr{B}$ *be* F*-vector spaces, and let* $\mathscr{B} \subseteq \mathscr{U} \oplus \mathscr{Y}$ *(not necessarily a subspace,* a priori*). Then the following statements are equivalent:*

(i) $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$ *is a linear general input/output system;*

(ii) *there exist an* F*-vector space* X_Σ *and linear mappings* $\mathsf{R}_{\Sigma,s} \in \mathrm{Hom}_\mathsf{F}(\mathsf{X}_\Sigma; \mathscr{Y})$ *and* $\mathsf{R}_{\Sigma,\mathrm{i}} \in \mathrm{Hom}_\mathsf{F}(\mathscr{U}; \mathscr{Y})$ *such that* X_Σ *is a state object and*

$$\begin{aligned} \mathsf{R}_\Sigma \colon \mathsf{X}_\Sigma \oplus \mathscr{U} &\to \mathscr{Y} \\ (\mathsf{x}, \mu) &\mapsto \mathsf{R}_{\Sigma,s}(\mathsf{x}) + \mathsf{R}_{\Sigma,\mathrm{i}}(\mu) \end{aligned}$$

is a response function.

Proof (i) $\Longrightarrow$ (ii) First we claim that there exists $\mathsf{R}_{\Sigma,i} \in \mathrm{Hom}_\mathsf{F}(\mathscr{U}; \mathscr{Y})$ such that

$$\{(\mu, \mathsf{R}_{\Sigma,i}(\mu)) \mid \mu \in \mathscr{U}\} \subseteq \mathscr{B}.$$

We will first prove this for $\mathrm{dom}(\Sigma) = \mathscr{U}$.

Let

$$\mathscr{L}_\mathscr{B} = \{\mathsf{L} \in \mathrm{Hom}_\mathsf{F}(\mathscr{U}'; \mathscr{Y}) \mid \mathscr{U}' \subseteq \mathscr{U},\ \{(\mu', \mathsf{L}(\mu')) \mid \mu' \in \mathscr{U}'\} \subseteq \mathscr{B}\}.$$

We define a partial order on $\mathscr{L}_\mathscr{B}$ by requiring that $\mathsf{L}_1 \preceq \mathsf{L}_2$ if $\mathrm{dom}(\mathsf{L}_1) \subseteq \mathrm{dom}(\mathsf{L}_2)$ and $\mathsf{L}_2|\mathrm{dom}(\mathsf{L}_1) = \mathsf{L}_1$. We claim that $\mathscr{L}_\mathscr{B} \neq \emptyset$. Indeed, let $(\mu, \eta) \in \mathscr{B}$ and define $\mathscr{U}' = \mathrm{span}_\mathsf{F}(\mu)$ and then define $\mathsf{L} \in \mathrm{Hom}_\mathsf{F}(\mathscr{U}'; \mathscr{Y})$ by the requirement that $\mathsf{L}(\mu) = \eta$. Then $\mathsf{L} \in \mathscr{L}_\mathscr{B}$. Let $\mathscr{P} \subseteq \mathscr{L}_\mathscr{B}$ be a totally ordered subset. For $\mathsf{L} \in \mathscr{P}$ let $\mathscr{U}_\mathsf{L} = \mathrm{dom}(\mathsf{L})$. Define $\mathscr{U}_\mathscr{P} = \cup_{\mathsf{L} \in \mathscr{P}} \mathscr{U}_\mathsf{L}$ and define $\mathsf{L}_\mathscr{P} \in \mathrm{Hom}_\mathsf{F}(\mathscr{U}_\mathscr{P}; \mathscr{Y})$ by requiring that $\mathsf{L}_\mathscr{P} = \mathsf{L}|\mathscr{U}_\mathsf{L}$ for $\mathsf{L} \in \mathscr{P}$. The definition of the partial order ensures that this definition makes sense. We claim that $\mathsf{L}_\mathscr{P} \in \mathscr{P}$. We first show that $\mathsf{L}_\mathscr{P}$ is indeed linear. Let $\mu_1, \mu_2 \in \mathscr{U}_\mathscr{P}$ and let $\mathsf{L} \in \mathscr{P}$ be such that $\mu_1, \mu_2 \in \mathscr{U}_\mathsf{L}$. Then

$$\mathsf{L}_\mathscr{P}(\mu_1 + \mu_2) = \mathsf{L}(\mu_1 + \mu_2) = \mathsf{L}_\mathscr{P}(\mu_1) + \mathsf{L}_\mathscr{P}(\mu_2).$$

Similarly, $\mathsf{L}_\mathscr{P}(a\mu) = a\mathsf{L}_\mathscr{P}(\mu)$ for $a \in \mathsf{F}$ and $\mu \in \mathscr{U}_\mathscr{P}$. Moreover, if $\mu \in \mathscr{U}_\mathscr{P}$. then for $\mathsf{L} \in \mathscr{P}$ such that $\mu \in \mathrm{dom}(\mathsf{L})$, we have

$$(\mu, \mathsf{L}_\mathscr{P}(\mu)) = (\mu, \mathsf{L}(\mu)) \in \mathscr{B}.$$

Thus $\mathsf{L}_\mathscr{P} \in \mathscr{P}$, as claimed. This shows that the totally ordered set $\mathscr{P}$ has an upper bound. Thus, by Zorn's Lemma, $\mathscr{L}_\mathscr{B}$ has a maximal element, which we denote by $\mathsf{R}_{\Sigma,i}$. We claim that $\mathrm{dom}(\mathsf{R}_{\Sigma,i}) = \mathscr{U}$. Suppose otherwise. Then there exists $\mu' \notin \mathrm{dom}(\mathsf{R}_{\Sigma,i})$. Define

$$\mathscr{U}' = \mathrm{dom}(\mathsf{R}_{\Sigma,i}) \oplus \mathrm{span}_\mathsf{F}(\mu').$$

Since $\mathrm{dom}(\Sigma) = \mathscr{U}$, there exists $\eta' \in \mathscr{Y}$ so that $(\mu', \eta') \in \mathscr{B}$. Now, it $\mu + a\mu' \in \mathscr{U}'$ with $\mu \in \mathrm{dom}(\mathsf{R}_{\Sigma,i})$, then define $\mathsf{L}' \in \mathrm{Hom}_\mathsf{F}(\mathscr{U}'; \mathscr{Y})$ by

$$\mathsf{L}'(\mu + a\mu') = \mathsf{R}_{\Sigma,i}(\mu) + a\eta'.$$

We easily verify that L' is linear. Also,

$$(\mu + a\mu', \mathsf{R}_{\Sigma,i}(\mu) + a\eta') \in \mathscr{B}, \qquad \mu' \in \mathscr{U}',\ a \in \mathsf{F}.$$

Since $\mathsf{L}'|\operatorname{dom}(\mathsf{R}_{\Sigma,i}) = \mathsf{R}_{\Sigma',i}$, we thus contradict the maximality of $\mathsf{R}_{\Sigma,i}$, and so conclude that $\operatorname{dom}(\mathsf{R}_{\Sigma,i}) = \mathscr{U}$.

Now we prove that there exists $\mathsf{R}_{\Sigma,i} \in \operatorname{Hom}_{\mathsf{F}}(\mathscr{U};\mathscr{Y})$ such that

$$\{(\mu, \mathsf{R}_{\Sigma,i}(\mu)) \mid \mu \in \mathscr{U}\} \subseteq \mathscr{B}$$

even when $\operatorname{dom}(\Sigma) \subset \mathscr{U}$. By Theorem I-4.5.52, we let $\mathscr{U}_1$ be a complement to $\operatorname{dom}(\Sigma) \subseteq \mathscr{U}$. Thus $\mathscr{U} = \operatorname{dom}(\Sigma) \oplus \mathscr{U}_1$. Then, by the previous paragraph, there exists $\mathsf{R}'_{\Sigma,i} \in \operatorname{Hom}_{\mathsf{F}}(\operatorname{dom}(\Sigma);\mathscr{Y})$ such that $(\mu, \mathsf{R}'_{\Sigma,i}(\mu)) \in \mathscr{B}$ for every $\mu \in \operatorname{dom}(\Sigma)$. Then define $\mathsf{R}_{\Sigma,i} \in \operatorname{Hom}_{\mathsf{F}}(\mathscr{U};\mathscr{Y})$ by

$$\mathsf{R}_{\Sigma,i}(\mu + \mu_1) = \mathsf{R}_{\Sigma,i}(\mu), \qquad \mu \in \operatorname{dom}(\Sigma),\ \mu_1 \in \mathscr{U}_1.$$

Clearly $\mathsf{R}_{\Sigma,i}$ has the desired property.

Now we define X_Σ and $\mathsf{R}_{\Sigma,s}$. We take

$$\mathsf{X}_\Sigma = \{(0_{\mathscr{U}}, \eta) \mid (0_{\mathscr{U}}, \eta) \in \mathscr{B}\} = \mathscr{B} \cap (\{0_{\mathscr{U}}\} \oplus \mathscr{Y}).$$

By Proposition I-4.5.34, X_Σ is an F-vector space. Define

$$\begin{aligned} \mathsf{R}_{\Sigma,s} \colon \mathsf{X}_\Sigma &\to \mathscr{Y} \\ (0_{\mathscr{U}}, \eta) &\mapsto \eta. \end{aligned}$$

Obviously $\mathsf{R}_{\Sigma,s}$ is linear.

Taking

$$\mathsf{R}_\Sigma(x, \mu) = \mathsf{R}_{\Sigma,s}(x) + \mathsf{R}_{\Sigma,i}(\mu),$$

let us prove that R_Σ is a response function. Indeed, let $(\mu, \eta) \in \mathscr{B}$ so that $(\mu, \mathsf{R}_{\Sigma,i}(\mu)) \in \mathscr{B}$. Therefore,

$$(0_{\mathscr{U}}, \eta - \mathsf{R}_{\Sigma,i}(\mu)) \in \mathscr{B}.$$

Thus

$$\eta - \mathsf{R}_{\Sigma,i}(\mu) = \mathsf{R}_{\Sigma,s}(x) \quad \Longrightarrow \quad \eta = R_\Sigma(x, \mu)$$

for some $x \in \mathsf{X}_\Sigma$ by definition of $\mathsf{R}_{\Sigma,s}$. Thus

$$\mathscr{B} \subseteq \{(\mu, \eta) \in \mathscr{U} \oplus \mathscr{Y} \mid \text{there exists } x \in \mathsf{X}_\Sigma \text{ such that } \eta = \mathsf{R}_\Sigma(x, \mu)\}.$$

Next let $(\mu, \eta) \in \mathscr{U} \oplus \mathscr{Y}$ has the property that there exists $x \in \mathsf{X}_\Sigma$ such that $\eta = R_\Sigma(x, \mu)$. Thus

$$\eta = \mathsf{R}_{\Sigma,s}(x) + \mathsf{R}_{\Sigma,i}(\mu).$$

Note that

$$(0_{\mathscr{U}}, \mathsf{R}_{\Sigma,s}(x)) \in \mathscr{B}, \quad (\mu, \mathsf{R}_{\Sigma,i}(\mu)) \in \mathscr{B}$$

by definition of $\mathsf{R}_{\Sigma,s}$ and $\mathsf{R}_{\Sigma,i}$. Since $\mathscr{B}$ is a subspace we have

$$(\mu, \mathsf{R}_{\Sigma,s}(x) + \mathsf{R}_{\Sigma,i}(\mu)) \in \mathscr{B}.$$

Thus

$$\{(\mu, \eta) \in \mathscr{U} \oplus \mathscr{Y} \mid \text{there exists } x \in \mathsf{X}_\Sigma \text{ such that } \eta = \mathsf{R}_\Sigma(x, \mu)\} \subseteq \mathscr{B}$$

and so R_Σ is indeed a response function.

(ii) $\implies$ (i) Let $(\mu_1, \eta_1), (\mu_2, \eta_2) \in \mathscr{B}$. Then, since X_Σ is a state object and R_Σ is a response function, there exists $x_1, x_2 \in \mathsf{X}_\Sigma$ such that

$$\eta_i = \mathsf{R}_\Sigma(x_i, \mu_i) = \mathsf{R}_{\Sigma,s}(x_i) + \mathsf{R}_{\Sigma,i}(\mu_i), \qquad i \in \{1, 2\}.$$

Then

$$\begin{aligned}\eta_1 + \eta_2 &= \mathsf{R}_{\Sigma,s}(x_1) + \mathsf{R}_{\Sigma,i}(\mu_1) + \mathsf{R}_{\Sigma,s}(x_2) + \mathsf{R}_{\Sigma,i}(\mu_2)\\ &= \mathsf{R}_\Sigma(x_1 + x_2, \mu_1 + \mu_2),\end{aligned}$$

showing that $(\mu_1 + \mu_2, \eta_1 + \eta_2) \in \mathscr{B}$. Similarly one shows that, if $(\mu, \eta) \in \mathscr{B}$ and $a \in \mathsf{F}$, then $a(\mu, \eta) \in \mathscr{B}$, and so $\mathscr{B}$ is a subspace. ■

Of course, a functional linear general input/output system is a linear general input/output system $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$ for which $\mathscr{B}$ is the graph of a linear function $\mathsf{F}_\Sigma \in \mathrm{Hom}_{\mathsf{F}}(\mathscr{U}; \mathscr{Y})$. Thus, if such systems are going to be interesting, it will be because of their specific structure since the general structure is pretty featureless. Let us give an interesting example of a specific functional linear general input/output system.

2.1.14 Example (Linear general input/output system) Let $k \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$ and define a mapping $F_k \colon \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R}) \to \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$ by

$$F_k(f)(t) = \int_0^t f(\tau) k(t - \tau) \, \mathrm{d}\tau$$

(see Theorem IV-4.1.13). By linearity of the integral, we see that

$$\Sigma = (\mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R}), \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R}), \mathrm{graph}(F_k))$$

is a functional linear general input/output system. This will be an important class of systems in this volume, and they are known as "convolution systems" (see). • what?

2.1.6 Notes

[Mesarovic and Takahara 1975, Mesarovic and Takahara 1989]

Exercises

2.1.1 Show that a functional input/output system possesses an "obvious" state object X_Σ and response function ρ_Σ.

2.1.2 Let $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$ be a general input/output system with state object X_Σ and corresponding response function ρ_Σ. Given $x_0 \in X_\Sigma$, define

$$\begin{aligned}F_{\Sigma,x_0} \colon \mathscr{U} &\to \mathscr{Y}\\ \mu &\mapsto \rho_\Sigma(x_0, \mu)^{\cdot}\end{aligned}$$

Answer the following questions.

(a) Show that F_{Σ,x_0} corresponds to a functional input/output system.
(b) Carry this out for the Example 2.1.5–2; that is, for each state, what is the corresponding functional input/output system? what does it mean physically?

2.1.3 Consider a

Section 2.2

General time systems

The focus in this volume is on systems whose inputs, outputs, and states are functions of time. In this section we consider in our abstract setting systems with time. We shall subsequently be primarily interested in systems described by differential and difference equations. However, it is illustrative to develop some of the properties of these systems in a more abstract setting where, in a certain sense, they are easier to motivate and understand.

2.2.1 General time-domains

We have previously carefully considered "time," particularly when working with transform theory. In this previous setting, we had denoted by $\mathbb{T}$ a "time-domain," by which we meant a subset of $\mathbb{R}$ of the form $\mathbb{T} = \mathbb{S} \cap I$ where $\mathbb{S}$ is a semigroup in $(\mathbb{R}, +)$ and I is an interval (see Definition IV-1.1.2). In this chapter, and a few times subsequently when we reference the general theory of systems, we shall consider a more general notion of time.

2.2.1 Definition (General time-domain) A *general time-domain* is a totally ordered set $(\mathbb{T}, \leq)$. When we make reference to a general time-domain, we shall sometimes simply write "$\mathbb{T}$," assuming the partial order. •

Of course, by Zermelo's Well Ordering Theorem (Theorem I-1.5.16), *any* set has a partial order making it a totally ordered set, so the notion of a time-domain apparently places no restriction on the sets of things we can consider as time. But the order is just as crucial in the definition as is the set. In any case, we shall mostly work with the simple sorts of time-domains described by Definition IV-1.1.2.

It will be convenient to introduce some notation regarding general time-domains. To this end, we introduce the following terminology associated with a general time-domain $(\mathbb{T}, \leq)$ and for $t, t_0, t_1 \in \mathbb{T}$ with $t_0 \leq t_1$:

$$
\begin{aligned}
\mathbb{T}_{<t} &= \{\tau \in \mathbb{T} \mid \tau < t\}, & \mathbb{T}_{[t_0,t_1]} &= \{\tau \in \mathbb{T} \mid t_0 \leq \tau \leq t_1\},\\
\mathbb{T}_{\leq t} &= \{\tau \in \mathbb{T} \mid \tau \leq t\}, & \mathbb{T}_{[t_0,t_1)} &= \{\tau \in \mathbb{T} \mid t_0 \leq \tau < t_1\},\\
\mathbb{T}_{>t} &= \{\tau \in \mathbb{T} \mid \tau > t\}, & \mathbb{T}_{(t_0,t_1]} &= \{\tau \in \mathbb{T} \mid t_0 < \tau \leq t_1\},\\
\mathbb{T}_{\geq t} &= \{\tau \in \mathbb{T} \mid \tau \geq t\}, & \mathbb{T}_{(t_0,t_1)} &= \{\tau \in \mathbb{T} \mid t_0 < \tau < t_1\}.
\end{aligned}
$$

These are, of course, generalisations of the notion of an interval, and we call them ***sub-time-domains***. Note that each sub-time-domain is itself a general time-domain with the partial order induced from $\mathbb{T}$.

We shall sometimes want to concatenate two general time-domains to form a new one. The following definition pertains to this idea.

2.2.2 Definition (Concatenation of general time-domains) Let $(\mathbb{T}, \leq)$ be a general time-domain.

(i) Given two sub-time-domains $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{T}$, we say that $\mathbb{S}_2$ ***follows*** $\mathbb{S}_1$ if

(a) $\mathbb{T}_{(\sup(\mathbb{S}_1),\inf(\mathbb{S}_2))} = \emptyset$ and

(b) $\mathbb{S}_1 \cap \mathbb{S}_2 = \emptyset$

(see Definition I-1.5.11 for notation).

(ii) If the sub-time-domain $\mathbb{S}_2$ follows the sub-time-domain $\mathbb{S}_1$, the ***concatenation*** of $\mathbb{S}_1$ and $\mathbb{S}_2$ is

$$\mathbb{S}_1 * \mathbb{S}_2 = \mathbb{S}_1 \cup \mathbb{S}_2.$$ •

We ask the reader to prove in Exercise 2.2.1 that the concatenation of two sub-time-domains is itself a sub-time-domain.

The usual notion of time-domain as we previously used it had important additional structure, namely the group structure inherited from $\mathbb{R}$. We generalise this as follows.

2.2.3 Definition (Additive general time-domain, stationary time-domain)

(i) An ***additive general time-domain*** is a pair $((\mathbb{S}, \leq), \mathbb{T})$ where

(a) $(\mathbb{S}, \leq)$ is a general time-domain with the structure of an Abelian group—with group operation denoted "+" and with identity element denoted by "0"—with the property that $t_0 \leq t_1$ if and only if $0 \leq t_1 - t_0$, and

(b) $\mathbb{T} \subseteq \mathbb{S}$ is a segment.

When referring to an additive time-domain we shall often simply write "$\mathbb{T}$," assuming the partial order and the Abelian group containing $\mathbb{T}$.

(ii) A ***stationary time-domain*** is an additive time-domain $((\mathbb{S}, \leq), \mathbb{T})$ with the property that, for every $\bar{t} \in \mathbb{T}$, there is a bijective mapping

$$\begin{aligned} \tau_{-\bar{t}} \colon \mathbb{T} &\to \mathbb{T}_{\geq \bar{t}} \\ t &\mapsto t + \bar{t}. \end{aligned}$$

The mapping $\tau_{-\bar{t}}$ is the ***time shift*** by $\bar{t}$.[2] •

The additive time-domains we will consider in practice are $((\mathbb{R}, \leq), \mathbb{T})$ and $((\mathbb{Z}(\Delta)), \mathbb{T})$, where $\mathbb{R}$ and $\mathbb{Z}(\Delta)$ have the usual additive group structure. In Exercise 2.2.4 the reader can verify that these are simply the time-domains considered previously in Definition IV-1.1.2. In Exercise 2.2.4 the reader can think about what are stationary time-domains in these typical cases.

[2]Note that the sign of the shift here is the opposite of that from Example IV-1.1.6–1.

2.2.2 Functions on general time-domains

Now that we have laid out the properties of time in a general way, let us consider functions on general time-domains. It is necessary[3] to be able to discuss functions defined only on a sub-time-domain of a general time-domain, and this complicates things.

2.2.4 Definition (Partial time function on a general time-domain) Let S be a set and let $(\mathbb{T}, \leq)$ be a general time-domain.

(i) A ***partial time function*** on $\mathbb{T}$ with values in S is a map $f\colon \mathbb{T}' \to S$ where $\mathbb{T}' \subseteq \mathbb{T}$ is a sub-time-domain.

(ii) The domain of a partial time function f on $\mathbb{T}$ with values in S is denoted by $\mathrm{dom}(f)$.

(iii) By $S^{(\mathbb{T})}$ we denote the set of partial time functions on $\mathbb{T}$ with values in S.

(iv) If $f \in S^{(\mathbb{T})}$, we denote by $\mathrm{dom}(f) \subseteq \mathbb{T}$ the domain of f. •

Note that we can think of dom as being a mapping $\mathrm{dom}\colon S^{(\mathbb{T})} \to \mathbf{2}^{\mathbb{T}}$. Doing this, $\mathrm{dom}^{-1}(\mathbb{S})$ denotes the set of mappings from $\mathbb{S}$ to S.

Just as one can concatenate sub-time-domains, one can concatenate functions defined on sub-time-domains.

2.2.5 Definition (Concatenation, restriction, and extension of partial time functions) Let S be a set and let $(\mathbb{T}, \leq)$ be a general time-domain.

(i) If $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{T}$ are sub-time-domains with $\mathbb{S}_2$ following $\mathbb{S}_1$ and if $f_1, f_2 \in S^{(\mathbb{T})}$ satisfy $\mathrm{dom}(f_i) = \mathbb{S}_i$, $i \in \{1, 2\}$, the ***concatenation*** of f_1 and f_2 is $f_1 * f_2 \in S^{(\mathbb{T})}$ with $\mathrm{dom}(f_1 * f_2) = \mathbb{S}_1 * \mathbb{S}_2$ and given by

$$f_1 * f_2(t) = \begin{cases} f_1(t), & t \in \mathbb{S}_1, \\ f_2(t), & t \in \mathbb{S}_2. \end{cases}$$

We shall say that the pair (f_1, f_2) is ***concatenatable*** if the conditions just preceding apply.

(ii) If $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{T}$ are sub-time-domains satisfying $\mathbb{S}_2 \subseteq \mathbb{S}_1$ and if $f \in S^{(\mathbb{T})}$ satisfies $\mathrm{dom}(f) = \mathbb{S}_1$, then the ***restriction*** of f to $\mathbb{S}_2$ is $f|\mathbb{S}_2 \in S^{(\mathbb{T})}$ being the usual set theoretic restriction.

(iii) If $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{T}$ are sub-time-domains satisfying $\mathbb{S}_2 \subseteq \mathbb{S}_1$ and if $f_1, f_2 \in S^{(\mathbb{T})}$ satisfy $\mathrm{dom}(f_i) = \mathbb{S}_i$, $i \in \{1, 2\}$, then f_1 is an ***extension*** of f_2 if $f_1|\mathbb{S}_2 = f_2$.

We sometimes wish to consider classes of partial time functions with various properties relative to concatenation or restriction. To this end, let $\mathscr{S} \subseteq S^{(\mathbb{T})}$. The collection $\mathscr{S}$ of partial time functions is:

[3] This necessity arises, for example, because an ordinary differential equation may only possess solutions for finite times (see Example 2.2.21).

(iv) ***closed under concatenation*** if, for every concatenatable pair (f_1, f_2) of partial time functions from $\mathscr{S}$, we have $f_1 * f_2 \in \mathscr{S}$;

(v) ***closed under restriction*** if, for every $f \in \mathscr{S}$ with $\mathrm{dom}(\mathbb{S}_1)$ and every sub-time-domain $\mathbb{S}_2 \subseteq \mathbb{S}_1$, $f|\mathbb{S}_2 \in \mathscr{S}$;

(vi) ***extendible*** if, for every $f \in \mathscr{S}$, there exists $g \in \mathrm{dom}^{-1}(\mathbb{T}) \cap \mathscr{S}$ such that g is an extension of f;

(vii) ***regular*** if it is closed under concatenations, closed under restrictions, and extendible. •

If $\mathscr{S} \subseteq S^{(\mathbb{T})}$ is a collection of partial time functions and if $\mathbb{S} \subseteq \mathbb{S}$ is a sub-time-interval, we denote

$$\mathscr{S}_{\mathbb{S}} = \{f|(\mathbb{S} \cap \mathrm{dom}(f)) \mid f \in \mathscr{S}\}.$$

Following our notation above for sub-time-intervals, we abbreviate

$$\mathscr{S}_{<t} = \{f|(\mathbb{T}_{<t} \cap \mathrm{dom}(f)) \mid f \in \mathscr{S}\}, \qquad \mathscr{S}_{>t} = \{f|(\mathbb{T}_{>t} \cap \mathrm{dom}(f)) \mid f \in \mathscr{S}\},$$
$$\mathscr{S}_{\leq t} = \{f|(\mathbb{T}_{\leq t} \cap \mathrm{dom}(f)) \mid f \in \mathscr{S}\}, \qquad \mathscr{S}_{\geq t} = \{f|(\mathbb{T}_{\geq t} \cap \mathrm{dom}(f)) \mid f \in \mathscr{S}\}.$$

Note that these subsets of partial time functions are not required to be in the original collection $\mathscr{S}$. We also denote

$$\mathscr{S}_t = \{f(t) \mid f \in \mathscr{S}\}, \tag{2.3}$$

noting that $\mathscr{S}_t \subseteq S$.

While these notions are simple, it is worthwhile to consider a few familiar classes of functions that do or do not satisfy the various conditions of the preceding definition.

2.2.6 Examples (Concatenation, restriction, extension)

1. We take $\mathbb{T} = \mathbb{R}$, $S = \mathbb{R}$, and consider the class $\mathscr{S}$ of partial time functions that are continuous on their domain. We then have the following attributes of the collection $\mathscr{S}$.
   (a) $\mathscr{S}$ is not closed under concatenation: Consider the two partial functions f_1 with domain $[0, 1]$ and f_2 with domain $(1, 2]$ given by $f_1(t) = 1$ and $f_2 = 0$. Clearly $f_1, f_2 \in \mathscr{S}$ and (f_1, f_2) are concatenatable. However, $f_1 * f_2$ is not continuous.
   (b) $\mathscr{S}$ is closed under restriction: This is clear.
   (c) $\mathscr{S}$ is not extendible: Consider $f \in \mathscr{S}$ with domain $(-\frac{\pi}{2}, \frac{\pi}{2})$ and given by $f = \tan^{-1}$. There is no continuous function $g \in C^0(\mathbb{R}; \mathbb{R})$ satisfying $f = g|(-\frac{\pi}{2}, \frac{\pi}{2})$ (what would be the value of g at $\pm\frac{\pi}{2}$?).
2. We take $\mathbb{T} = \mathbb{R}$, $S = \mathbb{R}$, and consider the class $\mathscr{S}$ of partial time functions that are Lebesgue integrable on their domain. For this collection of partial time functions, we have the following attributes.

(a) $\mathscr{S}$ is closed under concatenation: This follows from Proposition III-2.7.22.
(b) $\mathscr{S}$ is closed under restriction: This is clear.
(c) $\mathscr{S}$ is extendible: A function that is integrable on some interval $\mathbb{S}$ can be extended to an integrable function on $\mathbb{R}$ by taking its value to be zero on $\mathbb{R} \setminus \mathbb{S}$.

3. We take $\mathbb{T} = \mathbb{R}$, $S = \mathbb{R}$, and let $\mathscr{S}$ be the collection of partial time functions with the property that, if $f \in \mathscr{S}$ is such that $\mathrm{dom}(f) \subseteq \mathbb{R}_{\geq 0}$, then $f(t) = 0$ for all $t \in \mathrm{dom}(f)$. For this class of partial time functions, we have the following properties.
(a) $\mathscr{S}$ is closed under concatenations: This is clear.
(b) $\mathscr{S}$ is not closed under restriction: If we take f to have domain $\mathbb{R}$ and be defined by $f(t) = 1$ for all $t \in \mathrm{dom}(f)$, then the restriction $f|\mathbb{R}_{\geq 0}$ is not in $\mathscr{S}$.
(c) $\mathscr{S}$ is extendible: This is clear. •

The examples suggest some interesting classes of partial time functions that we will subsequently encounter. Let us give some notation for this.

2.2.7 Notation (Common classes of partial time functions) We work with continuous- and discrete-time. We consider scalar-valued signals here, with signals taking values in finite-dimensional spaces following straightforwardly as in Section IV-1.4.

We begin with discrete-time signals, letting $\mathbb{T}$ be a discrete time-domain as in Definition IV-1.1.2. Then we denote:

(i) $\mathbb{F}^{(\mathbb{T})}$ is the set of all partial time signals on $\mathbb{T}$;
(ii) $\mathsf{c}_{\mathrm{fin}}((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \in \mathsf{c}_{\mathrm{fin}}(\mathrm{dom}(f);\mathbb{F})\}$;
(iii) $\mathsf{c}_0((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \in \mathsf{c}_0(\mathrm{dom}(f);\mathbb{F})\}$;
(iv) $\ell^p((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \in \ell^p(\mathrm{dom}(f);\mathbb{F})\}$, $p \in [1,\infty]$.

For $f \in \ell^p((\mathbb{T});\mathbb{F})$, we denote by $\|f\|_p$ the p-norm of f as an element of $\ell^p(\mathrm{dom}(f);\mathbb{F})$. For $f \in \mathsf{c}_0((\mathbb{T});\mathbb{F})$, we denote by $\|f\|_\infty$ the ∞-norm of f as an element of $\mathsf{c}_0(\mathrm{dom}(f);\mathbb{F})$. Of course, all norms can be defined for $f \in \mathsf{c}_{\mathrm{fin}}((\mathbb{T});\mathbb{F})$.

Now we work with continuous-time signals, letting $\mathbb{T}$ be a continuous time-domain as in Definition IV-1.1.2. To define the appropriate Lebesgue spaces, we need equivalence classes of signals in $\mathbb{F}^{(\mathbb{T})}$ that agree almost everywhere. The way to do this is clear, but let us enunciate this. Two signals $f_1, f_2 \in \mathbb{F}^{(\mathbb{T})}$ are ***equivalent*** if $\mathrm{dom}(f_1) = \mathrm{dom}(f_2)$ and if

$$\lambda(\{t \in \mathbb{T} \mid f_1(t) - f_2(t) \neq 0\}) = 0.$$

We denote the equivalence class containing f by $[f]$. Then we denote:

(i) $\mathbb{F}^{(\mathbb{T})}$ is the set of all partial time signals on $\mathbb{T}$;
(ii) $\mathsf{C}^r_{\mathrm{cpt}}((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \in \mathsf{C}^r_{\mathrm{cpt}}(\mathrm{dom}(f);\mathbb{F})\}$, $r \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$;

(iii) $\mathsf{C}_0^r((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \in \mathsf{C}_0^r(\mathrm{dom}(f);\mathbb{F})\}$, $r \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$;
(iv) $\mathsf{C}_{\mathrm{bdd}}^r((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \in \mathsf{C}_{\mathrm{bdd}}^r(\mathrm{dom}(f);\mathbb{F})\}$, $r \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$;
(v) $\mathsf{C}^r((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \in \mathsf{C}^r(\mathrm{dom}(f);\mathbb{F})\}$, $r \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$;
(vi) $\mathsf{L}^{(p)}((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \in \mathsf{L}^{(p)}(\mathrm{dom}(f);\mathbb{F})\}$, $p \in [1,\infty]$;
(vii) $\mathsf{L}^p((\mathbb{T});\mathbb{F}) = \{[f] \in \mathbb{F}^{(\mathbb{T})} \mid [f]_p \in \mathsf{L}^p(\mathrm{dom}(f);\mathbb{F})\}$, $p \in [1,\infty]$;
(viii) $\mathsf{L}^p_{\mathrm{loc}}((\mathbb{T});\mathbb{F}) = \{[f] \in \mathbb{F}^{(\mathbb{T})} \mid [f]_p \in \mathsf{L}^p_{\mathrm{loc}}(\mathrm{dom}(f);\mathbb{F})\}$, $p \in [1,\infty]$;
(ix) $\mathsf{AC}((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \text{ is absolutely continuous}\}$;[4]
(x) $\mathsf{AC}_{\mathrm{loc}}((\mathbb{T});\mathbb{F}) = \{f \in \mathbb{F}^{(\mathbb{T})} \mid f \text{ is locally absolutely continuous}\}$.[5] •

Next we consider some particular constructions on stationary time-domains.

2.2.8 Definition (Partial functions on stationary time-domains) Let $((\mathbb{S},\leq),\mathbb{T})$ be a stationary time-domain, let S be a set, and let $\mathscr{S} \subseteq S^{(\mathbb{T})}$ be a collection of partial time functions.

(i) For $f \in \mathscr{S}$, a time $\bar{t} \in \mathbb{T}$ is **f-*admissible*** if

$$\{t + \bar{t} \mid t \in \mathrm{dom}(f)\} \subseteq \mathbb{T}.$$

We denote by $\tau(f)$ the set of f-admissible times.

(ii) For $f \in \mathscr{S}$ and $\bar{t} \in \tau(f)$, denote by $\tau_{\bar{t}}^* f \in S^{(\mathbb{T})}$ the partial function with

$$\mathrm{dom}(\tau_{\bar{t}}^* f) = \{t + \bar{t} \mid t \in \mathrm{dom}(f)\}$$

and $\tau_{\bar{t}}^* f(t) = f(t - \bar{t})$.

(iii) The collection $\mathscr{S}$ of partial time functions is ***stationary*** if

$$\tau_{\bar{t}}(\mathscr{S}) \triangleq \{\tau_{\bar{t}}^* f \mid f \in \mathscr{S},\ \bar{t} \in \tau(f)\} \subseteq \mathscr{S}.$$ •

2.2.3 Definition and basic properties of time systems

We can now give a definition of a general class of systems where objects depend on time.

2.2.9 Definition (General time system) A ***general time system*** is a sextuple

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

where

(i) U is a set (the ***input set***),
(ii) Y is a set (the ***output set***),
(iii) $\mathbb{T}$ is a general time-domain,

[4] See Definition III-2.9.23
[5] Ibid.

(iv) $\mathscr{U} \subseteq U^{(\mathbb{T})}$ (the ***admissible input signals***),
(v) $\mathscr{Y} \subseteq Y^{(\mathbb{T})}$ (the ***admissible output signals***), and
(vi) $\mathscr{B} \subseteq \mathscr{U} \times \mathscr{Y}$ (the ***behaviours*** of the system) has the property that, if $(\mu, \eta) \in \mathscr{B}$, then $\mathrm{dom}(\mu) = \mathrm{dom}(\eta)$. •

Note that, for a general time system $\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$, there is the associated general input/output system $(\mathscr{U}, \mathscr{Y}, \mathscr{B})$. In Figure 2.3 we depict how one can

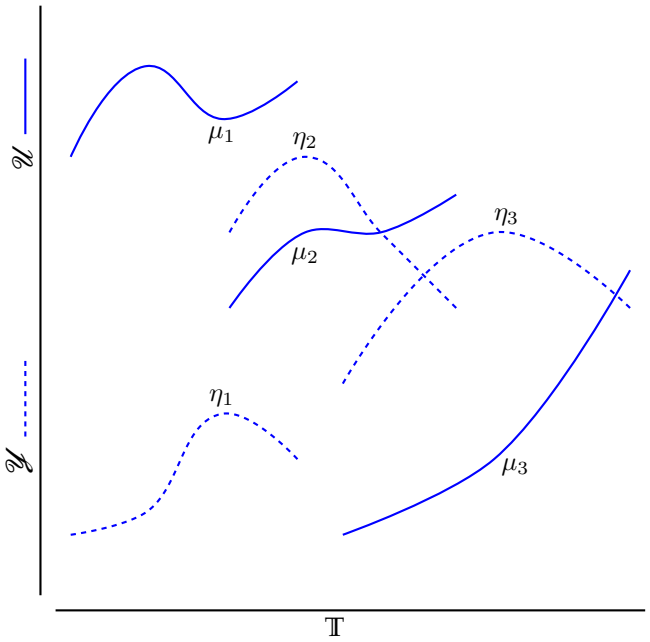


Figure 2.3 Depiction of general time system with behaviours defined on varying domains

think of a general time system.

General time systems can be restricted to sub-time-domains.

2.2.10 Definition (Restriction of a general time system) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system and let $\mathbb{S} \subseteq \mathbb{T}$ be a sub-time-domain. The ***restriction*** of Σ to $\mathbb{S}$ is the general time system

$$\Sigma_{\mathbb{S}} = (U, Y, \mathbb{S}, \mathscr{U}_{\mathbb{S}}, \mathscr{Y}_{\mathbb{S}}, \mathscr{B}_{\mathbb{T}}),$$

where

$$\mathscr{B}_{\mathbb{S}} = \{(\mu|(\mathbb{S} \cap \mathrm{dom}(\mu)), \eta|(\mathbb{S} \cap \mathrm{dom}(\eta))) \mid (\mu, \eta) \in \mathscr{B}\}.$$ •

We shall find it convenient to use some abbreviations to match those above for restrictions of partial time functions:

$$\begin{aligned}\mathscr{B}_{<t} &= \mathscr{B}_{\mathbb{T}_{<t}}, & \mathscr{B}_{>t} &= \mathscr{B}_{\mathbb{T}_{>t}}, & \Sigma_{<t} &= \Sigma_{\mathbb{T}_{<t}}, & \Sigma_{>t} &= \Sigma_{\mathbb{T}_{>t}},\\ \mathscr{B}_{\leq t} &= \mathscr{B}_{\mathbb{T}_{\leq t}}, & \mathscr{B}_{\geq t} &= \mathscr{B}_{\mathbb{T}_{\geq t}}. & \Sigma_{\leq t} &= \Sigma_{\mathbb{T}_{\leq t}}, & \Sigma_{\geq t} &= \Sigma_{\mathbb{T}_{\geq t}}.\end{aligned}$$

Note that we do not require these to be subsets of $\mathscr{B}$. That is to say, restricted behaviours need not be behaviours of the original system, but are behaviours of some new system.

Let us consider some examples of time systems.

2.2.11 Examples (General time system)

1. Let us consider a modification of the mass in a gravitational field from Example 2.1.2–1. In that previous example, we had considered the parameters for mass and gravitational acceleration as being inputs. Here we think of these as being fixed, and instead apply a force f to the mass which we think of as being the input.

   Thus we take $U = \mathbb{R}$ (the set where the force f takes its values), $Y = \mathbb{R}$ (the set of positions for the mass), $\mathbb{T} = \mathbb{R}_{\geq 0}$ (the set of times), $\mathscr{U} = \mathsf{L}^1_{\mathrm{loc}}((\mathbb{R}_{\geq 0}); \mathbb{R})$, and

   $$\mathscr{Y} = \{\xi \in C^1((\mathbb{R}_{\geq 0}); \mathbb{R}) \mid \dot{\xi} \text{ is locally absolutely continuous}\}$$

   (in Theorem 3.2.8 we shall see why this is the right space of outputs). The behaviours for the system are then given by

   $$\mathscr{B} = \{(f, \xi) \in \mathscr{U} \times \mathscr{Y} \mid m\ddot{\xi}(t) = -ma_g + f(t), \text{ a.e. } t \in \operatorname{dom}(f)\}$$

   (again, we shall see in Theorem 3.2.8 why we can only ask that the equality hold almost everywhere).

   Note that there is no *a priori* reason to not allow the domain for inputs and outputs to be any sub-time-domain of $\mathbb{R}_{\geq 0}$.
2. We give the general setting for Example 2.1.2–2. A ***deterministic finite state automaton*** is a quintuple $(Q, Y, \Lambda, \delta, \gamma)$ where
   (a) Q is a finite set (the ***state space***),
   (b) Y is a finite set (the ***output space***),
   (c) Λ is a finite set (the ***alphabet***),
   (d) $\delta\colon Q \times \Lambda \to Q$ (the ***transition function***), and
   (e) $\gamma\colon Q \to Y$ (the ***output function***).

   To consider this as a general time system, we take $\mathbb{T} = \mathbb{Z}_{\geq 0}$. An input signal is then a partial time function $\mu\colon \operatorname{dom}(\mu) \to \Lambda$, where $\operatorname{dom}(\mu) \subseteq \mathbb{Z}_{\geq 0}$ is a sub-time-domain of the form $\operatorname{dom}(\mu) = \{0, 1, \ldots, k\}$ or $\operatorname{dom}(\mu) = \mathbb{T}$. Thus inputs will be either finite sequences from the alphabet (in the case when $\operatorname{dom}(\mu) = \{0, 1, \ldots, k\}$) or infinite sequences in the alphabet (in the case when $\operatorname{dom}(\mu) =$

$\mathbb{Z}_{\geq 0}$). Corresponding to an input μ, we have a mapping $\theta\colon \operatorname{dom}(\mu) \to Q$ defined recursively by prescribing some initial $\theta(0) \in Q$, and then

$$\theta(t+1) = \delta(\theta(t), \mu(t)).$$

In turn, this determines a mapping $\eta\colon \operatorname{dom}(\mu) \to Y$ by $\eta(t) = \gamma \circ \theta(t)$.

We then have the set of inputs $\mathscr{U} \subseteq U^{(\mathbb{Z}_{\geq 0})}$ of partial time functions with domains as above and the set of outputs $\mathscr{Y} \subseteq Y^{(\mathbb{Z}_{\geq 0})}$ determined by the rules described above. This then defines the set $\mathscr{B} \subseteq \mathscr{U} \times \mathscr{Y}$ of behaviours.

3. The binary encoder and decoder can be rendered a general time system by allowing the inputs to vary as a function of discrete time $\mathbb{T} = \mathbb{Z}_{\geq 0}$. Let us work this out.

   For the encoder, we take

   $$U_{\text{enc}} = \{0,1\}^{\{0,1,\dots,2^n-1\}}, \quad Y_{\text{enc}} = \{0,1\}^{\{0,1,\dots,n-1\}}.$$

   The space of admissible inputs will be

   $$\mathscr{U}_{\text{enc}} = \{\mu \in U^{(\mathbb{Z}_{\geq 0})} \mid \mu(t) \text{ is one hot for every } t \in \operatorname{dom}(\mu)\}$$

   while the space of admissible outputs will be

   $$\mathscr{Y}_{\text{enc}} = Y^{(\mathbb{Z}_{\geq 0})}.$$

   Here it is natural to restrict the sub-time-domains of inputs and outputs to be of the form $\mathbb{S} = \{0,1,\dots,k\}$ for some $k \in \mathbb{Z}_{\geq 0}$. Then we have $(\mu,\eta) \in \mathscr{B}_{\text{enc}}$ if $\mu(t)$ and $\eta(t)$ satisfy the condition prescribed in Example 2.1.2–3.

   Of course, the construction adapts in the obvious way to the case of a binary decoder.

   If one thinks of this as a way of converting keyboard output into strings of 0's and 1's, then this is a way of converting text strings typed into a keyboard into a string of 0's and 1's.

4. The convolution system of Example 2.1.14 is a general time system with $\mathbb{T} = \mathbb{R}_{\geq 0}$, $U = Y = \mathbb{R}$, and $\mathscr{U} = \mathscr{Y} = \mathsf{L}^1_{\text{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$. •

The definition in this section of a general time system is not one that one can work with in practice simply because it is too general, and we shall need to add detail to it to arrive at something useful.

2.2.4 Completeness of general time systems

Let us begin with a few constructions concerning the compatibility of the admissible outputs with the admissible inputs.

2.2.12 Definition (Completeness, output completeness) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system.

(i) The system Σ is ***complete*** if the following are satisfied:
  (a) $\mathscr{U}$ is extendible;
  (b) if $(\mu, \eta) \in \mathscr{B}$ and if μ' is an extension of μ to $\mathbb{T}$, then there exists $\eta' \in \mathscr{Y}$ for which $(\mu', \eta') \in \mathscr{B}$.

(ii) The system Σ is ***output complete*** if, for any $\mu \in \mathscr{U}$ and for any family $(\eta_i)_{i\in I}$ in $\mathscr{Y}$ with the following properties:
  (a) I is a totally ordered set with partial order denoted by $\preceq$;[6]
  (b) $\operatorname{dom}(\eta_i) \subseteq \operatorname{dom}(\eta_j)$ if $i \preceq j$;
  (c) $\cap_{i\in I} \operatorname{dom}(\eta_i) \neq \emptyset$;
  (d) $\operatorname{dom}(\eta_i) \subseteq \operatorname{dom}(\mu)$, $i \in I$;
  (e) $(\mu_{\operatorname{dom}(\eta_i)}, \eta_i) \in \mathscr{B}$, $i \in I$;
  (f) there exists $\eta \in Y^{(\mathbb{T})}$ with $\operatorname{dom}(\eta) = \mathbb{S} \triangleq \cup_{i\in I} \operatorname{dom}(\eta_i)$ such that $\eta_{\operatorname{dom}(\eta_i)} = \eta_i$, $i \in I$,

  then $(\mu_{\mathbb{S}}, \eta) \in \mathscr{B}$. •

The idea of completeness is that all behaviours can be extended to be defined on the full time domain $\mathbb{T}$. This might not happen for two reasons. First, it may be the case that $\mathscr{U}$ is itself not extendible. Typically this is not the case, but an unwise choice of inputs might cause this to happen. Second, even if $\mathscr{U}$ is extendible, the outputs that appear as behaviours may not be. This has to do with the system itself; we refer the reader to Example 2.2.21 as an instance of this. This attribute is a true assumption in the sense that there are reasonable systems that do not have this property. In Figure 2.4 we depict how one can think about completeness.

For output completeness, let us explain the significance of the various conditions in the definition. The conditions (a) and (b) mean that the domains of the outputs $(\eta_i)_{i\in I}$ are "nested," while condition (c) means that there is a sub-time-domain contained in all of the domains. Condition (d) means that each of the outputs η_i is defined on a sub-time-domain of $\operatorname{dom}(\mu)$, while condition (e) means that the input μ, restricted to $\operatorname{dom}(\eta_i)$, gives a behaviour associated to η_i. The condition (f) is a compatibility condition on the outputs $(\eta_i)_{i\in I}$, declaring that, for every $i, j \in I$, η_i and η_j agree on the intersection of their domains, and agree with a function $\eta\colon \mathbb{S} \to Y$. Given all of this, the condition of output completeness is that η is, like all of the

[6] Given the remaining conditions, this condition is made without loss of generality. Indeed, we could instead *induce* a total order on I by requiring that $i \preceq j$ if $\operatorname{dom}(\eta_i) \subseteq \operatorname{dom}(\eta_j)$. The essential feature if that the domains $\operatorname{dom}(\eta_i)$, $i \in I$, should be "nested."

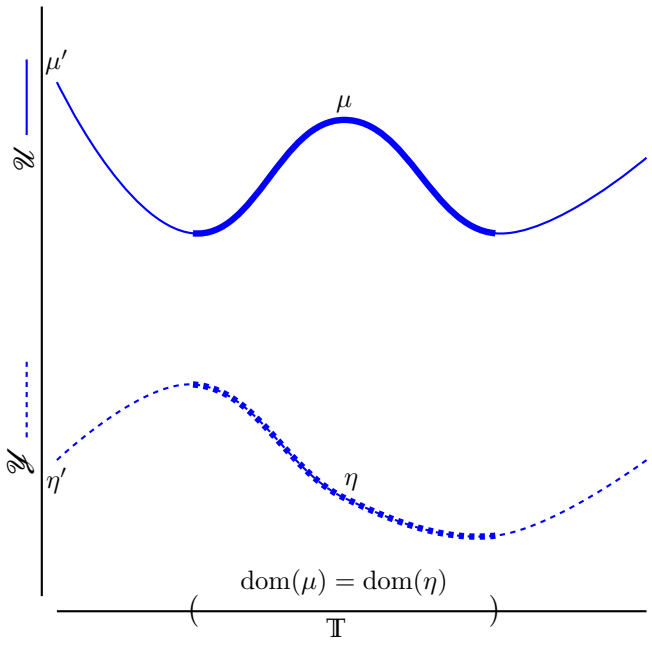


Figure 2.4 A depiction of completeness, showing the extension of an input μ to μ' on the full time domain and the existence of a corresponding output η' agreeing with η on the original domain

outputs η_i, $i \in I$, an output corresponding to an appropriate restriction of the input μ. Naïvely, one can think of this condition as being that

$$\left(\mu_S, \lim_{i \in I} \eta_i\right) \in \mathscr{B},$$

hence the "completeness" terminology. The attribute of output completeness is one that will be possessed by most systems of interest. In Figure 2.5 we depict how one can think about output completeness.

Let us consider completeness and output completeness in a few examples.

2.2.13 Examples (Completeness, output completeness)

1. For the falling mass subject to a force from Example 2.2.11–1, we claim that it is complete and output complete.
   For completeness, first note that, if $f \in \mathscr{U} = \mathsf{L}^1_{\mathrm{loc}}((\mathbb{T}); \mathbb{R})$, then f can be extended to a locally integrable signal on $\mathbb{R}_{\geq 0}$ in many different ways, e.g., by taking it to be zero off $\mathrm{dom}(f)$. Now suppose that $f \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$ and if $(f, \xi) \in \mathscr{B}$, then the relation $\ddot{\xi}(t) = -a_g + f(t)$ directly gives

$$\xi(t) = \xi(0) + \dot{\xi}(0)t - \frac{1}{2}a_g t^2 + \int_0^t \left(\int_0^s f(\tau)\, \mathrm{d}\tau \right) \mathrm{d}s.$$

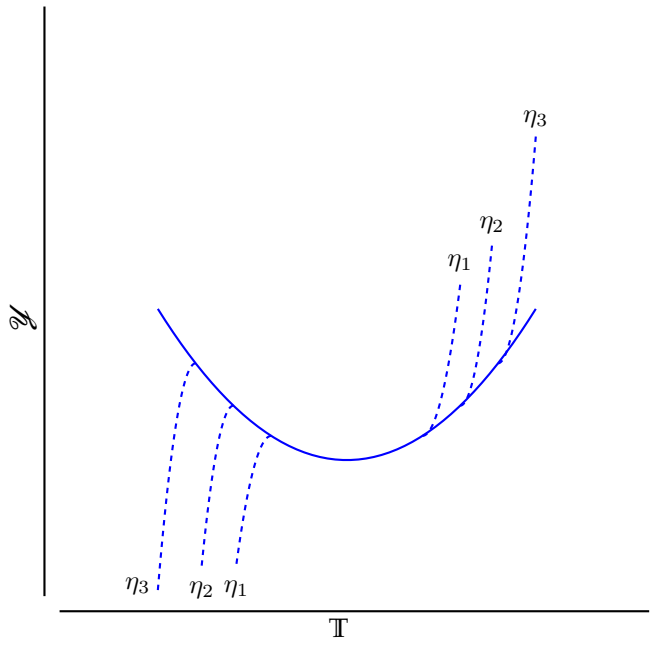


Figure 2.5 A depiction of output completeness with a family of outputs with nested domains giving rise to an output defined on the union of their domains

The indefinite integral

$$\int_0^s f(\tau)\,\mathrm{d}\tau$$

exists for all $s \in \mathbb{R}_{\geq 0}$ since f is locally integrable. Moreover, as a function of s, the resulting function is locally absolutely continuous, and so continuous. Thus the indefinite iterated integral

$$\int_0^t \left(\int_0^s f(\tau)\,\mathrm{d}\tau \right) \mathrm{d}s$$

exists for all $t \in \mathbb{R}_{\geq 0}$. This gives completeness.

For output completeness, let $f \in \mathrm{L}^1_{\mathrm{loc}}(\mathrm{dom}(f); \mathbb{R})$, let $(I, \preceq)$ be a totally ordered set, and let $(\xi_i)_{i \in I}$ be a family of outputs satisfying conditions (a)–(f) of Definition 2.2.12. Note that

$$\xi_i(t) = \xi_i(0) + \dot{\xi}_i(0)t - \frac{1}{2}a_g t^2 + \int_0^t \left(\int_0^s f(\tau)\,\mathrm{d}\tau \right) \mathrm{d}s, \qquad i \in I,\ t \in \mathrm{dom}(\xi_i).$$

Now let $\mathbb{S} = \cup_{i \in I} \mathrm{dom}(\xi_i)$ and let $\xi \colon \mathbb{S} \to \mathbb{R}$ be such that $\xi_{\mathrm{dom}(\xi_i)} = \xi_i$ for $i \in I$. If

$t \in \mathbb{S}$, then $t \in \operatorname{dom}(\xi_i)$ for some $i \in I$, and, therefore,

$$\begin{aligned}\xi(t) = \xi_i(t) &= \xi_i(0) + \dot{\xi}_i(0)t - \frac{1}{2}a_g t^2 + \int_0^t \left(\int_0^s f(\tau)\,\mathrm{d}\tau \right) \mathrm{d}s \\ &= \xi(0) + \dot{\xi}(0)t - \frac{1}{2}a_g t^2 + \int_0^t \left(\int_0^s f(\tau)\,\mathrm{d}\tau \right) \mathrm{d}s.\end{aligned}$$

This shows that $(f_{\mathbb{S}}, \xi) \in \mathscr{B}$, giving output completeness.

2. We claim that a deterministic finite state automaton is complete and output complete.
   Completeness is easy to see. Indeed, any input defined on a sub-time-domain can be extended to the entire time-domain. Also, if an input is defined on the entire time-domain, there is associated to it an admissible output, just because the procedure for producing an output is recursive.
   To see that the deterministic finite state automaton is also output complete, let μ be an admissible input and let $(\eta_i)_{i\in I}$ be a family of admissible outputs satisfying conditions (a)–(f) of Definition 2.2.12. Suppose first that $\mathbb{S} = \cup_{i\in I} \operatorname{dom}(\eta_i)$ is bounded. Then, since the time-domain is discrete and since all sub-time-domains $\operatorname{dom}(\eta_i)$ are of the form $\{0, 1, \dots, k_i\}$, there is some $i_* \in I$ such that $\cup_{i\in I} \operatorname{dom}(\eta_i) = \operatorname{dom}(\eta_{i_*})$. If we take $\eta = \eta_{i_*}$, we have $(\mu_{\mathbb{S}}, \eta) \in \mathscr{B}$.
   Next suppose that $\mathbb{S}$ is not bounded, implying that $\operatorname{dom}(\mu) = \mathbb{Z}_{\geq 0}$. Let $\Psi\colon Q \to Y^{\mathbb{Z}_{\geq 0}}$ be the mapping defined by asking that $\Psi(q)(t)$ to be the output at time t associated with choosing the initial state to satisfy $\theta(0) = q$. For $t \in \mathbb{Z}_{\geq 0}$, define an equivalence relation $\sim_t$ on Q by

$$q_1 \sim_t q_2 \iff \Psi(q_1)(t') = \Psi(q_2)(t') \text{ for } t' \leq t.$$

   Note that, if $t_1 \leq t_2$, then

$$q_1 \sim_{t_2} q_2 \implies q_1 \sim_{t_1} q_2$$

   Thus the number of equivalence classes associated with the equivalence relation $\sim_t$ increases with t. However, since Q is finite, there is a maximum number of equivalence classes. This means that there exists $t_* \in \mathbb{Z}_{\geq 0}$ such that, if, for all $q_1, q_2 \in Q$ and for all $t \geq t_*$, if $\Psi(q_1)(t') = \Psi(q_2)(t')$ for all $t' \leq t$, then it holds that $\Psi(q_1)(t) = \Psi(q_2)(t)$ for all $t \in \mathbb{Z}_{\geq 0}$. In particular, if we take $i_* \in I$ so that $\sup \operatorname{dom}(\eta_{i_*}) \geq t_*$, then taking $\eta = \eta_{i_*}$ gives $(\mu, \eta) \in \mathscr{B}$.
3. An entirely similar argument as we just saw for the finite state automaton shows that the binary encoder and binary decoder, as general time systems, are output complete. Note that, with the restriction that admissible inputs be of finite length in time, these systems are not complete since no input can be extended to an admissible input defined on the entire time-domain. •

2.2.5 Dynamical system representations and state space representations

In this section we give some important structure to general time systems, structure closely related to the notion of a state object and a response function explored in Section 2.1.3 for general input/output systems. Here the ideas are adapted in particular ways to account for time.

The first step in this process consists in making the following definition.

2.2.14 Definition (Initial state object, initial response function, subsequent response function, response family) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system and let $t_0 \in \mathbb{T}$.

(i) An ***initial response function*** for Σ at t_0 with ***initial state object*** $X^\Sigma_{t_0}$ is a map

$$\rho^\Sigma_{t_0} \colon X^\Sigma_{t_0} \times \mathscr{U}_{\geq t_0} \to \mathscr{Y}_{\geq t_0}$$

such that $(\mu, \eta) \in \mathscr{B}_{\geq t_0}$ if and only if there exists $x_{t_0} \in X^\Sigma_{t_0}$ such that $\rho^\Sigma_{t_0}(x_{t_0}, \mu) = \eta$.

(ii) A ***subsequent response function*** for Σ at $t > \mathbb{T}_{\geq t_0}$ from t_0 with ***subsequent state object*** X^Σ_{t,t_0} is a map

$$\rho^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times (\mathscr{U}_{\geq t_0})_{>t} \to (\mathscr{Y}_{\geq t_0})_{\geq t}$$

such that, if $(\mu, \eta) \in (\mathscr{B}_{\geq t_0})_{\geq t}$, then there exists $x_t \in X^\Sigma_{t,t_0}$ such that $\rho^\Sigma_{t,t_0}(x_t, \mu) = \eta$.

(iii) A ***response family*** for Σ at t_0 is a family $\boldsymbol{\rho}^\Sigma_{t_0} = (\rho^\Sigma_{t,t_0})_{t \in \mathbb{T}_{\geq t_0}}$ of subsequent response functions for Σ from t_0 with associated subsequent state objects $(X^\Sigma_{t,t_0})_{t \in \mathbb{T}_{\geq t_0}}$. •

These definitions are sufficiently subtle as to require some explanation.

1. $\rho^\Sigma_{t_0}$: The idea of an initial response function is rather similar to that for a response function for a general input/output system. A significant additional piece of information is the time t_0, which we think of as an "initial time." Thus, for a given input, $X^\Sigma_{t_0}$ parameterises all outputs as they pass through time t_0.
2. ρ^Σ_{t,t_0}: The idea here is similar, but different, to that for $\rho^\Sigma_{t_0}$. Here we restrict our attention to behaviours at time t that originated from time t_0.
3. $\boldsymbol{\rho}^\Sigma_{t_0}$: This gathers together all of the above.

In Figure 2.6 we illustrate how one might think of the initial and subsequent response functions.

We now introduce another important player in state descriptions of general time systems.

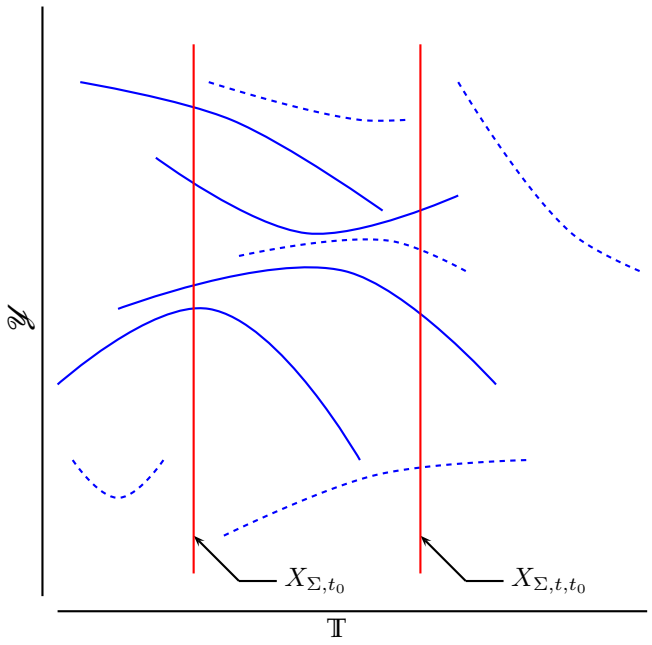


Figure 2.6 A depiction of initial and subsequent response functions for a fixed input, with the solid lines denoting outputs that pass through time t_0

2.2.15 Definition (State transition family) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system, let $t_0 \in \mathbb{T}$, and let $(X_{t,t_0})_{t \in \mathbb{T}_{\geq t_0}}$ be a family of sets. A ***family of state transition maps*** from t_0 is a collection $\Phi_{t_2,t_1} : \mathscr{U}_{[t_1,t_2)} \times X_{t_1,t_0} \to X_{t_2,t_0}$ with

(i) $t_1, t_2 \in \mathbb{T}_{>t_0}$, for $t_1 \leq t_2$,

(ii) $\Phi_{t,t}(\mu_{[t,t)}, x_t) = x_t$, $t \in \mathbb{T}_{>t_0}$, for $x_t \in X_{t,t_0}$,

(iii) $\Phi_{t_3,t_2}(\mu_{[t_2,t_3)}, \Phi_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1})) = \Phi_{t_3,t_1}(\mu_{[t_1,t_3)}, x_{t_1})$, for $t_1, t_2, t_3 \in \mathbb{T}_{>t_0}$, $t_1 \leq t_2 \leq t_3$. •

The idea of a family of state transition maps is that state objects at time t_1 are mapped to state objects at t_2 for $t_1 \leq t_2$. The composition property (iii) is called the ***semigroup property*** since it indicates how the different state transition maps interact with the forward movement of time. In Figure 2.7 we depict how one should think of a family of state transition maps.

Of course, a family of state transition maps is, itself, a meaningless thing for describing the system, since the system appears nowhere in the properties of the maps. To pull this all together, one needs a compatibility condition between a response family and a family of state transition maps.

2.2.16 Definition ((Pre-)dynamical system representation) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

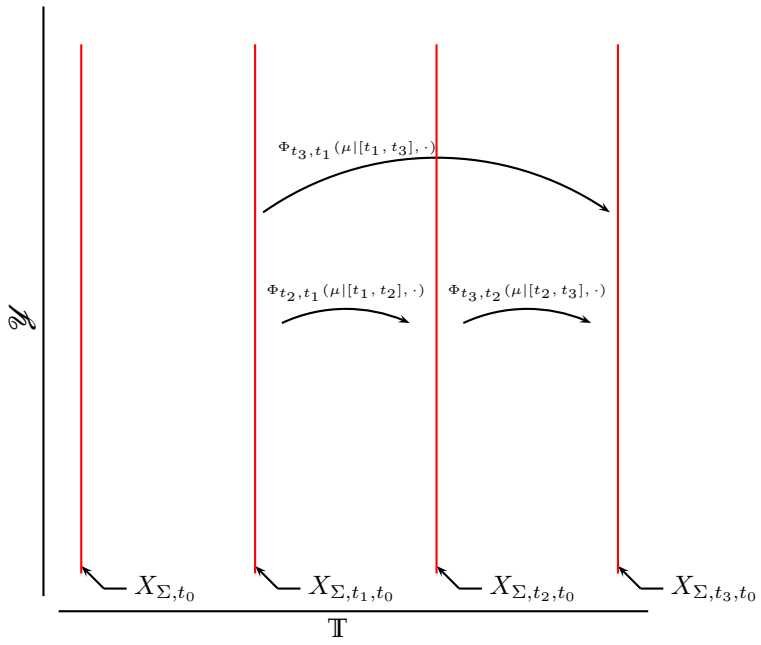


Figure 2.7 A depiction of a family of state transition maps, with the maps for a fixed input being depicted

be a general time system, let $t_0 \in \mathbb{T}$, let $(X^\Sigma_{t,t_0})_{t\in\mathbb{T}_{\geq t_0}}$, let

$$\rho^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{\geq t_0})_{\geq t}, \qquad t \in \mathbb{T}_{\geq t_0},$$

be a subsequent response family for Σ at t_0, and let

$$\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma_{t_1,t_0} \to X^\Sigma_{t_2,t_0}, \qquad t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2,$$

be a family of state transition maps.

(i) The subsequent response family and the family of state transition maps are ***compatible*** if

$$\rho^\Sigma_{t_1,t_0}(x_{t_1}, \mu_{\geq t_1})_{\geq t_2} = \rho^\Sigma_{t_2,t_0}(\Phi^\Sigma_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1}), \mu_{\geq t_2}).$$

(ii) A ***pre-dynamical system representation*** for Σ at t_0 consists of the three pieces of data

$$\begin{aligned}
&X^\Sigma_{t,t_0}, && t \in \mathbb{T}_{\geq t_0},\\
&\rho^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0},\\
&\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma_{t_1,t_0} \to X^\Sigma_{t_2,t_0}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2,
\end{aligned}$$

where the subsequent response family ρ^Σ_{t,t_0}, $t \in \mathbb{T}_{\geq t_0}$, and the family of state transition maps $\Phi^\Sigma_{t_2,t_1}$, $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, $t_1 \leq t_2$, is compatible. •

Let us discuss the matter of existence of pre-dynamical system representations.

2.2.17 Theorem (Existence of pre-dynamical system representations) *If*

$$\Sigma = (\mathrm{U}, \mathrm{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

is a complete general time system for which $\mathscr{U}$ is closed under concatenation and for which $\mathrm{dom}(\Sigma) = \mathscr{U}$*, then Σ possesses a pre-dynamical system representation.*

Proof Since Σ is complete, we can assume that every behaviour is the restriction of some behaviour defined on all of $\mathbb{T}$. For this reason, we shall take, without loss of generality,

$$\mathscr{U} \subseteq U^{\mathbb{T}}, \ \mathscr{Y} \subseteq Y^{\mathbb{T}},$$

i.e., we consider only behaviours defined on the entirety of the time-domain. In this case

$$(\mathscr{U}_{\geq t_0})_{\geq t} = \mathscr{U}_{\geq t}, \ (\mathscr{Y}_{\geq t_0})_{\geq t} = \mathscr{Y}_{\geq t}, \ (\mathscr{B}_{\geq t_0})_{\geq t} = \mathscr{B}_{\geq t}.$$

Since $\mathrm{dom}(\Sigma) = \mathscr{U}$, $\mathrm{dom}(\Sigma_{\geq t}) = \mathscr{U}_{\geq t}$ for every $t \in \mathbb{T}_{\geq t_0}$.

Define

$$X^{\Sigma}_{t,t_0} = \{x_t \colon \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t} \mid \ \mathrm{graph}(x_t) \subseteq \mathscr{B}_{\geq t}\}.$$

We claim that, for any $t \in \mathbb{T}_{\geq t_0}$ and for any $(\mu_{\geq t}, \eta_{\geq t}) \in \mathscr{B}_{\geq t}$, there exists $x_t \in X^{\Sigma}_{t,t_0}$ such that $\eta_{\geq t} = x_t(\mu_{\geq t})$. Thus, for $\mu'_{\geq t} \in \mathscr{U}_{\geq t}$, there exists $\eta'_{\geq t} \in \mathscr{Y}_{\geq t}$ such that $(\mu'_{\geq t}, \eta'_{\geq t}) \in \mathscr{B}_{\geq t}$. Therefore, by the Axiom of Choice, there is a choice function

$$\mathscr{U}_{\geq t} \ni \mu'_{\geq t} \mapsto \eta'_{\geq t} \in \mathscr{Y}_{\geq t};$$

denote this map by x'_t. Now define $x_t \colon \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t}$ by

$$x_t(\mu'_{\geq t}) = \begin{cases} x'_t(\mu'_{\geq t}), & \mu'_{\geq t} \neq \mu_{\geq t}, \\ \eta_{\geq t}, & \mu'_{\geq t} = \mu_{\geq t}. \end{cases}$$

Clearly $x_t \in X^{\Sigma}_{t,t_0}$ and $x_t(\mu_{\geq t}) = \eta_{\geq t}$.

Now define

$$\rho^{\Sigma}_{t,t_0} \colon X^{\Sigma}_{t,t_0} \times \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t}$$
$$(x_t, \mu_t) \mapsto x_t(\mu_t).$$

Note that, if $(\mu_{\geq t}, \eta_{\geq t}) \in \mathscr{B}_{\geq t}$, then (as we just showed) there exists $x_t \in X^{\Sigma}_{t,t_0}$ such that

$$x_t(\mu_{\geq t}) = \eta_{\geq t} \implies \rho^{\Sigma}_{t,t_0}(x_t, \mu_{\geq t}) = \eta_{\geq t}.$$

Conversely, if $x_t \in X^{\Sigma}_{t,t_0}$ and if $\eta_{\geq t} = \rho^{\Sigma}_{t,t_0}(x_t \mu_{\geq t})$, then $x_t(\mu_{\geq t}) = \eta_{\geq t}$, and so $(\mu_{\geq t}, \eta_{\geq t}) \in \mathscr{B}_{\geq t}$. Thus ρ^{Σ}_{t,t_0}, $t \in \mathbb{T}_{\geq t_0}$, is a subsequent response family.

Next, for $t_1, t_2 \in \mathbb{T}_{\geq t_0}$ with $t_1 \leq t_2$, define

$$\Phi^{\Sigma}_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^{\Sigma}_{t_1,t_0} \to X_{t_2,t_0}$$

by

$$\Phi^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1})(\mu'_{\geq t_2}) = (x_{t_1}(\mu_{[t_1,t_2)} * \mu'_{\geq t_2}))_{\geq t_2}.$$

Let us verify that this potential family of state transition maps satisfies the compatibility condition with the subsequent response family defined above. Let $(\mu_{\geq t_1}, \eta_{\geq t_1}) \in \mathscr{B}_{\geq t_1}$ and compute

$$\begin{aligned}\rho^{\Sigma}_{t_1,t_0}(x_{t_1}, \mu_{\geq t_1})_{\geq t_2} &= (x_{t_1}(\mu_{\geq t_1}))_{\geq t_2}\\ &= (x_{t_1}(\mu_{[t_1,t_2)} * \mu_{\geq t_2})_{\geq t_2}\\ &= \Phi^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1})(\mu_{\geq t_2})\\ &= \rho^{\Sigma}_{t_2,t_0}(\Phi^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1}), \mu_{\geq t_2}),\end{aligned}$$

as desired. Next we show that the mappings

$$\Phi^{\Sigma}_{t_2,t_1}, \qquad t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2,$$

define a family of state transition maps. First of all, since $\mu_{[t,t)} * \mu_{\geq t} = \mu_{\geq t}$, we have that $\Phi^{\Sigma}_{t,t}(\mu_{\geq t}, x_t) = x_t$. Also,

$$\begin{aligned}\Phi^{\Sigma}_{t_3,t_2}(\mu_{[t_2,t_3)}, \Phi^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1}))(\mu'_{\geq t_3}) &= (\Phi^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1})(\mu_{[t_2,t_3)} * \mu'_{\geq t_3}))_{\geq t_3}\\ &= ((x_{t_1}(\mu_{[t_1,t_2)} * \mu_{[t_2,t_3)} * \mu'_{\geq t_3}))_{\geq t_2})_{\geq t_3}\\ &= (x_{t_1}(\mu_{[t_1,t_3)} * \mu'_{\geq t_3}))_{\geq t_3}\\ &= \Phi^{\Sigma}_{t_3,t_1}(\mu_{[t_1,t_3)}, x_{t_1})(\mu'_{\geq t_3}),\end{aligned}$$

as desired. ■

While cute, the theorem is quite useless.

2.2.18 Remark (On the existence of pre-dynamical system representations) Here is a short list of reasons why the theorem gets you nowhere in practice.

1. It is not constructive. In the proof we use the Axiom of Choice, always the hallmark of an unsatisfying proof.
2. There are not well-defined states. The state object parameterising outputs at time t vary as t varies. This means, for example, that one cannot compare states at different times.
3. In practice, representations have structure. For specific and interesting classes of systems, there are typically structured state objects, subsequent response families, and state transition maps. We shall see this as we add more structure to our systems. •

Let us address, albeit in an unsatisfying way, the issue pointed out above about not being able to compare states at different times. The following definition captures this scenario.

2.2.19 Definition (Dynamical system representation) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system. A pre-dynamical system representation for Σ at t_0 prescribed by the data

$$\begin{aligned}
&X^\Sigma_{t,t_0}, && t \in \mathbb{T}_{\geq t_0},\\
&\rho^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0},\\
&\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma_{t_1,t_0} \to X^\Sigma_{t_2,t_0}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2,
\end{aligned}$$

is a ***dynamical system representation*** if there exists a set X^Σ such that $X^\Sigma = X^\Sigma_{t,t_0}$, $t \in \mathbb{T}_{\geq t_0}$. •

Let us show that dynamical system representations exist.

2.2.20 Theorem (Existence of dynamical system representations) *If*

$$\Sigma = (\mathrm{U}, \mathrm{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

is a complete general time system for which $\mathscr{U}$ is closed under concatenation and for which $\mathrm{dom}(\Sigma) = \mathscr{U}$*, then Σ possesses a dynamical system representation.*

Proof We make the same notational simplifications resulting from completeness as we did at the beginning of the proof of Theorem 2.2.17.

By Theorem 2.2.17, we suppose that we have a pre-dynamical system representation prescribed by the data

$$\begin{aligned}
&X^\Sigma_{t,t_0}, && t \in \mathbb{T}_{\geq t_0},\\
&\overline{\rho}^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t}, && t \in \mathbb{T}_{\geq t_0},\\
&\overline{\Phi}^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma_{t_1,t_0} \to X^\Sigma_{t_2,t_0}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2,
\end{aligned}$$

We let $X^\Sigma = \mathring{\cup}_{t \in \mathbb{T}} X^\Sigma_{t,t_0}$ and, for each $t \in \mathbb{T}$, arbitrarily choose $x^*_t \in X^\Sigma_{t,t_0}$. Define

$$\rho^\Sigma_{t,t_0} \colon X^\Sigma \times \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t}$$

by

$$\rho^\Sigma_{t,t_0}((t', x_{t'}), \mu_{\geq t}) = \begin{cases} \overline{\rho}^\Sigma_{t,t_0}(x_t, \mu_{\geq t}), & t' = t,\\ \overline{\rho}^\Sigma_{t,t_0}(x^*_t, \mu_{\geq t}), & t' \neq t. \end{cases}$$

Let us show that this defines a family of subsequent response functions. Let $t \in \mathbb{T}_{\geq t_0}$ and let $(\mu_{\geq t}, \eta_{\geq t}) \in \mathscr{B}_{\geq t}$. Then there exists $x_t \in X^\Sigma_{t,t_0}$ such that

$$\eta_{\geq t} = \overline{\rho}^\Sigma_{t,t_0}(x_t, \mu_{\geq t}) = \rho^\Sigma_{t,t_0}((t, x_t), \mu_{\geq t}).$$

Conversely, suppose that $\eta_{\geq t} = \rho^\Sigma_{t,t_0}((t', x_{t'}), \mu_{\geq t})$ for $(t', x_{t'}) \in X^\Sigma$. The definition of ρ^Σ_{t,t_0} then gives $(\mu_{\geq t}, \eta_{\geq t}) \in \mathscr{B}_{\geq t}$, showing that ρ^Σ_{t,t_0} is a response family.

Next, for $t_1, t_2 \in \mathbb{T}$ with $t_1 \leq t_2$, define

$$\Phi^{\Sigma}_{t_2,t_1} : \mathscr{U}_{[t_1,t_2)} \times X^{\Sigma} \to \mathscr{Y}_{\geq t}$$

by

$$\Phi^{\Sigma}_{t_2,t_1}\Big(\mu_{[t_1,t_2)}, (t', x_{t'})\Big) = \begin{cases} (t_2, \overline{\Phi}^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1})), & t' = t_1, \\ \Big(t_2, \overline{\Phi}^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x^*_{t_1})\Big), & t' \neq t_1. \end{cases}$$

We then have

$$\Phi^{\Sigma}(\mu_{[t,t)}, (t, x_t)) = \Big(t, \overline{\Phi}^{\Sigma}_{t_2,t_1}(\mu_{[t,t)}, x_t)\Big) = (t, x_t).$$

We also compute

$$\begin{aligned} \Phi^{\Sigma}_{t_3,t_2}(\mu_{[t_2,t_3)}, \Phi^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, (t', x_{t'}))) &= \Phi^{\Sigma}_{t_3,t_2}(\mu_{[t_2,t_3)}, \Phi^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, (t_1, x'_{t_1}))) \\ &= \Phi^{\Sigma}_{t_3,t_2}\Big(\mu_{[t_2,t_3)}, \Big(t_2, \overline{\Phi}^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x'_{t_1})\Big)\Big) \\ &= \Big(t_3, \overline{\Phi}^{\Sigma}_{t_3,t_2}\Big(\mu_{[t_2,t_3)}, \overline{\Phi}^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x'_{t_1})\Big)\Big) \\ &= \Big(t_3, \overline{\Phi}^{\Sigma}_{t_3,t_1}(\mu_{[t_1,t_3)}, x'_{t_1}\Big) \\ &= \Phi^{\Sigma}_{t_3,t_1}(\mu_{[t_1,t_3)}, (t', x_{t'})), \end{aligned}$$

where

$$x'_{t_1} = \begin{cases} x_{t_1}, & t' = t_1, \\ x^*_{t_1}, & t' \neq t_1. \end{cases}$$

Finally, we show the compatibility of the family of subsequent response functions and the family of state transition maps. To this end, we compute

$$\begin{aligned} \rho^{\Sigma}_{t_1,t_0}((t', x_{t'}), \mu_{\geq t_1})_{\geq t_2} &= \rho^{\Sigma}_{t_1,t_0}((t_1, x'_{t_1}), \mu_{\geq t_1})_{\geq t_2} \\ &= \Big(\overline{\rho}^{\Sigma}_{t_1,t_0}(x'_{t_1}, \mu_{\geq t_1})\Big)_{\geq t_2} \\ &= \overline{\rho}^{\Sigma}_{t_2,t_0}\Big(\overline{\Phi}^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x'_{t_1}), \mu_{\geq t_2}\Big) \\ &= \rho^{\Sigma}_{t_2,t_0}\Big(\Big(t_2, \overline{\Phi}^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, x'_{t_1})\Big), \mu_{\geq t_2}\Big) \\ &= \rho^{\Sigma}_{t_2,t_0}(\Phi^{\Sigma}_{t_2,t_1}(\mu_{[t_1,t_2)}, (t', x_{t'})), \mu_{\geq t_2}), \end{aligned}$$

where

$$x'_{t_1} = \begin{cases} x_{t_1}, & t' = t_1, \\ x^*_{t_1}, & t' \neq t_1, \end{cases}$$

as desired. ∎

The same criticisms of Remark 2.2.18, with the exception of 2, apply to the preceding theorem.

The assumption of completeness in the preceding theorems is necessary, and indeed a lack of completeness makes any definition of a dynamical system representation such as that in Definition 2.2.16 problematic. The following example illustrates this.

2.2.21 Example (Lack of completeness and nonexistence of pre-dynamical system representations) We define a general time system with the following data:

1. $U = \mathbb{R}$,
2. $Y = \mathbb{R}$,
3. $\mathbb{T} = \mathbb{R}_{\geq 0}$,
4. $\mathscr{U} = \mathsf{L}^1([0,T];\mathbb{R})$, $T \in \mathbb{R}_{>0}$,
5. $\mathscr{Y} = \mathsf{C}^0([0,T];\mathbb{R})$, $T \in \mathbb{R}_{>0}$, and
6. $\mathscr{B} = \{(\mu,\eta) \in \mathscr{U} \times \mathscr{Y} \mid \dot{\eta}(t) = \mu(t)\eta(t)^2, \text{ a.e. } t \in \operatorname{dom}(\mu)\}$.

We take $t_0 = \mathbb{R}$. We can determine that behaviours are given by pairs (μ,η) where $\mu \in \mathsf{L}^1([0,T];\mathbb{R})$ and

$$\eta(t) = \frac{\eta_0}{1 - \eta_0 \int_0^t \mu(\tau)\,\mathrm{d}\tau} \tag{2.4}$$

for some $\eta_0 \in \mathbb{R}$ (note that $\eta(0) = \eta_0$). Thus this initial condition η_0 at $t = 0$ prescribes the solution for any $t \in [0,T]$, and so we can take $X_0^\Sigma = \mathbb{R}$ as parameterising all outputs for every input μ.

Note that this solution will be defined as long as there exists no $t \in [0,T]$ for which $\eta_0 \int_0^t \mu(\tau)\,\mathrm{d}\tau = 1$. To illustrate the point we wish to make, it will suffice to consider the case where μ is a constant function, say $\mu(t) = \mu_0$. In this case, we require that there exist no $t \in [0,T]$ for which $\eta_0\mu_0 t = 1$. The essential observation for our purposes is the following:

> *For every $T \in \mathbb{R}_{>0}$ and every $\eta_0 \in X_0^\Sigma = \mathbb{R}$, there exists $\mu \in \mathsf{L}^1([0,T];\mathbb{R})$ such that, if*
>
> $$\eta(t) = \frac{\eta_0}{1 - \eta_0 \int_0^t \mu(\tau)\,\mathrm{d}\tau},$$
>
> *then $\lim_{t\uparrow T}|\eta(t)| = \infty$. Indeed, we can take $\mu(t) = \frac{1}{\eta_0 T}$.*

The main point of this is that there can be no map

$$\mathscr{U}_{[0,T]} \times \underbrace{X_{t_0}^\Sigma}_{\mathbb{R}} \to \underbrace{X_{T,t_0}^\Sigma}_{\mathbb{R}}$$

that maps the initial condition at $t = 0$ to a final condition at $t = T$ for *every* input. •

The preceding example notwithstanding, the class of systems that will be of most interest to us in this volume—linear systems—will have the completeness property, and so pre-dynamical system and dynamical system representations are worth thinking about.

Let us consider a few examples of dynamical system representations.

2.2.22 Examples (Dynamical system representations)

1. Let us consider the mass in a gravitational field in the guise of Example 2.2.11–1. Let us take $t_0 = 0$ and provide a dynamical system representation for this system by taking
   (a) $X^\Sigma = \mathbb{R}^2$,
   (b) $\rho^\Sigma_{t,0} \colon \mathbb{R}^2 \times \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t}$ is defined by

$$\rho^\Sigma_{t,0}((x_0, v_0), f_{\geq t})(s) = x_0 + (s-t)v_0 - \frac{1}{2}a_g(s-t)^2 + \int_t^s \left(\int_t^\tau f(r)\, \mathrm{d}r \right) \mathrm{d}\tau,$$

   and
   (c) $\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times \mathbb{R}^2 \to \mathbb{R}^2$ is defined by

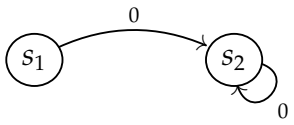
$$\begin{aligned}\Phi^\Sigma_{t_2,t_1}(f_{[t_1,t_2)}, (x_0, v_0)) \\ = \Bigg(x_0 + (t_2 - t_1)v_0 - \frac{1}{2}a_g(t_2 - t_1)^2 + \int_{t_1}^{t_2} \left(\int_{t_1}^s f(\tau)\, \mathrm{d}\tau \right) \mathrm{d}s, \\ v_0 - \frac{1}{2}a_g(t_2 - t_1) + \int_{t_1}^{t_2} f(\tau)\, \mathrm{d}\tau \Bigg).\end{aligned}$$

   The idea here is that the state is (x, v) where x is the position of the mass and v is its velocity. The subsequent response function returns the position of the mass if it starts at time t in a prescribed state and is subject to a force. The state transition map returns the state at time t_2 if one starts in a prescribed state at time t_1.
   We note that the subsequent response functions and the state transition maps are all defined for all states at every time. We shall see in the next example that this is not always the case.
2. We now consider a deterministic finite state automaton $(Q, Y, \Lambda, \delta, \gamma)$ as a general time system as in Example 2.2.11–2. We define a dynamical system representation for this general time system from $t_0 = 0$ by taking
   (a) $X^\Sigma = Q$,
   (b) $\rho^\Sigma_{t,0} \colon Q \times \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t}$ is defined by the specification of the automaton, taking a state at time $s \in \mathbb{Z}_{\geq 0}$ with $s \geq t$ to the output at that time according to the rules.
   (c) In a similar manner, $\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{\geq t} \times Q \to Q$ is defined by the dynamics of the automaton.

   We note that $\rho^\Sigma_{t,0}$, and therefore also $\Phi^\Sigma_{t_2,t_1}$, may not have a meaningful definition for all states in Q. As an example of this, consider the deterministic finite state automaton with

$$Q = \{s_1, s_2\}, \quad Y = \{s_1, s_2\}, \quad \Lambda = \{0\},$$

with γ the identity map, and with dynamics defined by the diagram



(2.5)

In this case, the state will end up at s_2 after at most one time step, and will remain there forever after. Thus the value of $\rho^{\Sigma}_{t,0}$ on state s_1 is immaterial for $t \geq 1$. •

We see that there is an important distinction between the two dynamical system representations in the above examples, so let us characterise this.

2.2.23 Definition (Full pre-dynamical system representation) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system with pre-dynamical system representation at t_0 prescribed by the data

$$\begin{aligned} &X^{\Sigma}_{t,t_0}, && t \in \mathbb{T}_{\geq t_0}, \\ &\rho^{\Sigma}_{t,t_0} : X^{\Sigma}_{t,t_0} \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0}, \\ &\Phi^{\Sigma}_{t_2,t_1} : \mathscr{U}_{[t_1,t_2)} \times X^{\Sigma}_{t_1,t_0} \to X^{\Sigma}_{t_2,t_0}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2. \end{aligned}$$

For $t \in \mathbb{T}_{\geq t_0}$, denote

$$(\overline{\mathscr{B}}_{\geq t_0})_{\geq t} = \{(\mu_{\geq t}, \eta_{\geq t}) \in (\mathscr{B}_{\geq t_0})_{\geq t} \times (\mathscr{Y}_{\geq t_0})_{\geq t} \mid \eta_{\geq t} = \rho^{\Sigma}_{t,t_0}(x_t, \mu_{\geq t}) \text{ for some } x_t \in X^{\Sigma}_{t,t_0}\}.$$

The pre-dynamical system representation is ***full*** if $(\overline{\mathscr{B}}_{\geq t_0})_{\geq t} = (\mathscr{B}_{\geq t_0})_{\geq t}$ for every $t \in \mathbb{T}_{\geq t_0}$. •

One easily sees that the dynamical system representation for the mass system of Example 2.2.22–1 is full, while that for the deterministic finite state automaton from Example 2.2.22–2 is full in some cases (e.g., for the example depicted in (2.1)) but not in others (e.g., for the example depicted in (2.5)).

As our final fleshing out of these more concrete representations of general time systems, we have the following notion.

2.2.24 Definition ((Pre-)state space representation) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system and let $t_0 \in \mathbb{T}$. A ***pre-state space representation*** for Σ at t_0 is prescribed by the data

$$\begin{aligned} &X^{\Sigma}_{t,t_0}, && t \in \mathbb{T}_{\geq t_0}, \\ &\gamma^{\Sigma}_{t,t_0} : X^{\Sigma}_{t,t_0} \times U \to Y, && t \in \mathbb{T}_{\geq t_0}, \\ &\Phi^{\Sigma}_{t_2,t_1} : \mathscr{U}_{[t_1,t_2)} \times X^{\Sigma}_{t_1,t_0} \to X^{\Sigma}_{t_2,t_0}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2, \end{aligned}$$

such that

(i) the maps $\Phi^{\Sigma}_{t_1,t_2}$, $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, $t_1 \leq t_2$, are a family of state transition maps and

(ii) $(\mu, \eta) \in \mathscr{B}_{\geq t_0}$ if and only if there exists $x \in X^{\Sigma}_{t_0,t_0}$ such that

$$\eta(t) = \gamma^{\Sigma}_{t,t_0}(\Phi^{\Sigma}_{t,t_0}(\mu_{[t_0,t)}, x), \mu(t))$$

for $t \in \mathbb{T}_{\geq t_0}$.

The maps γ^{Σ}_{t,t_0}, $t \in \mathbb{T}_{\geq t_0}$ are ***output functions***. If there exists a set X^{Σ} such that $X^{\Sigma}_{t,t_0} = X^{\Sigma}$, then the pre-state space representation is a ***state space representation*** with ***state space*** X^{Σ}. •

The idea with a state space representation is that the time evolution is prescribed by the state transition maps, while the output is merely prescribed as a function of the state. Thus the determination of the state involves dynamics, while the determination of the output from state is static, i.e., done for a fixed time.

Both examples we have of dynamical systems representations are, in fact, amenable to state space representations.

2.2.25 Examples (State space representations)

1. For the mass system from Example 2.2.22–1, we have the state space $X^{\Sigma} = \mathbb{R}^2$, the input space $U = \mathbb{R}$, and the output space $Y = \mathbb{R}$, with the output function defined by $\gamma((x_0, v_0), u) = x_0$ giving a state space representation. Thus the output function simply returns the evaluation of the output signal ξ at time t.
2. For the deterministic finite state automaton from Example 2.2.22–2, the state space is $X^{\Sigma} = Q$, the input space is $U = \Lambda$, and the output space is Y. This system has a state space representation if we take γ to simply be the output function defined as part of the system data. •

Many of the systems we consider in detail in this volume will come with natural state space representation; indeed, Chapter 5 is devoted to a detailed study of a class of such systems, while Chapters 7 and 8 consider specialised techniques for the class of these systems that are linear and time-invariant.

It turns out that the existence of a state space representation for a system is intimately connected to a specific character of the system in terms of how it depends on time. We now turn our attention to such matters.

2.2.6 Causality in time systems

As the word "causal" suggests, the idea of a causal system is that inputs (causes) determine outputs (effects). This general idea can be represented in a variety of different, but interconnected, ways. In this section we shall explore these.

2.2.26 Assumption (Completeness assumption and consequences) It is often convenient in this section to work with complete systems, since our notions of causality will be often connected with (pre-)dynamical systems representations. As a consequence, in this situation all behaviours signals are restrictions to their domain

of behaviours defined on the entire time-domain. For this reason, for a complete general time system, we can assume that $\mathscr{U} \subseteq U^{\mathbb{T}}$ and $\mathscr{Y} \subseteq Y^{\mathbb{T}}$. •

Let us start with the initial notions of causality for a system.

2.2.27 Definition (Causal, strongly causal system) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system and let $t_0 \in \mathbb{T}$.

(i) The system Σ is ***causal*** from t_0 if, for any $\mu_1, \mu_2 \in \mathscr{U}$ and for any $t \in \mathbb{T}_{\geq t_0}$, the following implication holds:

$$(\mu_1)_{[t_0,t]} = (\mu_2)_{[t_0,t]} \quad \Longrightarrow \quad \mathscr{B}(\mu_1)_{[t_0,t]} = \mathscr{B}(\mu_2)_{[t_0,t]}.$$

(ii) The system Σ is ***strongly causal*** from t_0 if, for any $\mu_1, \mu_2 \in \mathscr{U}$ and for any $t \in \mathbb{T}_{\geq t_0}$, the following implication holds:

$$(\mu_1)_{[t_0,t)} = (\mu_2)_{[t_0,t)} \quad \Longrightarrow \quad \mathscr{B}(\mu_1)_{[t_0,t]} = \mathscr{B}(\mu_2)_{[t_0,t]}. \quad \bullet$$

The idea, then, of a causal system is that the set of outputs corresponding to inputs that agree up to time t, also agrees up to time t. Said otherwise, the character of an input up to time t determines all possible outputs up to time t.

Next we consider these notions of causality applied to initial response functions.

2.2.28 Definition (Causal, strongly causal subsequent response function) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system satisfying Assumption 2.2.26, let $t_0 \in \mathbb{T}$, and let $\rho^{\Sigma}_{t,t_0} : X^{\Sigma}_{t,t_0} \times \mathscr{U}_{\geq t_0} \to \mathscr{Y}_{\geq t_0}$ be a family of subsequent response functions from t_0 with state space X^{Σ}. Let $\tau \in \mathbb{T}_{\geq t_0}$.

(i) The subsequent response function ρ^{Σ}_{τ,t_0}, is ***causal*** if, for any $x_\tau \in X^{\Sigma}_{\tau,t_0}$, for any $t \in \mathbb{T}_{\geq \tau}$, and for any $\mu_1, \mu_1 \in \mathscr{U}_{\geq \tau}$, the following implication holds:

$$(\mu_1)_{[\tau,t]} = (\mu_2)_{[\tau,t]} \quad \Longrightarrow \quad \rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_1)_{\geq \tau}))_{[\tau,t]} = \rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_2)_{\geq \tau})_{[\tau,t]}.$$

(ii) The subsequent response function ρ^{Σ}_{τ,t_0} is ***strongly causal*** if, for any $x_\tau \in X^{\Sigma}_{\tau,t_0}$, for any $t \in \mathbb{T}_{\geq \tau}$, and for any $\mu_1, \mu_1 \in \mathscr{U}_{\geq \tau}$, the following implication holds:

$$(\mu_1)_{[\tau,t)} = (\mu_2)_{[\tau,t)} \quad \Longrightarrow \quad \rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_1)_{\geq \tau})_{[\tau,t]} = \rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_2)_{\geq \tau})_{[\tau,t]}. \quad \bullet$$

The idea of these definitions is that, for every time $t \geq t_0$, equal inputs up to time t with the same initial state give equal outputs from time t. One might expect there to be some relationship between causal systems and systems with causal initial response functions.

2.2.29 Theorem (Relationship between causality and causal initial response functions) *For a general time system*

$$\Sigma = (\mathrm{U}, \mathrm{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B}),$$

satisfying Assumption 2.2.26 and for $\mathrm{t}_0 \in \mathbb{T}$*, the following statements hold:*

(i) if Σ *has a causal (resp. strongly causal) initial response function from* t_0*, then* Σ *is causal (resp. strongly causal) from* t_0*;*

(ii) if Σ *is output complete and causal (resp. strongly causal) from* t_0*, then* Σ *has a causal (resp. strongly causal) initial response function from* t_0*.*

Proof (i) Let

$$\rho^{\Sigma}_{t_0} \colon X^{\Sigma}_{t_0} \times \mathscr{U}_{\geq t_0} \to \mathscr{Y}_{\geq t_0}$$

be a causal initial response function. Let $\mu_1, \mu_2 \in \mathscr{U}$ be such that, for $t \in \mathbb{T}_{\geq t_0}$, $(\mu_1)_{\leq t} = (\mu_2)_{\leq t}$. By hypothesis,

$$\rho^{\Sigma}_{t_0}(x, (\mu_1)_{\geq t_0})_{\leq t} = \rho^{\Sigma}_{t_0}(x, (\mu_2)_{\geq t_0})_{\leq t},$$

and so

$$\begin{aligned}
\mathscr{B}(\mu_1)_{[t_0,t]} &= \{\eta_{[t_0,t]} \mid \eta \in \mathscr{B}(\mu_1)\} \\
&= \{\eta_{[t_0,t]} \mid \eta_{\geq t_0} = \rho^{\Sigma}_{t_0}(x, (\mu_1)_{\geq t_0})\} \\
&= \{\eta_{[t_0,t]} \mid \eta_{\geq t_0} = \rho^{\Sigma}_{t_0}(x, (\mu_1)_{\geq t_0})\} \\
&= \mathscr{B}(\mu_2)_{[t_0,t]},
\end{aligned}$$

showing that Σ is causal from t_0. An entirely similar argument shows that Σ is strongly causal if it possesses a strongly causal initial response function.

(ii) In this part of the proof we assume that $\operatorname{dom}(\Sigma) = \mathscr{U}$. This is done for convenience, and one can readily see that the conclusions are easily adapted to the general case.

Let $(\mu^*, \eta^*) \in \mathscr{B}$ be arbitrarily chosen. Denote by

$$\mathscr{R}_{\Sigma} = \{F \colon \mathscr{U}'_{\geq t_0} \to \mathscr{Y}_{\geq t_0} \mid \mathscr{U}' \subseteq \mathscr{U},\ \operatorname{graph}(F) \subseteq \mathscr{B}_{\geq t_0},\ F \text{ is causal (resp. strongly causal)}\}$$

the set of functional input/output systems that take values in $\mathscr{B}$ and that are causal (resp. strongly causal) as general time systems. Note that $\mathscr{R}_{\Sigma}$ is nonempty since we can take $\mathscr{U}' = \{\mu^*\}$ and $F(\mu^*) = \eta^*$. Since Σ is causal (resp. strongly causal), it is evident that F is a causal (resp. strongly causal) general time system. We define a partial order on $\mathscr{R}_{\Sigma}$ by $F_1 \preceq F_2$ if $\operatorname{dom}(F_1) \subseteq \operatorname{dom}(F_2)$ and $F_2|\operatorname{dom}(F_1) = F_1$. Let $\mathscr{P} \subseteq \mathscr{R}_{\Sigma}$ be a totally ordered subset and let $\mathscr{U}_{\mathscr{P}} = \cup_{F \in \mathscr{P}} \operatorname{dom}(F)$ and define $F_{\mathscr{P}} \colon \mathscr{U}_{\mathscr{P}} \to \mathscr{Y}_{\geq t_0}$ by requiring that $F_{\mathscr{P}}|\operatorname{dom}(F) = F$ for every $F \in \mathscr{P}$. We claim that $F_{\mathscr{P}} \in \mathscr{R}_{\Sigma}$. First of all, if $\mu_{\geq t_0} \in \mathscr{U}_{\mathscr{P}}$, then for $F \in \mathscr{P}$ such that $\mu_{\geq t_0} \in \operatorname{dom}(F)$, we have

$$(\mu_{\geq t_0}, F_{\mathscr{P}}(\mu_{\geq t_0})) = (\mu_{\geq t_0}, F(\mu_{\geq t_0})) \in \mathscr{B}_{\geq t_0}.$$

Thus $\operatorname{graph}(F_{\mathscr{P}}) \subseteq \mathscr{B}_{\geq t_0}$. Now let $(\mu_1)_{\geq t_0}, (\mu_2)_{\geq t_0} \in \mathscr{U}_{\mathscr{P}}$ be such that $(\mu_1)_{[t_0,t]} = (\mu_2)_{[t_0,t]}$ (resp. $(\mu_1)_{[t_0,t)} = (\mu_2)_{[t_0,t)}$). Let $F \in \mathscr{P}$ be such that $(\mu_1)_{\geq t_0}, (\mu_2)_{\geq t_0} \in \operatorname{dom}(F)$. Since F is causal (resp. strongly causal), it follows that $\mathscr{B}(\mu_1)_{[t_0,t]} = \mathscr{B}(\mu_2)_{[t_0,t]}$. Thus $F_{\mathscr{P}}$ is causal,

and this shows that the totally ordered set $\mathscr{P}$ has an upper bound. By Zorn's Lemma, $\mathscr{R}_\Sigma$ has a maximal element that we denote by F_Σ. We claim that $\mathrm{dom}(F_\Sigma) = \mathscr{U}_{\geq t_0}$. This we prove separately in the causal and strongly causal cases.

First let us consider the case where Σ is strongly causal. Assume that $\mathrm{dom}(F_\Sigma) \neq \mathscr{U}_{\geq t_0}$ and let $\mu_0 \in \mathscr{U}$ be such that $(\mu_0)_{\geq t_0} \notin \mathrm{dom}(F_\Sigma)$. Let $\mu_{\geq t_0} \in \mathrm{dom}(F_\Sigma)$.

First suppose that there is no $t \in \mathbb{T}_{\geq t_0}$ for which $\mu_{[t_0,t]} = (\mu_0)_{[t_0,t]}$. Define F'_Σ so that $\mathrm{dom}(F'_\Sigma) = \mathrm{dom}(F_\Sigma) \cup \{\mu_0\}$ and so that

$$F'_\Sigma(\mu_{\geq t_0}) = \begin{cases} F_\Sigma(\mu_{\geq t_0}), & \mu \in \mathrm{dom}(F_\Sigma), \\ \eta_0, & \mu_{\geq t_0} = (\mu_0)_{\geq t_0}. \end{cases}$$

By our hypothesis, the causality of F_Σ implies F'_Σ is causal.

Thus we may consider the situation where there exists $t \in \mathbb{T}_{\geq t_0}$ such that $\mu_{[t_0,t]} = (\mu_0)_{[t_0,t]}$. Denote

$$\mathbb{T}(\mu, \mu_0) = \cup\{[t_0, t) \mid \ \mu_{[t_0,t)} = (\mu_0)_{[t_0,t)}\}$$

and $\tau(\mu, \mu_0) = \sup \mathbb{T}(\mu, \mu_0)$. If $t \in \mathbb{T}(\mu, \mu_0)$ we have $\mu_{[t_0,t]} = (\mu_0)_{[t_0,t]}$, and so causality of Σ gives

$$\mathscr{B}(\mu)_{[t_0,t]} = \mathscr{B}(\mu_0)_{[t_0,t]}.$$

Since $\mathrm{graph}(F_\Sigma) \subseteq \mathscr{B}_{\geq t_0}$, there exists $\eta_{\mu,t} \in \mathscr{B}(\mu_0)_{\geq t_0}$ such that

$$(\eta_{\mu,t})_{[t_0,t]} = F_\Sigma(\mu_{\geq t_0})_{[t_0,t]}.$$

We claim that there exists

$$\eta_\mu \colon \ \cup_{t \in \mathbb{T}(\mu,\mu_0)} [t_0, t] \to Y$$

such that $\eta_\mu(s) = \eta_{\mu,t}(s)$ for all $s \in [t_0, t]$. Indeed, let $t_1, t_2 \in \mathbb{T}(\mu, \mu_0)$ and let $t \in [t_0, t_1] \cap [t_0, t_2]$. Then

$$\eta_{\mu,t_1}(t) = F_\Sigma(\mu_{\geq t_0})(t) = \eta_{\mu_1,t_2}(t),$$

showing that $\eta_{\mu,t_1} = \eta_{\mu,t_2}$ agree on the intersection of their domains. Thus the asserted function η_μ exists. By output completeness of Σ, there exists $\eta_\mu \in \mathscr{B}(\mu_0)_{\geq t_0}$ such that

$$(\eta_\mu)_{[t_0,t]} = (\eta_{\mu,t})_{[t_0,t]}, \qquad t \in \mathbb{T}(\mu, \mu_0).$$

Now we claim that there exists

$$\eta \colon \ \cup \{[t_0, t] \mid \ t \in \mathbb{T}(\mu, \mu_0), \ \mu_{\geq t_0} \in \mathrm{dom}(F_\Sigma)\} \to Y$$

such that

$$\eta_{[t_0,t]} = (\eta_\mu)_{[t_0,t]}, \qquad t \in \mathbb{T}(\mu, \mu_0), \ \mu_{\geq t_0} \in \mathrm{dom}(F_\Sigma).$$

Indeed, let $t_1, t_2 \in \mathbb{T}(\mu, \mu_0)$, let $t \in [t_0, t_1] \cap [t_0, t_2]$, and let $(\mu_1)_{\geq t_0}, (\mu_2)_{\geq t_0} \in \mathscr{B}(\mu_0)_{\geq t_0}$ be such that

$$(\eta_{\mu_1})_{[t_0,t']} = (\eta_{\mu_1,t_1})_{[t_0,t']}, \qquad t' \in \mathbb{T}(\mu_1, \mu_0),$$

and

$$(\eta_{\mu_2})_{[t_0,t']} = (\eta_{\mu_2,t_2})_{[t_0,t']}, \qquad t' \in \mathbb{T}(\mu_2, \mu_0).$$

By our constructions above, this gives

$$(\eta_{\mu_1})_{[t_0,t]} = F_\Sigma((\mu_1)_{\geq t_0})_{[t_0,t]}, \quad (\eta_{\mu_2})_{[t_0,t]} = F_\Sigma((\mu_2)_{\geq t_0})_{[t_0,t]}.$$

By definition of $\mathbb{T}(\mu_1, \mu_0)$ and $\mathbb{T}(\mu_2, \mu_0)$,

$$(\mu_1)_{[t_0,t]} = (\mu_0)_{[t_0,t]}, \quad (\mu_2)_{[t_0,t]} = (\mu_0)_{[t_0,t]}.$$

Thus causality of F_Σ from t_0 gives

$$(\eta_{\mu_1})_{[t_0,t]} = (\eta_{\mu_2})_{[t_0,t]}.$$

Thus there exists the asserted function η. Moreover, by output completeness, there exists $\eta_0 \in \mathscr{B}(\mu_0)_{\geq t_0}$ such that

$$(\eta_0)_{[t_0,t]} = (\eta_\mu)_{[t_0,t]}, \qquad t \in \mathbb{T}(\mu, \mu_0),\ \mu_{\geq t_0} \in \operatorname{dom}(F_\Sigma).$$

Now define F'_Σ so that $\operatorname{dom}(F'_\Sigma) = \operatorname{dom}(F_\Sigma) \cup \{(\mu_0)_{\geq t_0}\}$ and

$$F'_\Sigma(\mu_{\geq t_0}) = \begin{cases} F_\Sigma(\mu_{\geq t_0}), & \mu_{\geq t_0} \in \operatorname{dom}(F_\Sigma), \\ \eta_0, & \mu = \mu_0. \end{cases}$$

We claim that F'_Σ is causal. Let $(\mu_1)_{\geq t_0}, (\mu_2)_{\geq t_0} \in \operatorname{dom}(F'_\Sigma)$ and suppose that $(\mu_1)_{[t_0,t]} = (\mu_2)_{[t_0,t]}$. If $(\mu_1)_{\geq t_0}, (\mu_2)_{\geq t_0} \in \operatorname{dom}(F_\Sigma)$, then causality of F_Σ gives

$$(F'_\Sigma(\mu_1))_{[t_0,t]} = (F'_\Sigma(\mu_2))_{[t_0,t]}.$$

If $(\mu_1)_{\geq t_0} \in \operatorname{dom}(F_\Sigma)$ and $(\mu_2)_{\geq t_0} = (\mu_0)_{\geq t_0}$, then, since $(\mu_1)_{[t_0,t]} = (\mu_0)_{[t_0,t]}$, we must have $t \leq \tau(\mu_1, \mu_0)$. Therefore, making reference to our constructions above,

$$(F_\Sigma((\mu_1)_{\geq t_0})_{[t_0,t]} = (\eta_{\mu_1})_{[t_0,t]} = (\eta_0)_{[t_0,t]}.$$

From this we have

$$(F'_\Sigma((\mu_1)_{\geq t_0})_{[t_0,t]} = (F_\Sigma((\mu_1)_{\geq t_0})_{[t_0,t]} = (\eta_0)_{[t_0,t]} = (F'_\Sigma((\mu_2)_{\geq t_0})_{[t_0,t]},$$

giving causality of F'_Σ. Thus $F'_\Sigma \in \mathscr{R}_\Sigma$ and $F_\Sigma \preceq F'_\Sigma$, which contradicts the maximality of F_Σ, and so we must have $\operatorname{dom}(F_\Sigma) = \mathscr{U}_{\geq t_0}$.

Now we show that this same conclusion holds when Σ is assumed to be strongly causal. Assume that $\operatorname{dom}(F_\Sigma) \neq \mathscr{U}_{\geq t_0}$ and let $\mu_0 \in \mathscr{U}$ be such that $(\mu_0)_{\geq t_0} \notin \operatorname{dom}(F_\Sigma)$. For $\mu_{\geq t_0} \in \operatorname{dom}(F_\Sigma)$, denote

$$\mathbb{T}(\mu, \mu_0) = \cup\{[t_0, t) \mid \ \mu_{[t_0,t)} = (\mu_0)_{[t_0,t)}\}$$

and $\tau(\mu, \mu_0) = \sup \mathbb{T}(\mu, \mu_0)$. Note that, if $\mu_{\geq t_0} \in \operatorname{dom}(F_\Sigma)$, then

$$\mu_{[t_0,\tau(\mu,\mu_0))} = (\mu_0)_{[t_0,\tau(\mu,\mu_0))},$$

and, since Σ is strongly causal from t_0,

$$\mathscr{B}(\mu)_{[t_0,\tau(\mu,\mu_0)]} = \mathscr{B}(\mu_0)_{[t_0,\tau(\mu,\mu_0)]}.$$

Since $\text{graph}(F_\Sigma) \subseteq \mathscr{B}_{\geq t_0}$, there exists $\eta_\mu \in \mathscr{B}(\mu_0)_{\geq t_0}$ such that

$$(\eta_\mu)_{[t_0,\tau(\mu,\mu_0)]} = F_\Sigma(\mu_{\geq t_0})_{[t_0,\tau(\mu,\mu_0)]}.$$

We claim that there exists

$$\eta\colon \cup_{\mu_{\geq t_0} \in \text{dom}(F_\Sigma)} [t_0, \tau(\mu, \mu_0)) \to Y \tag{2.6}$$

such that $\eta(t) = \eta_\mu(t)$ for every $\mu_{\geq t_0} \in \text{dom}(F_\Sigma)$ and $t \in [t_0, \tau(\mu, \mu_0))$. Suppose that

$$t \in [t_0, \tau(\mu_1, \mu_0)) \cap [t_0, \tau(\mu_2, \mu_0))$$

for $(\mu_1)_{\geq t_0}, (\mu_2)_{\geq t_0} \in \text{dom}(F_\Sigma)$. Then

$$(\mu_1)_{[t_0,t)} = (\mu_2)_{[t_0,t)} = (\mu_0)_{[t_0,t)}.$$

By strong causality of F_Σ,

$$(F_\Sigma((\mu_1)_{\geq t_0}))_{[t_0,t]} = (F_\Sigma((\mu_2)_{\geq t_0}))_{[t_0,t]}.$$

Since

$$(\eta_{\mu_1})_{[t_0,\tau(\mu_1,\mu_0)]} = (F_\Sigma((\mu_1)_{\geq t_0}))_{[t_0,\tau(\mu_1,\mu_0)]}.$$

and

$$(\eta_{\mu_2})_{[t_0,\tau(\mu_2,\mu_0)]} = (F_\Sigma((\mu_2)_{\geq t_0}))_{[t_0,\tau(\mu_2,\mu_0)]},$$

we have

$$(\eta_{\mu_1})_{[t_0,t]} = (F_\Sigma((\mu_1)_{\geq t_0}))_{[t_0,t]} = (F_\Sigma((\mu_2)_{\geq t_0}))_{[t_0,t]} = (\eta_{\mu_2})_{[t_0,t]}.$$

Thus η exists as in (2.6). Now, by output completeness, we conclude that there exists $\eta_0 \in \mathscr{B}(\mu_0)$ such that

$$(\eta_0)_{[t_0,\tau(\mu,\mu_0)]} = (\eta_\mu)_{[t_0,\tau(\mu,\mu_0)]}, \qquad \mu_{\geq t_0} \in \text{dom}(F_\Sigma).$$

Now define F'_Σ so that $\text{dom}(F'_\Sigma) = \text{dom}(F_\Sigma) \cup \{(\mu_0)_{\geq t_0}\}$ and

$$F'_\Sigma(\mu_{\geq t_0}) = \begin{cases} F_\Sigma(\mu_{\geq t_0}), & \mu_{\geq t_0} \in \text{dom}(F_\Sigma), \\ \eta_0, & \mu = \mu_0. \end{cases}$$

We claim that F'_Σ is strongly causal. Let $(\mu_1)_{\geq t_0}, (\mu_2)_{\geq t_0} \in \text{dom}(F'_\Sigma)$ and suppose that $(\mu_1)_{[t_0,t)} = (\mu_2)_{[t_0,t)}$. If $(\mu_1)_{\geq t_0}, (\mu_2)_{\geq t_0} \in \text{dom}(F_\Sigma)$, then strong causality of F_Σ gives

$$(F'_\Sigma(\mu_1))_{[t_0,t]} = (F'_\Sigma(\mu_2))_{[t_0,t]}.$$

If $(\mu_1)_{\geq t_0} \in \text{dom}(F_\Sigma)$ and $(\mu_2)_{\geq t_0} = (\mu_0)_{\geq t_0}$, then, since $(\mu_1)_{[t_0,t)} = (\mu_0)_{[t_0,t)}$, we must have $t \leq \tau(\mu_1, \mu_0)$. Therefore, making reference to our constructions above,

$$(F_\Sigma((\mu_1)_{\geq t_0})_{[t_0,\tau(\mu_1,\mu_0)]} = (\eta_{\mu_1})_{[t_0,\tau(\mu_1,\mu_0)]} = (\eta_0)_{[t_0,\tau(\mu_1,\mu_0)]}.$$

From this we have

$$(F'_\Sigma((\mu_1)_{\geq t_0})_{[t_0,t]} = (F_\Sigma((\mu_1)_{\geq t_0})_{[t_0,t]} = (\eta_0)_{[t_0,t]} = (F'_\Sigma((\mu_2)_{\geq t_0})_{[t_0,t]},$$

giving strong causality of F'_Σ. Thus $F'_\Sigma \in \mathscr{R}_\Sigma$ and $F_\Sigma \preceq F'_\Sigma$, which contradicts the maximality of F_Σ, and so we must have $\mathrm{dom}(F_\Sigma) = \mathscr{U}_{\geq t_0}$.

Now that we have shown that $\mathrm{dom}(F_\Sigma) = \mathscr{U}_{\geq t_0}$, we can conclude the proof of the theorem. We let

$$X^\Sigma_{t_0} = \{F_\Sigma \colon \mathscr{U}_{\geq t_0} \to \mathscr{Y}_{\geq t_0} \mid F_\Sigma \in \mathscr{R}_\Sigma,\ \mathrm{dom}(F_\Sigma) = \mathscr{U}_{\geq t_0}\},$$

this set having been shown to be nonempty. We then define

$$\begin{aligned}\rho^\Sigma_{t_0} \colon X^\Sigma_{t_0} \times \mathscr{U}_{\geq t_0} &\to \mathscr{Y}_{\geq t_0}\\ (F_\Sigma, \mu_{\geq t_0}) &\mapsto F_\Sigma(\mu_{\geq t_0}).\end{aligned}$$

It is the easy to see that $\rho^\Sigma_{t_0}$ is causal (resp. strongly causal) by definition of $\mathscr{R}_\Sigma$. ■

We shall consider in Chapter 5 some large classes of causal general time systems. Many of these systems arise as a consequence of the following theorem.

2.2.30 Theorem (Causal systems and state space representations) *Let*

$$\Sigma = (\mathrm{U}, \mathrm{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system, let $\mathrm{t}_0 \in \mathbb{T}$*, and suppose that*

$$\{\mu(\mathrm{t}) \mid \mu \in \mathscr{U}\} = \mathrm{U}, \qquad \mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0}.$$

Then Σ *is causal from* t_0 *if and only it is has a state space representation at* t_0*.*

Proof Before we begin the specific proof, let us engage in a discussion of some general concepts. Suppose that we have a pre-dynamical system representation for Σ at t_0 prescribed by the data

$$\begin{aligned}&X^\Sigma_{t,t_0}, && t \in \mathbb{T}_{\geq t_0},\\ &\rho^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0},\\ &\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma_{t_1,t_0} \to X^\Sigma_{t_2,t_0}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2.\end{aligned}$$

Let us say that the pre-dynamical system representation is ***surjective*** if

$$\Phi^\Sigma_{t,t_0}(X^\Sigma_{t_0,t_0} \times \mathscr{U}_{[t_0,t)}) = X^\Sigma_{t,t_0}, \qquad t \in \mathbb{T}_{\geq t_0}.$$

The relevance, for our purposes, of the notion of a surjective pre-dynamical system representation is the following result.

1 Lemma *Let*

$$\Sigma = (\mathrm{U}, \mathrm{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system with surjective pre-dynamical system representation at t_0 *prescribed by the data*

$$\begin{aligned}&\mathrm{X}^\Sigma_{\mathrm{t},\mathrm{t}_0}, && \mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0},\\ &\rho^\Sigma_{\mathrm{t},\mathrm{t}_0} \colon \mathrm{X}^\Sigma_{\mathrm{t},\mathrm{t}_0} \times (\mathscr{U}_{\geq \mathrm{t}_0})_{\geq \mathrm{t}} \to (\mathscr{Y}_{>\mathrm{t}_0})_{\geq \mathrm{t}}, && \mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0},\\ &\Phi^\Sigma_{\mathrm{t}_2,\mathrm{t}_1} \colon \mathscr{U}_{[\mathrm{t}_1,\mathrm{t}_2)} \times \mathrm{X}^\Sigma_{\mathrm{t}_1,\mathrm{t}_0} \to \mathrm{X}^\Sigma_{\mathrm{t}_2,\mathrm{t}_0}, && \mathrm{t}_1, \mathrm{t}_2 \in \mathbb{T}_{\geq \mathrm{t}_0},\ \mathrm{t}_1 \leq \mathrm{t}_2.\end{aligned}$$

Then the following statements are equivalent:

(i) $\rho^{\Sigma}_{\mathrm{t}_0,\mathrm{t}_0}$ *is a causal initial response function from* t_0*;*

(ii) $\rho^{\Sigma}_{\tau,\mathrm{t}_0}$ *is a causal initial response function from* τ *for every* $\tau \in \mathbb{T}_{\geq \mathrm{t}_0}$*.*

Proof (i) $\Longrightarrow$ (ii) Let $\tau \in \mathbb{T}_{\geq t_0}$, let $(\mu_1)_{\geq\tau}, (\mu_2)_{\geq\tau} \in \mathscr{U}_{\geq\tau}$ and let $x_\tau \in X^{\Sigma}_{\tau,t_0}$. By surjectivity of the pre-dynamical system representation, let $x^* \in X^{\Sigma}_{t_0,t_0}$ and $\mu^* \in \mathscr{U}_{\geq t_0}$ be such that $\Phi^{\Sigma}_{\tau,t_0}(x^*, \mu^*_{[t_0,\tau)}) = x_\tau$. Then

$$\rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_1)_{\geq\tau}) = \rho^{\Sigma}_{\tau,t_0}(\Phi^{\Sigma}_{\tau,t_0}(x^*, \mu^*_{[t_0,\tau)}), (\mu_1)_{\geq\tau}) = \rho^{\Sigma}_{\tau,t_0}(x^*, \mu^*_{[t_0,\tau)} * (\mu_1)_{\geq\tau})_{\geq\tau}$$

and, similarly,

$$\rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_2)_{\geq\tau}) = \rho^{\Sigma}_{\tau,t_0}(x^*, \mu^*_{[t_0,\tau)} * (\mu_2)_{\geq\tau})_{\geq\tau}.$$

Now let $t \in \mathbb{T}_{\geq\tau}$ and suppose that

$$((\mu_1)_{\geq\tau})_{[\tau,t]} = ((\mu_2)_{\geq\tau})_{[\tau,t]}.$$

Then

$$(\mu^*_{[t_0,\tau)} * (\mu_1)_{\geq\tau})_{[t_0,t]} = (\mu^*_{[t_0,\tau)} * (\mu_2)_{\geq\tau})_{[t_0,t]},$$

and so causality of $\rho^{\Sigma}_{t_0,t_0}$ implies that

$$\rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_1)_{\geq\tau})_{[t_0,t]} = \rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_2)_{\geq\tau})_{[t_0,t]}.$$

Combining all of this,

$$\rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_1)_{\geq\tau})_{[t_0,t]} = \rho^{\Sigma}_{\tau,t_0}(x_\tau, (\mu_2)_{\geq\tau})_{[t_0,t]},$$

showing that ρ^{Σ}_{τ,t_0} is causal, as claimed.

(ii) $\Longrightarrow$ (i) This is self-evident. ▼

Now we prove the theorem.

First of all, suppose that Σ has a state space representation at t_0 prescribed by the data

$$\begin{aligned} &X^{\Sigma}, \\ &\gamma^{\Sigma}_{t,t_0} \colon X^{\Sigma} \times U \to Y, && t \in \mathbb{T}_{\geq t_0}, \\ &\Phi^{\Sigma}_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^{\Sigma} \to X^{\Sigma}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2. \end{aligned}$$

Suppose that $(\mu_1)_{[t_0,t]} = (\mu_2)_{[t_0,t]}$. Then

$$\eta_{[t_0,t]} \in \mathscr{B}(\mu_1)_{[t_0,t]} \iff \eta_1(t) = \gamma^{\Sigma}_{t,t_0}(\Phi^{\Sigma}_{t,t_0}((\mu_1)_{[t_0,t)}, x), \mu_1(t))$$

and

$$\eta_{[t_0,t]} \in \mathscr{B}(\mu_2)_{[t_0,t]} \iff \eta_1(t) = \gamma^{\Sigma}_{t,t_0}(\Phi^{\Sigma}_{t,t_0}((\mu_2)_{[t_0,t)}, x), \mu_2(t)).$$

From this we conclude that $\mathscr{B}(\mu_1)_{[t_0,t]} = \mathscr{B}(\mu_2)_{[t_0,t]}$ and so Σ is causal from t_0.

Next suppose that Σ is causal from t_0 and, by Theorem 2.2.29, let $\rho^\Sigma_{t_0}$ be a causal initial response function with initial state object X^Σ. We now construct a surjective pre-dynamical system representation for Σ. Let

$$\begin{aligned}
&X^\Sigma_{t,t_0} = X^\Sigma \times \mathscr{U}_{[t_0,t)}, && t \in \mathbb{T}_{\geq t_0},\\
&\rho^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t} && t \in \mathbb{T}_{t,t_0},\\
&\qquad ((x, \mu_{[t_0,t)}), \mu'_{\geq t}) \mapsto \rho^\Sigma_{t_0}(x, \mu_{[t_0,t)} * \mu'_{\geq t})_{\geq t},\\
&\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma_{t_1,t_0} \to X^\Sigma_{t_2,t_0} && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2.\\
&\qquad (\mu'_{\geq t_2}, (x, \mu_{[t_1,t_2)})) \mapsto (\mu_{[t_1,t_2)}) * \mu'_{\geq t_2}, x),
\end{aligned}$$

It is a straightforward exercise to see that this is a pre-dynamical system representation (Exercise 2.2.6). It is also evident that this pre-dynamical system representation is surjective, and so, by the lemma above, ρ^Σ_{τ,t_0} is a causal initial response function from τ for every $\tau \in \mathbb{T}_{\geq t_0}$. Define

$$\begin{aligned}
\gamma^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times U &\to Y\\
((x, \mu_{[t_0,t)}), u) &\mapsto \rho^\Sigma_{t,t_0}((x, \mu_{[t_0,t)}), \mu'_{\geq t})(t),
\end{aligned}$$

where μ' satisfies $\mu'(t) = u$. Note that

$$\gamma^\Sigma_{t,t_0}((x, \mu_{[t_0,t)}), u) = \rho^\Sigma_{t_0}(x, \mu_{[t_0,t)}) * \mu'_{\geq t})(t).$$

From causality of $\rho^\Sigma_{t_0}$ we easily conclude that γ^Σ_{t,t_0} is well-defined, i.e., independent of $\mu_{[t_0,t)}$). One can readily verify that this data defines a pre-state space representation for Σ from t_0.

To define a state space representation, we follow closely the procedure from the proof of Theorem 2.2.20. We let $\overline{X}^\Sigma = \mathring{\cup}_{t \in \mathbb{T}} X^\Sigma_{t,t_0}$ and, for each $t \in \mathbb{T}$, arbitrarily choose $x^*_t \in X^\Sigma_{t,t_0}$. For $t_1, t_2 \in \mathbb{T}$ with $t_1 \leq t_2$, define

$$\overline{\Phi}^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times \overline{X}^\Sigma \to \overline{X}^\Sigma$$

by

$$\overline{\Phi}^\Sigma_{t_2,t_1}\big(\mu_{[t_1,t_2)}, (t', x_{t'})\big) = \begin{cases} (t_2, \Phi^\Sigma_{t_2,t_1}(\mu_{[t_1,t_2)}, x_{t_1})), & t' = t_1,\\ \big(t_2, \Phi^\Sigma_{t_2,t_1}(\mu_{[t_1,t_2)}, x^*_{t_1})\big), & t' \neq t_1. \end{cases}$$

Also define

$$\gamma^\Sigma_{t,t_0} \colon X^\Sigma \times U \to Y$$

by

$$\overline{\gamma}^\Sigma_{t,t_0}((t', x), u) = \begin{cases} \gamma^\Sigma_{t,t_0}(x, u), & t' = t,\\ \gamma^\Sigma_{t,t_0}(x^*_t, u), & t' \neq t. \end{cases}$$

We can proceed rather as in the proof of Theorem 2.2.20 to show that the data

$$\begin{aligned}
&\overline{X}_\Sigma,\\
&\overline{\Phi}^\Sigma_{t_2,t_1}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\\
&\gamma^\Sigma_{t,t_0}, && t \in \mathbb{T}_{\geq t_0},
\end{aligned}$$

defines a state space representation for Σ from t_0. ■

Let us give an example of a system that is not causal and a class of systems that are causal.

2.2.31 Examples (Causal and non-causal systems)

1. Since the notion of a general time system is so flexible, it is easy to give cooked examples of such systems that are not causal. For example, let us take $U = \mathbb{R}$, $Y = \mathbb{R}$, and $\mathbb{T} = \mathbb{R}_{\geq 0}$. We take $t_0 = 0$. Define $\mathscr{U}$ to be set of signals of the form

$$\mu_a(t) = \begin{cases} 0, & t \in [0, a], \\ 1, & t \in (a, \infty). \end{cases}$$

for $a \in \mathbb{R}_{\geq 0}$. Also let $\eta_a(t) = a$, $a, t \in \mathbb{R}_{\geq 0}$. Let

$$\mathscr{B} = \{(\mu_a, \eta_a) \mid a \in \mathbb{R}_{\geq 0}\}.$$

We claim that this system is not causal. Indeed, let $t \in \mathbb{R}$ and let $a_1, a_2 > t$ be distinct. Then $(\mu_{a_1})_{[0,t]} = (\mu_{a_2})_{[0,t]}$, but $\mathscr{B}(\mu_1)_{[0,t]} \neq \mathscr{B}(\mu_2)_{[0,t]}$.

2. Let us consider a special class of general time systems. A general time system

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

is ***memoryless*** if, for each $t \in \mathbb{T}$, there exists a mapping $\Phi_t \colon \mathscr{U}_t \to Y$ such that $\mathscr{Y}_t = \Phi_t(\mathscr{U}_t)$, where we recall from (2.3) the meaning of $\mathscr{U}_t$ and $\mathscr{Y}_t$. Thus a memoryless system has the character that the outputs at time t are determined solely by the inputs at time t. It is clear that a memoryless system is causal, but not strongly causal.

3. We claim that a deterministic finite state automaton is causal as a time system. This is easily seen directly since an output up to time t is uniquely parameterised by (a) the input up to time t and (b) the initial state. Also, if one wants to hit this with a big hammer, one can note from Example 2.2.25–2 that a deterministic finite state automata possesses a state space representation and, from Theorem 2.2.30, a system with a state space representation is causal. •

Causal systems form a natural class, since systems that are not causal can be thought of as "nonphysical," in that the future needs to be known to know the present. In Chapter 6 we shall consider large classes of causal systems, and it is this class of systems whose explication will occupy us for much of this volume.

2.2.7 Past-determined time systems

We now turn to a more stringent variation of causality for time systems. It is a property that is not possessed by all interesting systems, but is useful, when applicable, because it allows one to determine whether future outputs are determined by past measurable quantities, namely inputs and outputs. Most importantly, it does not rely on the notion of an internal (and possibly not measurable) state.

The following definition is intended to capture this idea.

2.2.32 Definition (Past-determined, strongly past-determined) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system satisfying Assumption 2.2.26, let $t_0 \in \mathbb{T}$, and let $\tau \in \mathbb{T}_{\geq t_0}$.

(i) The system Σ is ***past-determined*** from τ if:

(a) for any $(\mu, \eta) \in \mathscr{B}_{[t_0,\tau)}$ and for any $\mu' \in \mathscr{U}_{\geq\tau}$, there exists $\eta' \in \mathscr{Y}_{\geq\tau}$ such that $(\mu * \mu', \eta * \eta') \in \mathscr{B}$;

(b) for any $(\mu_1, \eta_1), (\mu_2, \eta_2) \in \mathscr{B}$ and for any $t \in \mathbb{T}_{\geq\tau}$, the following implication holds:

$$(\mu_1, \eta_1)_{[t_0,\tau)} = (\mu_2, \eta_2)_{[t_0,\tau)},\ (\mu_1)_{[t_0,t]} = (\mu_2)_{[t_0,t]} \implies (\eta_1)_{[t_0,t]} = (\eta_2)_{[t_0,t]}.$$

(ii) The system Σ is ***strongly past-determined*** from τ if:

(a) for any $(\mu, \eta) \in \mathscr{B}_{[t_0,\tau)}$ and for any $\mu' \in \mathscr{U}_{\geq\tau}$, there exists $\eta' \in \mathscr{Y}_{\geq\tau}$ such that $(\mu * \mu', \eta * \eta') \in \mathscr{B}$;

(b) for any $(\mu_1, \eta_1), (\mu_2, \eta_2) \in \mathscr{B}$ and for any $t \in \mathbb{T}_{\geq\tau}$, the following implication holds:

$$(\mu_1, \eta_1)_{[t_0,\tau)} = (\mu_2, \eta_2)_{[t_0,\tau)},\ (\mu_1)_{[t_0,t)} = (\mu_2)_{[t_0,t)} \implies (\eta_1)_{[t_0,t]} = (\eta_2)_{[t_0,t]}.$$ •

In order to understand the property of being past-determinable, let us introduce the following idea.

2.2.33 Definition (Finitely observable[7]) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system and let $t_0 \in \mathbb{T}$ and $\tau \in \mathbb{T}_{\geq t_0}$. The system Σ is ***finitely observable*** from τ if, for any $\mu \in \mathscr{U}$ and for any $\eta_1, \eta_2 \in \mathscr{B}(\mu)_{\geq t_0}$, the following implication holds:

$$(\eta_1)_{[t_0,\tau)} = (\eta_2)_{[t_0,\tau)} \implies (\eta_1)_{\geq t_0} = (\eta_2)_{\geq t_0}.$$ •

We may now give a characterisation of the property of being past-determined.

2.2.34 Proposition (Characterisation of past-determinacy) *Let*

$$\Sigma = (\mathrm{U}, \mathrm{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system and let $\mathrm{t}_0 \in \mathbb{T}$ *and* $\tau \in \mathbb{T}_{\geq \mathrm{t}_0}$. *Suppose that* $\mathscr{U}$ *is closed under concatenation. Then* Σ *is past-determined (resp. strongly past-determined) from* τ *if and only if it is*

[7] We shall consider in Section 9.4 the notion of "observability" for systems, and we comment that this is not related to the notion of "finitely observable" we consider here.

(i) causal (resp. strongly causal) from τ and

(ii) finitely observable from τ.

Proof Assume that Σ is past-determined from τ. Let $\mu_1, \mu_2 \in \mathscr{U}_{\geq t_0}$ satisfy $(\mu_1)_{[\tau,t]} = (\mu_2)_{[\tau,t]}$ for all $t \in \mathbb{T}_{\geq\tau}$ and let $(\eta_1)_{\geq t_0} \in \mathscr{B}(\mu_1)_{\geq t_0}$. Then $((\mu_1)_{[t_0,\tau)}, (\eta_1)_{[t_0,\tau)}) \in \mathscr{B}_{[t_0,\tau)}$. Past-determinacy then ensures that there exists $(\eta_2)_{\geq\tau}$ such that

$$((\mu_1)_{[t_0,\tau)} * (\mu_2)_{\geq\tau}, (\eta_1)_{[t_0,\tau)} * (\eta_2)_{\geq\tau}) \in \mathscr{B}_{\geq t_0}.$$

Abbreviate $\eta^*_{\geq t_0} = (\eta_1)_{[t_0,\tau)} * (\eta_2)_{\geq\tau}$. Since $(\mu_1)_{[\tau,t]} = (\mu_2)_{[\tau,t]}$ for $t \in \mathbb{T}_{\geq\tau}$, we must have

$$(\mu_1)_{[t_0,\tau)} * (\mu_2)_{\geq\tau} = (\mu_2)_{\geq t_0},$$

and so $((\mu_2)_{\geq t_0}, \eta^*_{\geq t_0}) \in \mathscr{B}_{\geq t_0}$. By construction we have $(\mu_1, \eta_1)_{[t_0,\tau)} = (\mu_2, \eta^*)_{[t_0,\tau]}$. Therefore, past-determinacy of Σ from τ, along with the fact that $(\mu_1)_{[\tau,t]} = (\mu_2)_{[\tau,t]}$ for all $t \in \mathbb{T}_{\geq\tau}$, implies that $(\eta_1)_{[t_0,t]} = \eta^*_{[t_0,t]}$ for all $t \in \mathbb{T}_{\geq\tau}$. We thus conclude that

$$\mathscr{B}(\mu_1)_{[t_0,t]} \subseteq \mathscr{B}(\mu_2)_{[t_0,t]}.$$

The opposite inclusion follows similarly, and so we conclude that Σ is causal from τ.

One similarly shows that if Σ is strongly past-determined from τ, then Σ is strongly causal from τ.

To show that Σ is finitely observable from τ, let $\eta_1, \eta_2 \in \mathscr{B}(\mu)$ satisfy $(\eta_1)_{[t_0,\tau)} = (\eta_2)_{[t_0,\tau)}$. Therefore, $(\mu, \eta_1)_{[t_0,\tau)} = (\mu, \eta_2)_{[t_0,\tau]}$ and $\mu_{[t_0,t]} = \mu_{[t_0,t]}$, and past-determinacy of Σ implies that $(\eta_1)_{[t_0,t]} = (\eta_2)_{[t_0,t]}$ for all $t \in \mathbb{T}_{\geq\tau}$. Thus $(\eta_1)_{\geq t_0} = (\eta_2)_{\geq t_0}$, and we conclude that Σ is finitely observable from τ.

Now suppose that (i) and (ii) in the statement of the proposition hold. Assume that

$$(\mu_1, \eta_1)_{[t_0,\tau)} = (\mu_2, \eta_2)_{[t_0,\tau)}$$

and $(\mu_1)_{[t_0,t]} = (\mu_2)_{[t_0,t]}$ for all $t \in \mathbb{T}_{\geq\tau}$ for $(\mu_1, \eta_1), (\mu_2, \eta_2) \in \mathscr{B}$. Since Σ is causal from τ, there exists $\eta^* \in \mathscr{B}(\mu_2)$ such that $\eta^*_{[t_0,t]} = (\eta_1)_{[t_0,t]}$. Since $t \geq \tau$,

$$\eta^*_{[t_0,\tau)} = (\eta_1)_{[t_0,\tau)} = (\eta_2)_{[t_0,\tau)}.$$

Since $\eta_2, \eta^* \in \mathscr{B}(\mu_2)$, finite observability gives $\eta^* = \eta_2$ and so

$$\eta^*_{[t_0,t]} = (\eta_1)_{[t_0,t]} = (\eta_2)_{[t_0,t]}$$

for $t \in \mathbb{T}_{\geq\tau}$. This is the second half of the definition of past-determinacy. Similarly one proves the second half of the definition for strong past-determinacy when Σ is strongly causal from τ.

For the first half of the definition of past-determinacy, let $(\mu, \eta) \in \mathscr{B}_{[t_0,\tau)}$ and let $\mu' \in \mathscr{U}_{\geq\tau}$. Since $\mathscr{U}$ is closed under concatenation, $\mu^* \triangleq \mu_{[t_0,\tau]} * \mu'_{\geq\tau} \in \mathscr{U}$ Since Σ is causal from τ, we have

$$\mu_{[t_0,\tau)} = \mu^*_{[t_0,\tau)} \implies \mathscr{B}(\mu)_{[t_0,\tau)} = \mathscr{B}(\mu^*)_{[t_0,\tau)},$$

and so there exists $\eta^* \in \mathscr{B}(\mu^*)$ such that $\eta^*_{[t_0,\tau)} = \eta_{[t_0,\tau)}$. Now we write

$$(\mu^*_{\geq t_0}, \eta^*_{\geq t_0}) = (\mu_{[t_0,\tau]} * \mu'_{\geq\tau}, \eta_{[t_0,\tau)} * \eta^*_{\geq\tau}),$$

and this gives the desired condition. ■

Let us relate past-determinacy with causality.

2.2.35 Proposition (Past-determined systems are causal) *A general time system*

$$\Sigma = (\mathrm{U}, \mathrm{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

is causal (resp. strongly causal) from τ if it is past-determined (resp. strongly past-determined) from τ.

Proof Let $(\mu_{[t_0,\tau)}, \eta_{[t_0,\tau)}) \in \mathscr{B}_{[t_0,\tau)}$ and let $\mu'_{\geq\tau} \in \mathscr{U}_{\geq\tau}$. By past-determinacy, there exists η'_τ such that

$$(\mu_{[t_0,\tau)} * \mu'_{\geq\tau}, \eta_{[t_0,\tau)} * \eta'_{\geq\tau}) \in \mathscr{B}_{\geq t_0}.$$

By Proposition 2.2.34, Σ is finitely observable from τ and this implies that the previous condition uniquely defines $\eta'_{\geq\tau}$.

With this in mind, we construct a causal (resp. strongly causal) initial response function from τ. To this end we define

$$X^\Sigma = \{(\mu_{[t_0,\tau)}, \eta_{[t_0,\tau)}) \mid (\mu_{\geq t_0}, \eta_{\geq t_0}) \in \mathscr{B}_{\geq t_0}\}$$

and define

$$\rho^\Sigma_{\tau,t_0} \colon X^\Sigma \times \mathscr{U}_{\geq\tau} \to \mathscr{Y}_{\geq\tau}$$

by the requirement that

$$\eta'_{\geq\tau} = \rho^\Sigma_{\tau,t_0}((\mu_{[t_0,\tau)}, \eta_{[t_0,\tau)}), \mu'_{\geq\tau}) \iff (\mu_{[t_0,\tau)} * \mu'_{\geq\tau}, \eta_{[t_0,\tau)} * \eta'_{\geq\tau}) \in \mathscr{B}_{\geq t_0}.$$

It immediately follows from the second part of the definition of past-determinacy (resp. strong past-determinacy) that ρ^Σ_{τ,t_0} is a causal (resp. strongly causal) initial response function, showing that Σ is causal (resp. strongly causal) from τ. ■

what? In we shall see a large and important class of past-determined time systems. The following example shows that the converse of the preceding result is not true.

2.2.36 Example (A general time system that is not past-determined) We consider a deterministic finite state automaton $(Q, Y, \Lambda, \delta, \gamma)$ with

$$Q = \{s_0, s_1\}, \quad Y = \{0, 1\}, \quad \Lambda = \{0, 1\},$$

with the state transition map δ determined by the table

	0	1
s_0	s_0	s_0
s_1	s_1	s_1

and the output function γ determined by the table

	0	1
s_0	0	0
s_1	0	1

As we saw in Example 2.2.31–3, this system is causal. We claim that this system is not past-determined. To see this, note that, if the initial state is s_0, then the output is 0 for all time, no matter what the input. However, if the initial state is s_1, then an input string with a 1 in the kth slot will produce an output of 1 in the $(k+1)$st slot. From this we can conclude that this system is not past-determined from any τ. •

2.2.8 Stationarity in time systems

Stationarity is a property of time-invariance. As we shall see, the way in which one expresses this is to say that the system is invariant under shifts of time, in a suitable sense. Thus, in the development, one needs to have time-domains where shifts make sense. This is where we make use of the idea of a time-domain for stationarity, as in Definition 2.2.3.

Let us begin with the definitions. First we introduce some notation for time shifts. We consider a stationary time-domain $((\mathbb{S}, \leq), \mathbb{T})$, $t_0 \in \mathbb{T}$, and $a \in \mathbb{S}$. For a sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}_{\geq t_0}$, let us say that a is ***$\mathbb{T}'$-admissible*** if

$$\{t \in \mathbb{T}' \mid t - a \in \mathbb{T}_{\geq t_0}\} = \mathbb{T}'.$$

If a is $\mathbb{T}'$-admissible, let us denote

$$\mathbb{T}'_a = \{t - a \in \mathbb{T}_{\geq t_0} \mid t \in \mathbb{T}'\}$$

and so define

$$\begin{aligned} \hat{\tau}_a \colon \mathbb{T}'_a &\to \mathbb{T}_{\geq t_0} \\ t &\mapsto t + a. \end{aligned}$$

Now let X be a set and let $\mathscr{X} \subseteq X^{(\mathbb{T})}$. For $\xi \in \mathscr{X}_{\geq t_0}$, suppose that $a \in \mathbb{S}$ is $\mathrm{dom}(\xi)$-admissible. This being the case, we denote by $\hat{\tau}_a^* \xi$ the signal with domain $\mathrm{dom}(\xi)_a$ given by $\hat{\tau}_a^* \xi(t) = \xi(t + a)$. We depict, shifted signals in Figure 2.8. With these

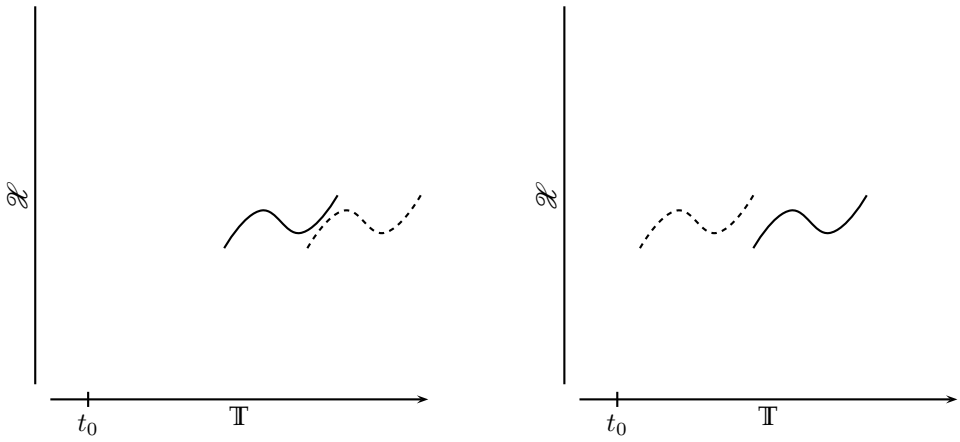


Figure 2.8 Shifting signals (dashed) in a stationary time-domain by positive time (left) and negative time (right)

constructions, let us define another shift operator that shifts a signal that starts at a time greater than t_0 back to the time t_0. Specifically, let $\xi \in \mathscr{X}_{\geq t_0}$, let $a \in \mathbb{T}_{\geq t_0}$, and denote $\tau^*_{t_0,a} \xi$ to be the signal with domain

$$\mathrm{dom}(\tau^*_{t_0,a} \xi) = \{t - (a - t_0) \mid t \in \mathrm{dom}(\xi)\}$$

and defined by

$$\tau^*_{t_0,a}\xi(t) = \xi(t + (a - t_0)).$$

In Figure 2.9 we depict this operation.

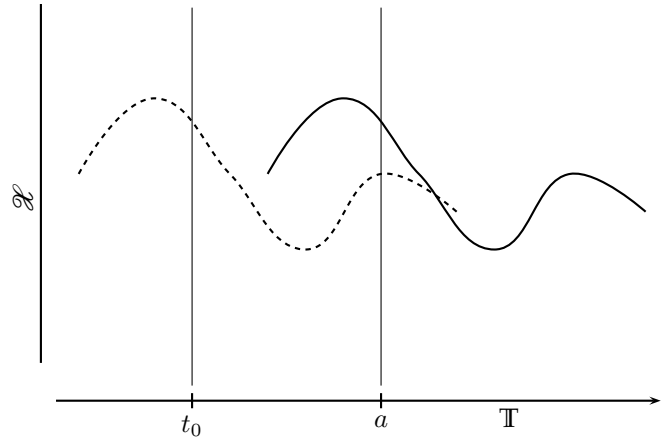


Figure 2.9 Shifting from an initial time a to an initial time t_0

With these rather elementary and somewhat cumbersome bits of notation at our disposal, we can easily give definitions of stationarity for time systems.

2.2.37 Definition ((Strongly) stationary general time system) Let $((\mathbb{S}, \leq), \mathbb{T})$ be a stationary time-domain, let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system, and let $t_0 \in \mathbb{T}$. The system

(i) is ***stationary*** from t_0 if $\tau^*_{t_0,t}(\mathscr{U}_{\geq t_0}) = \mathscr{U}_{\geq t_0}$ and $\tau^*_{t_0,t}(\mathscr{B}_{\geq t_0}) \subseteq \mathscr{B}_{\geq t_0}$ for every $t \in \mathbb{T}_{\geq t_0}$, and

(ii) is ***strongly stationary*** from t_0 if $\tau^*_{t_0,t}(\mathscr{B}_{\geq t_0}) = \mathscr{B}_{\geq t_0}$ for every $t \in \mathbb{T}_{\geq t_0}$. •

We wish to characterise stationary systems using dynamical systems representations. To this end, we make the following definitions.

2.2.38 Definition ((Strongly) time-invariant dynamical system representation) Let $((\mathbb{S}, \leq), \mathbb{T})$ be a stationary time-domain and let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system with dynamical system representation at t_0 prescribed by the data

$$\begin{aligned} &X^\Sigma, \\ &\rho^\Sigma_{t,t_0} : X^\Sigma \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0}, \\ &\Phi^\Sigma_{t_2,t_1} : \mathscr{U}_{[t_1,t_2)} \times X^\Sigma \to X^\Sigma, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2. \end{aligned}$$

The dynamical system representation is

(i) ***time-invariant*** from t_0 if, for $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, for $\mu \in \mathscr{U}_{\geq t_0}$, and for $x \in X^\Sigma$, it holds that
   (a) $\tau^*_{t_0,t_1}(\mathscr{U}_{\geq t_0}) \subseteq \mathscr{U}_{\geq t_0}$,
   (b) $\rho^\Sigma_{t_1,t_0}(x, \mu_{\geq t_1}) = \hat{\tau}^*_{t_1-t_0}\rho^\Sigma_{t_0}(x, \hat{\tau}^*_{-(t_1-t_0)}\mu_{\geq t_1})$,
   (c) $\Phi^\Sigma_{t_2,t_1}(x, \mu_{[t_1,t_2)}) = \Phi^\Sigma_{t_2-t_1+t_0,t_0}(x, \hat{\tau}^*_{-(t_1-t_0)}\mu_{[t_1,t_2)})$, and
   (d) $\tau^*_{t_0,t_1}(\mathscr{Y}_{\geq t_0}) = \mathscr{Y}_{\geq t_0}$.

(ii) ***strongly time-invariant*** from t_0 if, additionally, for every $t \in \mathbb{T}_{\geq t_0}$,

$$\mathscr{B}_{\geq t} = \{(\mu_{\geq t}, \eta_{\geq t}) \mid \eta_{\geq t} = \rho^\Sigma_{t,t_0}(x, \mu_{\geq t}) \text{ for some } x \in X^\Sigma\}. \qquad \bullet$$

Our main result is then the following.

2.2.39 Theorem ((Strongly) stationary systems and (strongly) time-invariant dynamical system representations) *Let* $((\mathbb{S}, \leq), \mathbb{T})$ *be a stationary time-domain, let*

$$\Sigma = (\mathrm{U}, \mathrm{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a complete general time system, and let $\mathrm{t}_0 \in \mathbb{T}$. *Then:*

(i) the following three statements are equivalent:
   (a) Σ *is stationary from* t_0*;*
   (b) for any $\mathrm{t}_1, \mathrm{t}_2 \in \mathbb{T}_{\geq \mathrm{t}_0}$ *with* $\mathrm{t}_1 < \mathrm{t}_2$,

$$\tau^*_{\mathrm{t}_0,\mathrm{t}_2}(\mathscr{B}_{\geq \mathrm{t}_0}) \subseteq \tau^*_{\mathrm{t}_0,\mathrm{t}_1}(\mathscr{B}_{\geq \mathrm{t}_0});$$

   (c) Σ *possesses a dynamical system representation that is time-invariant from* t_0*;*

(ii) the following two statements are equivalent:
   (a) Σ *is strongly stationary from* t_0*;*
   (b) Σ *possesses a dynamical system representation that is strongly time-invariant from* t_0.

Proof Since Σ is complete, we suppose that all behaviours are defined on $\mathbb{T}$.
(i)(a) $\Longrightarrow$ (i)(c) Suppose that Σ is stationary from t_0 and denote

$$X^\Sigma = \{x \colon \mathscr{U}_{\geq t_0} \to \mathscr{Y}_{\geq t_0} \mid \text{ graph}(x) \subseteq \mathscr{B}\}.$$

Define

$$\begin{aligned} \rho^\Sigma_{t_0} \colon X^\Sigma \times \mathscr{U}_{\geq t_0} &\to \mathscr{Y}_{\geq t_0} \\ (x, \mu_{\geq t_0}) &\mapsto x(\mu_{\geq t_0}). \end{aligned}$$

By definition, $\rho^\Sigma_{t_0}$ is an initial response function for Σ from t_0 with initial state object X^Σ. Now define $\Phi^\Sigma_{t,t_0} \colon X^\Sigma \times \mathscr{U}_{[t_0,t)} \to X^\Sigma$ by requiring that, if

$$\Phi^\Sigma_{t,t_0}(x, \mu_{[t_0,t)}) = \hat{x},$$

then $\hat{x}(\hat{\mu}_{\geq t_0}) = \hat{\eta}_{\geq t_0}$ if and only if

$$(x(\mu_{[t_0,t)} * (\hat{\tau}^*_{t-t_0}\hat{\mu}_{\geq t_0})))_{\geq t} = \hat{\tau}^*_{t-t_0}\hat{\eta}_{\geq t_0}.$$

We should show that this definition makes sense. First we verify that Φ^Σ_{t,t_0} takes values in X^Σ. To see this, note that, if $\hat{x}(\hat{\mu}_{\geq t_0}) = \hat{\eta}_{\geq t_0}$, then

$$x(\mu_{[t_0,t)} * \hat{\tau}^*_{t-t_0}\hat{\mu}_{\geq t_0}) = \eta_{[t_0,t)} * \hat{\tau}^*_{t-t_0}\hat{\eta}_{\geq t_0}$$

for some $\eta_{[t_0,t)} \in \mathscr{Y}_{[t_0,t)}$. Note that

$$\hat{\mu}_{\geq t_0} = \hat{\tau}^*_{t-t_0}(\mu_{[t_0,t)} * \hat{\tau}^*_{t-t_0}\hat{\mu}_{\geq t_0}), \quad \hat{\eta}_{\geq t_0} = \hat{\tau}^*_{t-t_0}(\eta_{[t_0,t)} * \hat{\tau}^*_{t-t_0}\hat{\eta}_{\geq t_0})$$

so that $(\hat{\mu}, \hat{\eta}) \in \hat{\tau}_{t-t_0}(\mathscr{B}) \subseteq \mathscr{B}$. Thus $\hat{x} \in \mathscr{B}$. Now let us show that Φ^Σ_{t,t_0} is a single-valued function. Suppose that $\hat{\eta}_{\geq t_0}, \bar{\eta}_{\geq t_0} \in \mathscr{Y}_{\geq t_0}$ satisfy

$$(x(\mu_{[t_0,t)} * (\hat{\tau}^*_{t-t_0}\hat{\mu}_{\geq t_0})))_{\geq t} = \hat{\tau}^*_{t-t_0}\hat{\eta}_{\geq t_0} = \hat{\tau}^*_{t-t_0}\bar{\eta}_{\geq t_0}.$$

Then, since $\hat{\tau}_{t-t_0}$ is a bijection, $\hat{\eta}_{\geq t\_0} = \bar{\eta}_{\geq t_0}$, showing that x^* is single-valued. Finally, since $\tau^*_{t_0,t}(\mathscr{U}_{\geq t_0}) = \mathscr{U}_{\geq t_0}$, it follows that $\hat{x}$ is defined for every $\mu_{\geq t_0} \in \mathscr{U}_{\geq t_0}$.

Next we define the dynamical system representation by requiring that

$$\rho^\Sigma_{t,t_0}(x, \mu_{\geq t_1}) = \hat{\tau}^*_{t_1-t_0}\rho^\Sigma_{t_0}(x, \hat{\tau}_{-(t_1-t_0)}\mu_{\geq t_1}),$$
$$\Phi^\Sigma_{t_2,t_1}(x, \mu_{[t_1,t_2)}) = \Phi^\Sigma_{t_2-t_1,t_0}(x, \hat{\tau}^*_{-(t_1-t_0)}\mu_{[t_1,t_2)}).$$

Thus, provided we verify that this does indeed define a dynamical system representation, it is by construction time-invariant.

First we show that ρ^Σ_{t,t_0}, $t \in \mathbb{T}_{\geq t_0}$, is a response family. Let $(\mu_{\geq t}, \eta_{\geq t}) \in \mathscr{B}_{\geq t}$ so that

$$(\hat{\tau}^*_{-(t-t_0)}\mu_{\geq t}, \hat{\tau}^*_{-(t-t_0)}\eta_{\geq t}) \in \hat{\tau}^*_{t_0,t}(\mathscr{B}) \subseteq \mathscr{B}.$$

Then there exists $x \in X^\Sigma$ such that

$$\hat{\tau}^*_{-(t-t_0)}\eta_{\geq t} = \rho^\Sigma_{t_0}(x, \hat{\tau}^*_{-(t-t_0)}\mu_{\geq t}),$$

giving

$$\eta_{\geq t} = \hat{\tau}^*_{t-t_0}\rho^\Sigma_{t_0}(x, \hat{\tau}^*_{-(t-t_0)}\mu_{\geq t}) = \rho^\Sigma_{t,t_0}(x, \mu_{\geq t}),$$

which shows that ρ^Σ_{t,t_0}, $t \in \mathbb{T}_{\geq t_0}$, is indeed a response family.

It is straightforward to check that $\Phi^\Sigma_{t_2,t_1}$, $t_1, t_2 \in \mathbb{T}_{\geq t_0}$ comprise a state transition function.

Next we claim that

$$(\rho^\Sigma_{t_0}(x, \mu_{\geq t_0}))_{\geq t} = \rho^\Sigma_{t,t_0}(\Phi^\Sigma_{t,t_0}(x, \mu_{[t_0,t)}), \mu_{\geq t})$$

for every $x \in X^\Sigma$ and $\mu_{\geq t_0} \in \mathscr{U}_{\geq t_0}$. Denote

$$\eta_{\geq t} = (\rho^\Sigma_{t_0}(x, \mu_{\geq t_0}))_{\geq t}$$

so that

$$\eta_{\geq t} = (x(\mu_{\geq t_0}))_{\geq t} = (x(\mu_{[t_0,t)} * \mu_{\geq t}))_{\geq t},$$

which gives, by definition of Φ^Σ_{t,t_0}, $\hat{x}(\hat{\tau}^*_{t-t_0}\mu_{\geq t}) = \hat{\tau}^*_{t-t_0}\eta_{\geq t}$, where $\hat{x} = \Phi^\Sigma_{t,t_0}(x, \mu_{[t_0,t_1)})$. Thus we compute

$$\begin{aligned}
\hat{\tau}^*_{-(t-t_0)}\eta_{\geq t} &= \rho^\Sigma_{t_0}(\Phi^\Sigma_{t,t_0}(x, \mu_{[t_0,t)}), \hat{\tau}^*_{-(t-t_0)}\mu_{\geq t}) \\
&= \hat{\tau}^*_{-(t-t_0)} \circ \hat{\tau}^*_{t-t_0}\rho^\Sigma_{t_0}(\Phi^\Sigma_{t,t_0}(x, \mu_{[t_0,t)}), \hat{\tau}^*_{-(t-t_0)}\mu_{\geq t}) \\
&= \hat{\tau}^*_{-(t-t_0)}\rho^\Sigma_{t,t_0}(\Phi^\Sigma_{t,t_0}(x, \mu_{[t_0,t)}), \mu_{\geq t}),
\end{aligned}$$

giving our claim since $\hat{\tau}_{t-t_0}$ is a bijection.

Next we claim that ρ^Σ_{t,t_0}, $t \in \mathbb{T}_{\geq t_0}$, and $\Phi^\Sigma_{t_2,t_1}$, $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, are compatible. For this we compute

$$\begin{aligned}
(\rho^\Sigma_{t_1,t_0}(x, \mu_{\geq t_1}))_{\geq t_2} &= (\hat{\tau}^*_{t_1-t_0}\rho^\Sigma_{t_0}(x, \hat{\tau}^*_{-(t_1-t_0)}\mu_{\geq t_1}))_{\geq t_2} \\
&= \hat{\tau}^*_{t_1-t_0}(\rho^\Sigma_{t_0}(x, \hat{\tau}^*_{-(t_1-t_0)}\mu_{\geq t_1})_{t_2-t_1+t_0}) \\
&= \hat{\tau}^*_{t_1-t_0}\rho^\Sigma_{t,t_0}(\Phi^\Sigma_{t_1,t_0}(x, \hat{\tau}^*_{-(t_1-t_0)}\mu_{[t_0,t_1)}), \hat{\tau}^*_{-(t_1-t_0)}\mu_{\geq t_2}) \\
&= \hat{\tau}^*_{t_1-t_0}\hat{\tau}^*_{t_2-t_1}\rho^\Sigma_{t_0}(\Phi^\Sigma_{t_1,t_0}(x, \hat{\tau}^*_{-(t_1-t_0)}\mu_{[t_1,t_2)}), \hat{\tau}^*_{t_1-t_0}\hat{\tau}^*_{t_2-t_1}\mu_{\geq t_2}) \\
&= \hat{\tau}^*_{t_2-t_0}\rho^\Sigma_{t_0}(\Phi^\Sigma_{t_2-t_1+t_0,t_0}(x, \hat{\tau}^*_{-(t-t_0)}\mu_{[t_1,t_2)}), \hat{\tau}_{-(t_2-t_0)}\mu_{\geq t_2}) \\
&= \rho^\Sigma_{t_2-t_0,t_0}(\Phi^\Sigma_{t_2,t_1}(x, \mu_{[t_1,t_2)}), \mu_{\geq t_2}),
\end{aligned}$$

as desired.

(i)(c) $\implies$ (i)(a) Let

$$\begin{aligned}
&X^\Sigma, \\
&\rho^\Sigma_{t,t_0} \colon X^\Sigma \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0}, \\
&\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma \to X^\Sigma, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2.
\end{aligned}$$

be a dynamical system representation that is time-invariant at t_0. Now we have

$$\begin{aligned}
&(\mu_{\geq t_0}, \eta_{\geq t_0}) \in \hat{\tau}^*_{t_0,t}(\mathscr{B}_{\geq t_0}) \\
\iff &(\hat{\tau}^*_{t-t_0}\mu_{\geq t_0}, \hat{\tau}^*_{t-t_0}\eta_{\geq t_0}) \in \mathscr{B}_{\geq t} \\
\implies &\hat{\tau}^*_{t-t_0}\eta_{\geq t_0} = \rho^\Sigma_{t,t_0}(x, \hat{\tau}^*_{t-t_0}\mu_{\geq t_0}) \text{ for some } x \in X^\Sigma \\
\iff &\hat{\tau}^*_{t-t_0}\eta_{\geq t_0} = \hat{\tau}^*_{t-t_0}\rho^\Sigma_{t_0}(x, \hat{\tau}^*_{(-t-t_0)}\hat{\tau}^*_{t-t_0}\mu_{\geq t_0}) \text{ for some } x \in X^\Sigma \\
\iff &\eta_{\geq t_0} = \rho^\Sigma_{t_0}(x, \mu_{\geq t_0}) \text{ for some } x \in X^\Sigma \\
\iff &(\mu_{\geq t_0}, \eta_{\geq t_0}) \in \mathscr{B},
\end{aligned}$$

and so $\hat{\tau}^*_{t_0,t}(\mathscr{B}) \subseteq \mathscr{B}$.

(i)(a) $\implies$ (i)(b) Let $t_2' = t_2 - t_1 + t_0$. We have

$$\tau^*_{t_0,t_2}(\mathscr{B}) \subseteq \mathscr{B} \implies \tau^*_{-(t_2-t_1)}\tau_{t_0,t_2}(\mathscr{B}) \subseteq \tau_{-(t_2-t_1)}(\mathscr{B}).$$

Since

$$\tau^*_{-(t_2-t_1)}\tau^*_{t_0,t_2}(\mathscr{B}) = \tau^*_{t_1-t_0}(\mathscr{B}),$$

this part of the theorem follows.

(i)(b) $\implies$ (i)(a) This is obvious.

(ii)(a) $\implies$ (ii)(b) We use the fact that we have a time-invariant dynamical system representation, as proved above. Using the fact that Σ is strongly stationary, we compute

$$\begin{aligned}
&\eta_{\geq t} = \rho^{\Sigma}_{t,t_0}(x, \mu_{\geq t}) \text{ for some } x \in X^{\Sigma} \\
\iff &\tau^*_{t-t_0}\eta_{\geq t} = \tau^*_{t-t_0}\rho^{\Sigma}_{t,t_0}(x, \tau^*_{-(t-t_0)}\tau^*_{t-t_0}\mu_{\geq t}) \text{ for some } x \in X^{\Sigma} \\
\iff &\hat{\tau}^*_{t-t_0}\eta_{\geq t} = \rho^{\Sigma}_{t_0}(x, \hat{\tau}^*_{t-t_0}\mu_{\geq t}) \text{ for some } x \in X^{\Sigma} \\
\iff &(\hat{\tau}^*_{t-t_0}\mu_{\geq t}, \hat{\tau}^*_{t-t_0}\eta_{\geq t}) \in \mathscr{B} \\
\iff &(\hat{\tau}^*_{t-t_0}\mu_{\geq t}, \hat{\tau}^*_{t-t_0}\eta_{\geq t}) \in \hat{\tau}^*_{t-t_0}(\mathscr{B}) = \mathscr{B} \\
\iff &(\mu_{\geq t}, \eta_{\geq t}) \in \hat{\tau}^*_{-(t-t_0)}\mathscr{B} = \mathscr{B}_{\geq t},
\end{aligned}$$

giving this part of the theorem.

(ii)(b) $\implies$ (ii)(a) Suppose that we have a dynamical system representation

$$\begin{aligned}
&X^{\Sigma}, \\
&\rho^{\Sigma}_{t,t_0} \colon X^{\Sigma} \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0}, \\
&\Phi^{\Sigma}_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^{\Sigma} \to X^{\Sigma}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2.
\end{aligned}$$

that is time-invariant at t_0 and that satisfies

$$\mathscr{B}_{\geq t} = \{(\mu_{\geq t}, \eta_{\geq t}) \mid \eta_{\geq t} = \rho^{\Sigma}_{t,t_0}(x, \mu_{\geq t}) \text{ for some } x \in X^{\Sigma}\}.$$

Then we have

$$\begin{aligned}
&\eta_{\geq t} = \rho^{\Sigma}_{t_0}(x, \mu_{\geq t}) \text{ for some } x \in X^{\Sigma} \\
\iff &\hat{\tau}^*_{-(t-t_0)}\eta_{\geq t} = \rho^{\Sigma}_{t,t_0}(x, \hat{\tau}^*_{-(t-t_0)}\mu_{\geq t}) \text{ for some } x \in X^{\Sigma} \\
\iff &(\hat{\tau}^*_{-(t-t_0)}\mu_{\geq t}, \hat{\tau}^*_{-(t-t_0)}\eta_{\geq t}) \in \{(\mu_{\geq t}, \eta_{\geq t}) \mid \eta_{\geq t} = \rho^{\Sigma}_{t,t_0}(x, \mu_{\geq t}) \text{ for some } x \in X^{\Sigma}\} \\
\iff &(\hat{\tau}^*_{-(t-t_0)}\mu_{\geq t}, \hat{\tau}^*_{-(t-t_0)}\eta_{\geq t}) \in \tau^*_{t_0,t}\mathscr{B}_{\geq t_0} \\
\iff &(\mu_{\geq t}, \eta_{\geq t}) \in \tau^*_{t-t_0}\tau^*_{t_0,t}(\mathscr{B}_{\geq t_0}) = (\mathscr{B})_{\geq t},
\end{aligned}$$

giving this part of the theorem. ∎

While we will not consider such representations in this section, it is possible to consider time-invariant state space representations.

2.2.40 Definition ((Strongly) time-invariant state space representation) Let $((\mathbb{S}, \leq), \mathbb{T})$ be a stationary time-domain and let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system with state space representation at t_0 prescribed by the data

$$\begin{aligned}
&X^{\Sigma}, \\
&\gamma^{\Sigma}_{t,t_0} \colon X^{\Sigma} \times U \to Y, && t \in \mathbb{T}_{\geq t_0}, \\
&\Phi^{\Sigma}_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^{\Sigma} \to X^{\Sigma}, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2.
\end{aligned}$$

The state space representation is

time-invariant from t_0 if, for $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, for $\mu \in \mathscr{U}_{\geq t_0}$, and for $x \in X^\Sigma$, it holds that

(i) $\tau^*_{t_0,t_1}(\mathscr{U}_{\geq t_0}) \subseteq \mathscr{U}_{\geq t_0}$,

(ii) $\gamma^\Sigma_{t_1,t_0}(x,u) = \gamma^\Sigma_{t_0,t_0}(x,u)$,

(iii) $\Phi^\Sigma_{t_2,t_1}(x, \mu_{[t_1,t_2)}) = \Phi^\Sigma_{t_2-t_1+t_0,t_0}(x, \hat{\tau}^*_{-(t_2-t_1)}\mu_{[t_1,t_2)})$, and

(iv) $\tau^*_{t_0,t_1}(\mathscr{Y}_{\geq t_0}) = \mathscr{Y}_{\geq t_0}$. •

Let us consider some examples of stationary and nonstationary systems.

2.2.41 Examples (Stationary and nonstationary systems)

1. We claim that deterministic finite state automata are stationary. To see this, suppose that we have a behaviour $(\mu, \eta) \in \mathscr{B}$ for such a system. This, then, is determined by an initial state $\theta(0) \in Q$, an input $\mu \in \Lambda^{\mathbb{Z}_{\geq 0}}$, the state sequence $(\theta(j))_{j\in\mathbb{Z}_{\geq 0}}$ determined by the dynamics δ, and the output sequence $(\eta(j))_{j\in\mathbb{Z}_{\geq 0}}$ determined by the output map γ. Given such a behaviour and given $k \in \mathbb{Z}_{\geq 0}$, we have $(\hat{\tau}^*_k\mu, \hat{\tau}^*_k\eta) \in \tau^*_k(\mathscr{B})$ defined by

$$\tau^*_k\mu(j) = \mu(j+k), \quad \tau^*_k\eta(j) = \eta(j+k).$$

If we define $\tau^*_k\theta(j) = \theta(j+k)$, then we see that the input sequence $\tau^*_k\mu$ and the initial state $\tau^*_k\theta(0)$ give rise to the output $\tau^*_k\eta$, and so we conclude that $\tau^*_k(\mathscr{B}) \subseteq \mathscr{B}$.

It is not generally the case that a deterministic finite state automaton is strongly stationary. To see this, we consider the system described by the diagram (2.5). We see that any behaviour (μ, η) of the form

$$\mu(j) = 0, \quad \eta(j) = \begin{cases} s_1, & j = 0, \\ s_2, & j \in \mathbb{Z}_{>0}, \end{cases} \qquad j \in \mathbb{Z}_{\geq 0},$$

cannot be recovered from a shifted behaviour $(\tau^*_k\mu', \tau^*_k\eta')$ for $k \in \mathbb{Z}_{>0}$. That is, $\tau^*_k(\mathscr{B}) \subset \mathscr{B}$ for $k \in \mathbb{Z}_{>0}$.

2. For the systems of most importance to us in this volume, those described by differential and difference equations, we shall give in Sections 6.1.1 and 6.3.1 a concise description of those that are stationary. Here we consider a simple example of this that gives some insight into the notion of stationarity.

Let $\mathscr{U} = \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$ and let $\mathscr{Y} \subseteq \mathsf{AC}(\mathbb{R}_{\geq 0}; \mathbb{R})$ be determined by

$$\dot{\eta}(t) = a(t)\eta(t) + \mu(t), \qquad \mu \in \mathscr{U},$$

for some $a \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$. We shall see in Example 4.3.6 that solutions of the preceding equation satisfying $\eta(t_0) = y_0$ are given by

$$\eta(t) = \Phi(t, t_0)y_0 + \Phi(t, t_0)\int_{t_0}^{t} \mu(\tau)\Phi(t_0, \tau)\mathrm{d}\tau$$

where

$$\Phi(t, t_0) = \int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau.$$

Let us first consider the case where a is constant, say $a(t) = \alpha$, $t \in \mathbb{R}_{\geq 0}$. In this case,

$$\Phi(t, t_0) = \mathrm{e}^{\alpha(t-t_0)},$$

and so, when $\alpha \neq 0$ (we shall leave the case of $\alpha = 0$ for the reader to work out),

$$\eta(t) = \mathrm{e}^{\alpha(t-t_0)} y_0 + \int_{t_0}^{t} \mathrm{e}^{\alpha(t-\tau)} \mu(\tau)\,\mathrm{d}\tau.$$

A behaviour is obtained by taking $t_0 = 0$, so that a general behaviour (μ, η) has the form

$$\mu \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R}), \quad \eta(t) = \mathrm{e}^{\alpha t} y_0 + \int_0^t \mathrm{e}^{\alpha(t-\tau)} \mu(\tau)\,\mathrm{d}\tau, \qquad t \in \mathbb{R}_{\geq 0}.$$

Now let $t_1 \in \mathbb{R}_{\geq 0}$ and compute

$$\tau^*_{0,t_1}\mu(t) = \mathrm{e}^{\alpha t} \underbrace{\left(\mathrm{e}^{-\alpha t_1} y_0 + \int_0^{t_1} \mathrm{e}^{-\alpha s} \mu(s - t_1)\,\mathrm{d}s \right)}_{y_0'} + \int_0^t \mathrm{e}^{\alpha(t-s)} \tau^*_{0,t_1}\mu(s)\,\mathrm{d}s.$$

Thus the shifted behaviour is again a behaviour with the new initial condition y_0'. This shows that $(\tau^*_{0,t_1}\mu, \tau^*_{0,t_1}\eta) \in \mathscr{B}$, and so the system is stationary. It is easy to see, by reversing the previous computations, that the system is, in fact, strongly stationary.

In the case when a is not constant, then one can see that the system is not stationary. We shall prove this in a more general setting in Proposition 6.1.7. •

2.2.9 Linear time systems

The next significant bit of structure we introduce for time systems is that of linearity. Special classes of linear time systems will receive substantial attention in subsequent chapters in this volume. Here we consider a fairly general framework for linear time systems.

An essential observation when dealing with linear time systems is that, for a field F, an F-vector space V, and a general time-domain $\mathbb{T}$, the set $\mathsf{V}^{\mathbb{T}}$ is an F-vector space, as demonstrated in Example I-4.5.2–8. Note that $\mathsf{V}^{(\mathbb{T})}$ is *not* an F-vector space since partial time functions cannot generally be added as they have different domains. Nonetheless, we can make the following definition.

2.2.42 Definition (Linearly closed set of partial time functions) Let F be a field, let V be an F-vector space, and let $\mathbb{T}$ be a general time domain. A subset $\mathscr{V} \subseteq \mathsf{V}^{(\mathbb{T})}$ of partial time functions is ***linearly closed*** if, for $a \in \mathsf{F}$ and for $\nu, \nu_1, \nu_2 \in \mathsf{V}^{(\mathbb{T})}$ with $\operatorname{dom}(\nu_1) = \operatorname{dom}(\nu_2)$, we have $\nu_1 + \nu_2 \in \mathscr{V}$ and $a\nu \in \mathscr{V}$, where

$$(\nu_1 + \nu_2)(t) = \nu_1(t) + \nu_2(t), \qquad t \in \operatorname{dom}(\nu_1) = \operatorname{dom}(\nu_2),$$

and $(a\nu)(t) = a(\nu(t))$ for $t \in \operatorname{dom}(\nu)$. •

With this in mind, we make the following definition.

2.2.43 Definition (Linear time system) Let F be a field. An ***F-linear time system*** is a general time system

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

for which

(i) U and Y are F-vector spaces,

(ii) $\mathscr{U} \subseteq \mathsf{U}^{(\mathbb{T})}$ and $\mathscr{Y} \subseteq \mathsf{Y}^{(\mathbb{T})}$ are linearly closed, and

(iii) $\mathscr{B} \subseteq (\mathsf{U} \oplus \mathsf{Y})^{(\mathbb{T})}$ is linearly closed. •

For linear time systems, it is reasonable and convenient to restrict the generality and work with signals all of which are restrictions of signals defined on the entire time-domain. Such systems will of necessity be complete. Let us single out such systems.

2.2.44 Definition (Full-time linear time system) Let F be a field. An F-linear time system

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

is ***full-time*** if

$$\operatorname{dom}(\mu) = \operatorname{dom}(\eta) = \operatorname{dom}(\mu, \eta) = \mathbb{T}$$

for every $\mu \in \mathscr{U}$, $\eta \in \mathscr{Y}$, and $(\mu, \eta) \in \mathscr{B}$. The ***core*** of such a system is the subspace

$$\mathscr{B}(0) = \{\eta \in \mathscr{Y} \mid (0, \eta) \in \mathscr{B}\}.$$ •

If Σ is a full-time linear time system with time-domain $\mathbb{T}$ and if $\mathbb{T}' \subseteq \mathbb{T}$ is a sub-time-domain, then the restrictions of inputs, outputs, and behaviours are subspaces of the vector spaces of inputs, outputs, and behaviours, respectively.

Linear time systems have the feature that, under some mild assumptions, one can draw useful conclusions about their system theoretic structure. Let us illustrate this by giving a class of linear time systems that turn out to be very structured.

2.2.45 Definition (Basic linear time system) Let F be a field. An F-linear time system

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

is ***basic*** if it is

(i) full-time,
(ii) strongly causal, and
(iii) strongly stationary, and
(iv) if it has a finite-dimensional core. •

We shall see that all of the linear systems we work with in this volume fit into the category of basic linear time systems. Thus anything we can prove about this class of systems will hold in generality for the linear systems we consider in detail subsequently. Let us, therefore, begin to establish a few facts about basic linear time systems.

We begin with the property of finite observability introduced in Definition 2.2.33.

2.2.46 Proposition (Basic linear time systems are finitely observable) *Let* F *be a field and let*

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B}),$$

be an F*-linear time system for which* $\operatorname{card}(\mathbb{T}) > 1$. *Let* $\mathrm{t}_0 \in \mathbb{T}$. *Then the following equivalent statements hold for* $\mathrm{t}_0 \in \mathbb{T}$*:*

(i) Σ *is finitely observable from some* $\tau \in \mathbb{T}_{\geq \mathrm{t}_0}$*;*

(ii) there exists $\tau \in \mathbb{T}_{>\mathrm{t}_0}$ *such that, if* $\eta \in \mathscr{B}(0)_{\geq \mathrm{t}_0}$ *satisfies* $\eta_{[\mathrm{t}_0,\tau)} = 0_{[\mathrm{t}_0,\tau)}$*, then* $\eta_{\geq \mathrm{t}_0} = 0_{\geq \mathrm{t}_0}$.

Proof Let us first show that the two conditions are equivalent. It is clear that (i) implies (ii). Now, suppose that (ii) holds for some $\tau \in \mathbb{T}_{>t_0}$ and let $\mu \in \mathscr{U}$. Suppose that, for $\eta_1, \eta_2 \in \mathscr{B}(\mu)_{\geq t_0}$, we have

$$(\eta_1)_{[t_0,\tau)} = (\eta_2)_{[t_0,\tau)}.$$

Then

$$(\eta_1)_{[t_0,\tau)} - (\eta_2)_{[t_0,\tau)} = 0.$$

Also,

$$(\mu, \eta_1), (\mu, \eta_2) \in \mathscr{B}_{\geq t_0} \implies (0, \eta_1 - \eta_2) \in \mathscr{B}_{\geq t_0} \implies \eta_1 - \eta_2 \in \mathscr{B}(0)_{\geq t_0},$$

and so we conclude that

$$\eta_1 - \eta_2 = 0 \implies \eta_1 = \eta_2.$$

Next we show that a basic linear time system satisfies (ii). We claim that the assumption that $\operatorname{card}(\mathbb{T}) > 1$ implies that $\operatorname{card}(\mathbb{T}_{t_0}) \geq \operatorname{card}(\mathbb{Z}_{>0})$. Indeed, we have $t \in \mathbb{T}_{t_0}$ such that $t \neq 0$. Therefore, $kt \in \mathbb{T}_{t_0}$ for all $k \in \mathbb{Z}_{>0}$. We claim that the times kt, $k \in \mathbb{Z}_{>0}$, are distinct. Suppose not, so that $k_1, k_2 \in \mathbb{Z}_{>0}$ satisfy $k_2 > k_1$ and

$$k_1 t = k_2 t \implies (k_2 - k_1)t = 0.$$

Since $t \neq 0$, the property of the total order gives $k_1 = k_2$.

Given the preceding, let $\tau \in \mathbb{T}_{\geq t_0}$ be such that $\operatorname{card}(\mathbb{T}_{[t_0,\tau)}) \geq \dim_{\mathsf{F}}(\mathscr{B}(0)_{\geq t_0})$. Let $\eta \in \mathscr{B}(0)_{\geq t_0}$ have the property that $\eta_{[t_0,\tau)} = 0_{[t_0,\tau)}$. Suppose that $\eta \neq 0$ so that there exists $T \in \mathbb{T}_{\geq \tau}$ such that

$$T = \sup\{t \in \mathbb{T}_{\geq t_0} \mid \eta(t) = 0\}.$$

Let $\delta > 0$ be such that $\eta(T + \delta) \neq 0$. Let $t_1, \ldots, t_k \in \mathbb{T}_{[t_0,T]}$ with $k \geq \dim_{\mathsf{F}}(\mathscr{B}(0)_{\geq t_0})$ and satisfying

$$t_0 < t_1 < \cdots < \tau_k = T.$$

Define $\eta_j = \tau^*_{T-t_{k-j}}\eta$, $j \in \{0, 1, \ldots, k\}$. For $j \in \{0, 1, \ldots, k\}$, we then have

$$\begin{aligned}
&\eta \in \mathscr{B}(0)_{\geq t_0}\\
\implies &(0, \eta) \in \mathscr{B}_{\geq t_0}\\
\implies &(0_{\geq t_0+T-t_{k-j}}, \eta_{\geq t_0+T-t_{k-j}}) \in \mathscr{B}_{\geq t_0+T-t_{k-j}} = \tau^*_{T-t_j}(\mathscr{B}_{\geq t_0})\\
\implies &(0_{\geq t_0+T-t_{k-j}}, \eta_{\geq t_0+T-t_{k-j}}) = (\tau^*_{T-t_j}\mu', \tau^*_{T-t_j}\eta') \text{ for some } (\mu', \eta') \in \mathscr{B}_{\geq t_0}\\
\implies &\mu' = \tau^*_{t_j-t_0}0_{\geq t_0+T-t_{k-j}} = 0,\ \eta' = \tau^*_{T-t_{k-j}}\eta_{\geq t_0+T-t_{k-j}}\\
\implies &\eta' = \eta_j \in \mathscr{B}(0)_{\geq t_0},
\end{aligned}$$

using the fact that Σ is strongly stationary. We claim that $\eta_0, \eta_1, \ldots, \eta_k$ are linearly independent. Indeed, suppose that

$$c_0\eta_0 + c_1\eta_1 + \cdots + c_k\eta_k = 0$$

for some $c_0, c_1, \ldots, c_k \in \mathsf{F}$. We then have

$$\eta_j(t_{k-j} + \delta) = \eta(T + \delta) = 0, \qquad j \in \{0, 1, \ldots, k\},$$

and

$$\eta_j(t_{k-l} + \delta) = \eta(T - (t_{k-j} - t_{k-l}) + \delta) = 0, \qquad l \in \{0, 1, \ldots, j-1\},$$

since $\tau_{k-j} - t_{k-l} > 0$. Thus

$$\eta_0(t_0 + \delta) = \eta_1(t_0 + \delta) = \cdots = \eta_{k-1}(t_0 + \delta) = 0$$

and so $c_k\eta_k(t_0 + \delta) = 0$, whence $c_k = 0$. In like manner,

$$\eta_0(t_1 + \delta) = \eta_1(t_1 + \delta) = \cdots = \eta_{k-2}(t_1 + \delta) = 0,$$

which gives $c_{k-1}\eta_{k-1}(t_1 + \delta) = 0$, and so $c_{k-1} = 0$. Continuing in this way, $c_0 = c_1 = \cdots = c_k = 0$, giving the asserted linear independence. This contradicts the fact that $\dim_{\mathsf{F}}(\mathscr{B}(0)_{\geq t_0}) \leq k$, and so we must have $\eta = 0$. ■

This gives the following important corollary about the characteristics of basic linear time systems.

2.2.47 Corollary (Basic linear time systems are past-determined) *Let* F *be a field, let*

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be an F*-linear time system, and let* $t_0 \in \mathbb{T}$*. Then there exists* $\tau \in \mathbb{T}_{>t_0}$ *such that* Σ *is strongly past-determined from* τ*.*

Proof This follows immediately from Proposition 2.2.34. ■

Next we consider dynamical system and state space representations for basic linear general time systems. One hopes that one can arrive at such representations that preserve the linearity of the system. So let us first encode the desired properties of these representations.

2.2.48 Definition (Linear dynamical system and state space representations) Let F be a field and let

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a full-time linear time system and let $t_0 \in \mathbb{T}$.

(i) A dynamical system representation determined by the data

$$\begin{aligned} &\mathsf{X}^\Sigma, \\ &\rho^\Sigma_{t,t_0} \colon \mathsf{X}^\Sigma \oplus (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0}, \\ &\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \oplus \mathsf{X}^\Sigma \to \mathsf{X}^\Sigma, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2, \end{aligned}$$

is **F-*linear*** if

(a) X^Σ is an F-vector space and

(b) the mappings ρ^Σ_{t,t_0}, $t \in \mathbb{T}_{\geq 0}$, and $\Phi^\Sigma_{t_2,t_1}$, $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, $t_1 \geq t_2$, are F-linear.

(ii) A state space representation determined by the data

$$\begin{aligned} &\mathsf{X}^\Sigma, \\ &\gamma^\Sigma_{t,t_0} \colon \mathsf{X}^\Sigma \oplus \mathsf{U} \to \mathsf{Y}, && t \in \mathbb{T}_{\geq t_0}, \\ &\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \oplus \mathsf{X}^\Sigma \to \mathsf{X}^\Sigma, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2, \end{aligned}$$

is **F-*linear*** if

(a) X^Σ is an F-vector space and

(b) the mappings γ^Σ_{t,t_0}, $t \in \mathbb{T}_{\geq 0}$, and $\Phi^\Sigma_{t_2,t_1}$, $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, $t_1 \geq t_2$, are F-linear.

A significant result for basic linear time systems is then the following.

2.2.49 Proposition (Representations for basic linear time systems) *Let* F *be a field. For a basic* F*-linear time system*

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B}),$$

the following statements hold for $\mathsf{t}_0 \in \mathbb{T}$*:*

(i) Σ *possesses a strongly causal, strongly stationary linear dynamical system representation from* t_0*;*

(ii) Σ *possesses a strongly causal, strongly stationary linear state space representation from* t_0*.*

Proof By Corollary 2.2.47 we have that Σ is past-determined from some $\tau \in \mathbb{T}_{\geq t_0}$. Define

$$\rho'_\tau \colon \mathscr{U}_{\geq\tau} \to \mathscr{Y}_{\geq\tau}$$

by requiring that

$$\rho'_\tau(\mu_{\geq\tau}) = \eta_{\geq\tau} \iff (0_{[t_0,\tau)} * \mu_{\geq\tau}, 0_{[t_0,\tau)} * \eta_{\geq\tau}) \in \mathscr{B}_{\geq t_0}.$$

We claim that ρ'_τ is well-defined. First of all, if $\mu_{\geq\tau} \in \mathscr{U}_{\geq\tau}$, let η' be such that

$$(0_{[t_0,\tau)} * \mu_{\geq\tau}, \eta') \in \mathscr{B}_{\geq t_0}.$$

Then, by strong causality of Σ, we must have $\eta'_{[t_0,\tau)} = 0_{[t_0,\tau)}$. Next, if

$$(0_{[t_0,\tau)} * \mu_{\geq\tau}, 0_{[t_0,\tau)} * \eta_{\geq\tau}), (0_{[t_0,\tau)} * \mu_{\geq\tau}, 0_{[t_0,\tau)} * \eta'_{\geq\tau}) \in \mathscr{B}_{\geq t_0} \implies \eta_{\geq\tau} = \eta'_{\geq\tau}$$

since Σ is past-determined from τ. This shows that ρ'_τ is well-defined. It is also evident that ρ'_τ is F-linear.

Next define $\mathsf{X}^\Sigma = \mathscr{B}(0)_{\geq t_0}$ and define

$$\begin{aligned} \rho''_\tau \colon \mathsf{X}^\Sigma &\to \mathscr{Y}_{\geq\tau} \\ \eta &\mapsto \hat{\tau}^*_{\tau - t_0}\eta. \end{aligned}$$

Clearly ρ''_τ is linear.

Now define

$$\begin{aligned} \rho^\Sigma_\tau \colon X^\Sigma \oplus \mathscr{U}_{\geq\tau} &\to \mathscr{Y}_{\geq\tau} \\ (\eta, \mu_{\geq\tau}) &\mapsto \rho'_\tau(\mu_{\geq\tau}) + \rho''_\tau(\eta). \end{aligned}$$

Linearity of Σ, and the fact that $\hat{\tau}^*_{t_0,\tau}(\mathscr{B}(0)_{\geq t_0}) = \mathscr{B}_\tau(0_{\geq\tau})$, shows that this mapping is well-defined and linear. Moreover, let $(\mu_{\geq\tau}, \eta_{\geq\tau}) \in \mathscr{B}_{\geq\tau}$. Then we write

$$(\mu_{\geq\tau}, \eta_{\geq\tau}) = (\mu_{\geq\tau}, \rho'_\tau(\mu_{\geq\tau})) + (0_{\geq\tau}, \eta_{\geq\tau} - \rho'_\tau(\mu_{\geq\tau})),$$

from which we deduce

$$(\mu_{\geq\tau}, \rho'_\tau(\mu_{\geq\tau})) \in \mathscr{B}_{\geq\tau} \implies \eta_{\geq\tau} - \rho'_\tau(\mu_{\geq\tau}) \in \mathscr{B}_{\geq\tau}(0_{\geq\tau}) = \hat{\tau}_{\tau - t_0}(\mathsf{X}^\Sigma),$$

and so

$$\eta_\tau = \rho'_\tau(\mu_{\geq\tau}) + \rho''_\tau(\eta') = \rho^\Sigma_\tau(\eta, \mu_{\geq\tau})$$

for some $\eta' \in \mathsf{X}^\Sigma$. This shows that ρ^Σ_τ is an initial response function for $\mathscr{B}_{\geq\tau}$.

We next claim that ρ^Σ_τ is strongly causal. Indeed,

$$\mu_{[\tau,t)} = \mu'_{[\tau,t)} \implies 0_{[t_0,\tau)} * \mu_{[\tau,t)} = 0_{[t_0,\tau)} * \mu'_{[\tau,t)} \implies (\rho'_\tau(\mu_{\geq\tau}))_{[\tau,t]} = (\rho'_\tau(\mu'_{\geq\tau}))_{[\tau,t]},$$

from which we deduce that ρ^Σ_τ is indeed strongly causal.

The above shows that ρ^Σ_τ is a linear, strongly causal initial response function for $\mathscr{B}_{\geq\tau}$.

Now define

$$\begin{aligned}\rho^\Sigma_{t_0} &: \mathsf{X}^\Sigma \oplus \mathscr{U}_{\geq t_0} \to \mathscr{Y}_{\geq t_0}\\ &\quad (\eta, \mu) \mapsto \hat\tau^*_{-(\tau-t_0)}\rho^\Sigma_\tau(\eta, \hat\tau^*_{\tau-t_0}\mu).\end{aligned}$$

Strong stationarity of Σ implies that $\rho^\Sigma_{t_0}$ is a linear strong causal initial response function for $\mathscr{B}_{\geq t_0}$.

Now define

$$\begin{aligned}\rho^\Sigma_{t,t_0} &: \mathsf{X}^\Sigma \oplus \mathscr{U}_{\geq\tau} \to \mathscr{Y}_{\geq\tau}\\ &\quad (\eta, \mu_{\geq\tau}) \mapsto \hat\tau^*_{-(\tau-t_0)}\rho^\Sigma_0(\eta, \hat\tau^*_{\tau-t_0}\mu_{\geq\tau})\end{aligned}$$

and

$$\begin{aligned}\Phi^\Sigma_{t,t_0} &: \mathscr{U}_{[t_0,t)} \oplus \mathsf{X}^\Sigma \to \mathsf{X}^\Sigma\\ &\quad (\mu_{[t_0,t)}, \eta) \mapsto \hat\tau^*_{-(t-t_0)}\rho^\Sigma_{t_0}(\eta, \mu_{[t_0,t)} * 0_{\geq t}),\end{aligned}$$

keeping in mind that $\mathsf{X}^\Sigma = \mathscr{B}(0)_{\geq t_0}$. Finally, for $t_1, t_2 \in \mathbb{T}_{\geq t_0}$ with $t_1 \leq t_2$, define

$$\begin{aligned}\Phi^\Sigma_{t_1,t_2} &: \mathscr{U}_{[t_1,t_2)} \oplus \mathsf{X}^\Sigma \to \mathsf{X}^\Sigma\\ &\quad (\mu_{[t_1,t_2)}, \eta) \mapsto \Phi^\Sigma_{t_2-t_1+t_0,t_0}(\hat\tau_{-(\tau-t_0)}(\mu_{[t_1,t_2)}), \eta).\end{aligned}$$

One then shows, by direct computation, that the data

$$\begin{aligned}&\mathsf{X}^\Sigma,\\ &\rho^\Sigma_{t,t_0} : \mathsf{X}^\Sigma \oplus (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0},\\ &\Phi^\Sigma_{t_2,t_1} : \mathscr{U}_{[t_1,t_2)} \oplus \mathsf{X}^\Sigma \to \mathsf{X}^\Sigma, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2,\end{aligned}$$

is an F-linear, strongly causal, and strongly stationary dynamical system representation for Σ from t_0.

Finally, if we define

$$\begin{aligned}\gamma^\Sigma_{t,t_0} &: \mathsf{X}^\Sigma \oplus \mathsf{U} \to \mathsf{Y}\\ &\quad (\eta, u) \mapsto \rho^\Sigma_{t,t_0}(\eta, \mu_{\geq t})(t),\end{aligned}$$

where $\mu_{\geq t}$ satisfies $\mu(t) = u$. This, then, gives the data

$$\begin{aligned}&\mathsf{X}^\Sigma,\\ &\gamma^\Sigma_{t,t_0} : \mathsf{X}^\Sigma \oplus \mathsf{U} \to \mathsf{Y}, && t \in \mathbb{T}_{\geq t_0},\\ &\Phi^\Sigma_{t_2,t_1} : \mathscr{U}_{[t_1,t_2)} \oplus \mathsf{X}^\Sigma \to \mathsf{X}^\Sigma, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2,\end{aligned}$$

which is an F-linear, strongly stationary state space representation for Σ. ■

Let us give some examples of linear time systems, noting that in Sections 6.6, 6.8, 6.7, and 6.9 we shall consider large classes of such systems.

2.2.50 Examples (Linear time systems)

1. The system of Example 2.2.41–2 is a $\mathbb{R}$-linear time system. Since it is not generally stationary, it is not generally a basic linear time system. It is, however, a basic $\mathbb{R}$-linear time system when a is constant.
2. The convolution system of Example 2.1.14 is a basic $\mathbb{R}$-linear time system. •

Exercises

2.2.1 Let $(\mathbb{T}, \leq)$ be a general time-domain with $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{T}$ sub-time-domains such that $\mathbb{S}_2$ follows $\mathbb{S}_1$. Show that $\mathbb{S}_1 * \mathbb{S}_2$ is a sub-time-domain of $\mathbb{T}$.

2.2.2 For each of the following general time-domains, describe explicitly the pairs of sub-time-domains $\mathbb{S}_1$ and $\mathbb{S}_2$ for which $\mathbb{S}_2$ follows $\mathbb{S}_1$:

(a) $(\mathbb{Z}(\Delta), \leq)$;
(b) $(\mathbb{Q}, \leq)$;
(c) $(\mathbb{R}, \leq)$.

2.2.3 Show that the continuous and discrete time domains of Definition IV-1.1.2 are additive time-domains in the sense of Definition 2.2.3.

2.2.4 Answer the following questions.

(a) Describe explicitly all possible stationary time-domains that are subsets of the additive time-domain $((\mathbb{R}, \leq), \mathbb{R})$.
(b) Describe explicitly all possible stationary time-domains that are subsets of the additive time-domain $((\mathbb{Z}(\Delta), \leq), \mathbb{Z}(\Delta))$.

2.2.5 (Only for students in countries that use dollars and cents as currency.) Consider a soda machine for which a can of soda costs 50¢ and which takes only quarters as inputs. Model this as a deterministic finite state automaton $(Q, \Lambda, Y, \delta, \mu)$.

2.2.6 Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system and let $\rho^\Sigma_{t_0} \colon X^\Sigma \times \mathscr{U}_{\geq t_0} \to \mathscr{Y}_{\geq t_0}$ be an initial response function with initial state object X^Σ. Define

$$\begin{aligned}
&X^\Sigma_{t,t_0} = X^\Sigma \times \mathscr{U}_{[t_0,t)}, && t \in \mathbb{T}_{\geq t_0},\\
&\rho^\Sigma_{t,t_0} \colon X^\Sigma_{t,t_0} \times \mathscr{U}_{\geq t} \to \mathscr{Y}_{\geq t} && \\
&\qquad ((x, \mu_{[t_0,t)}), \mu'_{\geq t}) \mapsto \rho^\Sigma_{t_0}(x, \mu_{[t_0,t)} * \mu'_{\geq t})_{\geq t}, && t \in \mathbb{T}_{t,t_0},\\
&\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma_{t_1,t_0} \to X^\Sigma_{t_2,t_0} && \\
&\qquad (\mu'_{\geq t_2}, (x, \mu_{[t_1,t_2)})) \mapsto (\mu_{[t_1,t_2)}) * \mu'_{\geq t_2}, x), && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2.
\end{aligned}$$

Show that this data describes a pre-dynamical system representation for Σ at t_0. (This is called the ***Nerode realisation***.)

Section 2.3

Some problems in general system theory

The preceding sections were dedicated to a description of a general system theory, and to looking at specific structures within this theory, but still in a general setting. While it is interesting to play around within this world of general system theory structure, the fact of the matter is that the theory is aimed at solving problems. In this section we describe such problems, but maintain our general and abstract setting. This generality and abstraction is helpful for understanding the problems themselves in the absence of distracting additional structure.

2.3.1 Goal-seeking

Goal-seeking can be seen as additional structure within a system that accounts for its behaviour. The additional structure is put in place to describe specific attributes of the system. We shall give a general definition of what we mean by a goal-seeking system, then illustrate the general concepts with a specific example.

We shall give the formal definition, which will be difficult to contextualise initially, and then we will discuss in broad terms how this works. We hope that the example that follows will make this more clear.

2.3.1 Definition (Goal-seeking system) A general input/output system $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$ is a ***goal-seeking system*** if there exist

(i) a set $\mathscr{M}$ (the ***internal model space***),

(ii) a totally ordered set $(V, \leq)$ (the ***value space***),

(iii) a mapping $G\colon \mathscr{M} \times \mathscr{U} \times \mathscr{Y} \to V$ (the ***goal function***),

(iv) a subset $\mathscr{F} \subseteq \mathscr{U} \times \mathscr{Y} \times \mathscr{M}$ (the ***internal goal-seeking system***),

(v) a subset $\mathscr{P} \subseteq \mathscr{M} \times \mathscr{U} \times \mathscr{Y}$ (the ***parameterised systems***), and

(vi) a subset $\mathscr{E} \subseteq \mathscr{U} \times \mathscr{Y} \times V \times \mathscr{M}$ (the ***the selection system***)

that satisfy the following conditions:

(vii) $\mathscr{E} = \{(\mu, \eta, v, \lambda) \mid v = G(\lambda, \mu, \eta),\ v \leq G(\lambda', \mu', \eta'),\ (\lambda', \mu', \eta') \in \mathscr{M} \times \mathscr{U} \times \mathscr{Y}\}$;

(viii) $\mathscr{F} = \{(\mu, \eta, \lambda) \mid (\mu, \eta, G(\lambda, \mu, \eta), \lambda) \in \mathscr{E}\}$;

(ix) $\mathscr{B} = \{(\mu, \eta) \mid (\lambda, \mu, \eta) \in \mathscr{P}$ and $(\mu, \eta, \lambda) \in \mathscr{F}$ for some $\lambda \in \mathscr{M}\}$. •

Let us discuss the definition, and how one should think of how the components fit together. In goal-seeking systems, one is seeking, given an input, an output that is optimal in the sense that it minimises the goal function G. However, the minimisation may be carried out over variables that are not strictly described by the inputs and outputs, and this explains the introduction of the internal model space $\mathscr{M}$. The selection system $\mathscr{E}$ describes the set of optimal inputs, outputs,

internal model variables, and values that are optimal in the sense set out in part (vii). The selection system feeds into the the system through the internal goal-seeking system $\mathscr{F}$. One should think of $\mathscr{F}$ as describing, given an input μ and an output η, the set of possible internal variables λ that are selected as optimal by the selection system. The set of parameterised systems should be thought of as prescribing, for each $\lambda \in \mathscr{M}$, a subset

$$\mathscr{B}_\lambda = \{(\mu, \eta) \in \mathscr{U} \times \mathscr{Y} \mid (\lambda, \mu, \eta) \in \mathscr{P}\}$$

of behaviours. Thus $\mathscr{P}$ selects, given an input μ, determines whether there exists an internal variable λ and an output η for which the data (μ, λ, η) is optimal.

Let us illustrate this with an example.

2.3.2 Example (Length minimisation as goal-seeking) Let us consider a system designed to produce a curve that connects points $x_0, x_1 \in \mathbb{R}^n$ while minimising the length of the curve. Thus we consider the input space to be the singleton $\mathscr{U} = \{(x_0, x_1)\} \subseteq \mathbb{R}^n \times \mathbb{R}^n$ and the output space is the set

$$\mathscr{Y} = \mathsf{H}^1([0,1]; \mathbb{R}^n) = \{\gamma \colon [0,1] \to \mathbb{R}^n \mid \gamma \text{ is locally absolutely continuous and } \gamma' \in \mathsf{L}^2([0,1]; \mathbb{R}^n)\}$$

of absolutely continuous curves with domain $[0,1]$ and whose derivative is square integrable. We know what the system $\mathscr{B} \subseteq \mathscr{U} \times \mathscr{Y}$ is because we know that straight lines minimise length; thus we know that the system is

$$\mathscr{B} = \{((x_0, x_1), \gamma) \in \mathscr{U} \times \mathscr{Y} \mid \gamma(t) = (1-t)x_0 + tx_1,\ t \in [0,1]\}.$$

What we will do, however, is make this length minimisation a part of the system by prescribing a goal-seeking structure.

We take

$$\mathscr{M} = \{\gamma \in \mathscr{Y} \mid \gamma(0) = x_0,\ \gamma(1) = x_1\}.$$

We take $V = \mathbb{R}$ with its standard total order and define the cost function G by the length:

$$G \colon \mathscr{M} \times \mathscr{U} \times \mathscr{Y} \to V$$
$$(\hat{\gamma}, (x_0, x_1), \gamma) \mapsto \int_0^1 \|\hat{\gamma}'(t)\| \, \mathrm{d}t.$$

We then take

$$\mathscr{E} = \left\{ ((x_0, x_1), \gamma, \ell, \hat{\gamma}) \in \mathscr{U} \times \mathscr{Y} \times V \times \mathscr{M} \,\middle|\, \int_0^1 \|\hat{\gamma}'(t)\| \, \mathrm{d}t = \ell,\ \ell \leq \int_0^1 \|\overline{\gamma}'(t)\| \, \mathrm{d}t,\ \overline{\gamma} \in \mathscr{M} \right\}$$

Thus $\mathscr{E}$ contains the solution to the problem: Find the curve connecting x_0 and x_1 of minimum length. This then gives

$$\mathscr{F} = \{((x_0, x_1), \gamma, \hat{\gamma}) \in \mathscr{U} \times \mathscr{Y} \times \mathscr{M} | \\ \hat{\gamma} \text{ is the curve of minimum length connecting } x_0 \text{ and } x_1\}.$$

We next connect the solution to the optimisation problem to the larger system by taking

$$\mathscr{P} = \{((x_0, x_1), \gamma, \hat{\gamma}) \in \mathscr{U} \times \mathscr{Y} \times \mathscr{M} \mid \hat{\gamma} = \gamma\},$$

which gives

$$\mathscr{B} = \{((x_0, x_1), \gamma) \in \mathscr{U} \times \mathscr{Y} | \\ \hat{\gamma} \text{ is the curve of minimum length connecting } x_0 \text{ and } x_1\}.$$

Note that the goal-seeking idea is contained in the definition of $\mathscr{E}$, and one must devise some method to show that

$$\mathscr{E} = \{((x_0, x_1), \gamma, \ell, \hat{\gamma} \in \mathscr{U} \times \mathscr{Y} \times V \times \mathscr{M} | \\ \hat{\gamma}(t) = (1-t)x_0 + tx_1,\ t \in [0,1], \|x_1 - x_1\| = \ell\}.$$

The idea is that this process is decoupled from the macroscopic input/output behaviour of the system. •

2.3.2 Decision problems

Decision problems are important in theoretical computer science, where they are a part of the theory of computational complexity. The definition of a decision problem is simple.

2.3.3 Definition (Decision problem) A ***decision problem*** is a functional input/output system $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$ where $\mathscr{Y} = \{\text{yes}, \text{no}\}$. •

Many problems arise as decision problems, or can be converted to decision problems.

2.3.4 Examples (Decision problem)

1. The automaton of Example 2.1.2–2 can be seen as providing an answer to the question "Does a given input string $\boldsymbol{w} = w_1 w_2 \cdots w_k$ contain an even number of number of zeros?" Indeed, the automaton is designed to return a "yes" answer when $q_k = s_1$ and a "no" answer when $q_k = s_2$.
2. General input/output systems correspond to decision problems. Indeed, if $\Sigma = (\mathscr{U}, \mathscr{Y}, \mathscr{B})$ is a general input/output system, then we associated to this the decision problem $\Sigma_{\text{dec}} = (\mathscr{U}_{\text{dec}}, \{\text{yes}, \text{no}\}, \mathscr{B}_{\text{dec}})$ by taking $\mathscr{U}_{\text{dec}} = \mathscr{U} \times \mathscr{Y}$ and

$$\mathscr{B}_{\text{dec}} = \{((\mu, \eta), \text{yes}) \mid (\mu, \eta) \in \mathscr{B}\} \cup \{((\mu, \eta), \text{no}) \mid (\mu, \eta) \notin \mathscr{B}\}.$$

3. There is a relationship between optimisation problems and decision problems. Let us illustrate this concretely. Let us consider a graph $(\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V}$ and edges $\mathcal{E}$. For $k \in \mathbb{Z}_{>0}$, we consider a system $\Sigma_k = (\mathscr{U}, \{\text{yes}, \text{no}\}, \mathscr{B}_k)$, where
   (a) $\mathscr{U}$ is the set of all pairs $(v_1, v_2) \in \mathcal{V} \times \mathcal{V}$ of vertices and
   (b) $((v_1, v_2), \text{yes}) \in \mathscr{B}_k$ if and only if there exists a path in the graph from v_1 to v_2 with length at most k.

   In this case, one can consider an optimisation problem: Given vertices (v_1, v_2), find the shortest path in the graph connecting v_1 and v_2. A solution to this optimisation problem, if it exists, can be obtained by considering in sequence the systems $\Sigma_1, \Sigma_k, \dots$ and then noting that a shortest path will be obtained by the smallest k for which $((v_1, v_2), \text{yes}) \in \mathscr{B}_k$. •

2.3.3 Reachability

Reachability, more or less, answers the question, "Where can I go from here?" There are many possible variations of reachability questions, and indeed the "correct" variation typically relies heavily on the precise properties of the system with which one is working. Here we consider some simple examples of the sorts of questions one might consider.

2.3.5 Example (Reachability problems)

1. A typical reachability problem can be posed as follows. Consider a dynamical system representation given by the data

$$\begin{aligned} &X^\Sigma, \\ &\rho^\Sigma_{t,t_0} \colon X^\Sigma \times (\mathscr{U}_{\geq t_0})_{\geq t} \to (\mathscr{Y}_{>t_0})_{\geq t}, && t \in \mathbb{T}_{\geq t_0}, \\ &\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma \to X^\Sigma, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2. \end{aligned}$$

   Suppose we are given $x_1, x_2 \in X^\Sigma$ and $t_1, t_2 \in \mathbb{T}_{\geq t_0}$ with $t_1 \leq t_2$. We say that the state x_2 is ***reachable*** in time t_2 from state x_1 at time t_1 if there exists $\mu_{\in[t_1,t_2)} \in \mathscr{U}_{[t_1,t_2)}$ such that $x_2 = \Phi^\Sigma_{t_2,t_1}(\mu_{[t_1,t_2)}, x_1)$. This sort of problem can be referred to as "state reachability."
2. If, additionally, we have a state space representation given by the data

$$\begin{aligned} &X^\Sigma, \\ &\gamma^\Sigma_{t,t_0} \colon X^\Sigma \times U \to Y, && t \in \mathbb{T}_{\geq t_0}, \\ &\Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma \to X^\Sigma, && t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2, \end{aligned}$$

   then the preceding notion of controllability can be adapted to outputs. Thus, given $y_1, y_2 \in Y$ and $t_1, t_2 \in \mathbb{T}_{\geq t_0}$ with $t_1 \leq t_2$. We say that the output y_2 is ***reachable*** in time t_2 from output y_1 at time t_1 if there exists $\mu_{\in[t_1,t_2)} \in \mathscr{U}_{[t_1,t_2)}$ and $x \in X^\Sigma$ such that

$$y_a = \gamma^\Sigma_{t_a,t_0}(\Phi^\Sigma_{t_a,t_0}(\mu_{[t_0,t_a)}, x)), \qquad a \in \{1, 2\}.$$

   This sort of problem can be referred to as "output reachability." •

2.3.4 Observability

Observability is concerned with the extent to which a knowledge of the input/output behaviour of a system determines its input/state behaviour. As with reachability, there are specific forms of this sort of property, depending on the exact attributes of the system. However, one can give a useful definition in the context of time systems, so let us do this.

2.3.6 Definition (Observable general time system) Let

$$\Sigma = (U, Y, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathscr{B})$$

be a general time system, let $t_0 \in \mathbb{T}$, and suppose that Σ possesses a state space representation at t_0 prescribed by the data

$$\begin{array}{ll} X^\Sigma, & \\ \gamma^\Sigma_{t,t_0} \colon X^\Sigma \times U \to Y, & t \in \mathbb{T}_{\geq t_0}, \\ \Phi^\Sigma_{t_2,t_1} \colon \mathscr{U}_{[t_1,t_2)} \times X^\Sigma \to X^\Sigma, & t_1, t_2 \in \mathbb{T}_{\geq t_0},\ t_1 \leq t_2. \end{array}$$

Then Σ is ***observable*** if, for $(\mu_1, \eta_1), (\mu_2, \eta_2) \in \mathscr{B}_{\geq t_0}$, we have

$$(\mu_1, \eta_1) = (\mu_2, \eta_2) \implies \Phi^\Sigma_{t,t_0}((\mu_1)_{[t_0,t)}, x) = \Phi^\Sigma_{t,t_0}((\mu_2)_{[t_0,t)}, x)$$

for every $x \in X^\Sigma$. •

The idea is that, if one knows the input/output behaviour, then the state behaviour is completely determined by the initial condition.

2.3.5 Stability

Stability has to do, loosely speaking, with the problem: Do behaviours that are close at some time remain close for subsequent times? As with a number of the topics in this section, the precise nature of stability depends on the structure present in a specific context. A crucial such structure that often shows up in the theory of stability is the nature of what "close" might mean. As well as this, there are many sorts of refinements of what stability properties might be useful in any given system theoretic problem. What we shall do, therefore, is consider a few examples of stability problems that indicate the kinds of things one often looks for in problems of stability.

2.3.7 Examples (Stability)

1. Let us consider a simple, but not perfectly concrete, example that shows the sort of behaviour that is of interest in the theory of stability. Consider a ball rolling along a hilly terrain prescribed by a function f as in Figure 2.10. We suppose f is bounded from below.

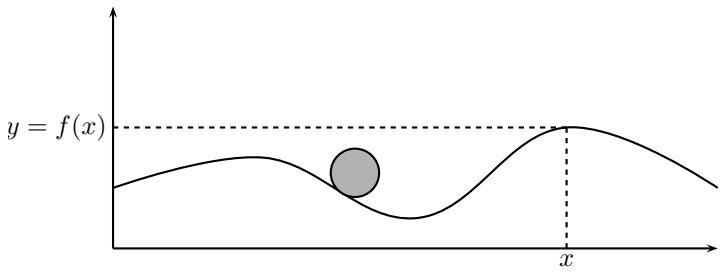


Figure 2.10 Ball rolling over hilly terrain

Suppose first that the ball rolls without friction. Then one can easily imagine that, at positions x where f has a strict local minimum, the position of the ball will be "locally stable," in the sense that, as long as one does not move far away from the minimum point, the ball will be trapped by the terrain, and so remain close to the minimum point. If, on the other hand, one starts near a strict local maximum of f, then the position of the ball will be "unstable," in the sense that the only motion that will not move away from the local maximum is the special motion where the ball is at rest at exactly the local maximum. With no friction present, the ball will never get closer to a strict local minimum of f for longer times.

If one has friction, however, eventually almost all motions of the ball will result in the ball approaching a strict local minimum of f. The only motions that will not have those attributes are those that approach an unstable position as time goes to infinity. The stability possessed by the local minima of f with friction is called "asymptotic stability."

The preceding discussion had to do with stability of equilibria. One can also talk about stability of nonequilibrium motions. Consider, for example, a motion of the ball while it is trapped near a local minimum of f. This motion will be periodic. One can wonder whether it is "stable," in the sense that motions starting nearby remain nearby. It turns out that this is interesting, and depends on the exact shape of the graph of f. Generally speaking, these motions are *not* stable, because nearby motions have different periods, and so will eventually "separate" from the specific trajectory one is considering.

2. The stability discussed in the preceding example of the rolling ball refers to the stability of the states of the system. Another sort of stability is that of input/output stability. Here one is interested in the behaviour of the output for certain sorts of inputs. A typical such property is what is called "bounded-input, bounded-output" stability. As the name suggests, it has to do with whether, given any bounded input, all resulting outputs are also bounded. In this case, one has to be specific about what "bounded" might mean. •

We shall discuss stability at length in Chapters 10 and 11.

2.3.6 Stabilisation

Stabilisation is, as one might expect, rather related to stability. However, here one wishes to make use of the inputs of a system to render a possibly unstable behaviour a stable behaviour. As with stability, making this clear typically requires a careful consideration of system structure. We shall, therefore, examine the problem of stabilisation for a simple, widely used, example.

2.3.8 Example (Stabilisation) We consider a pendulum atop a cart as depicted in Figure 2.11. The cart is subject to a force F as shown, and the objective is to choose F

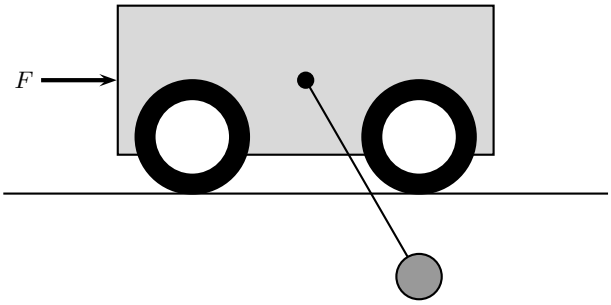


Figure 2.11 Pendulum atop a cart

in such a way that the pendulum is balanced in the upright position, despite this position being naturally unstable. Often what one wants is to design F to be an explicit function of the states of the system (these being the position of the cart, the angle of the pendulum arm, and the velocities of these), and this is known as "feedback stabilisation."

A few important questions present themselves:

1. is it even possible to stabilise the pendulum in the upright position?
2. in designing a feedback stabiliser, how regular a function of the states is the input?
3. from what set of states can the pendulum be stabilised? •

2.3.7 Classification and comparison

There are various sorts of questions that fall under the umbrella of the problems we have in mind in this section. Here are a few of these:

1. given two systems, are they equivalent in some way, namely in some way that they can be mapped one to the other according to some nice class of mappings?
2. for a given system, is there, among all systems equivalent to it, a "nice" one?
3. can the behaviours of a system be regarded as a subset of behaviours of another system?

4. are the behaviours of a system "projections" of the behaviours of a larger system?

These sorts of questions all require some sort of notion of a transformation of systems. The exact sort of transformation one wishes to allow will depend substantially on the structure of the systems under consideration, and the properties of these systems in which one is interested.

Here we shall consider a simple sort of transformation associated with a general input/output system, and this will allow us to provide a little context to the general questions posed above.

2.3.9 Definition (Transformations for general input/output systems) Let $\Sigma_1 = (\mathscr{U}_1, \mathscr{Y}_1, \mathscr{B}_1)$ and $\Sigma_2 = (\mathscr{U}_2, \mathscr{Y}_2, \mathscr{B}_2)$ be general input/output systems. A ***system transformation*** from Σ_1 to Σ_2 is a pair of mappings $\Phi = (\phi_{\text{in}}, \phi_{\text{out}})$ satisfying

$$\phi_{\text{in}} \colon \mathscr{U}_1 \to \mathscr{U}_2, \quad \phi_{\text{out}} \colon \mathscr{Y}_1 \to \mathscr{Y}_2$$

and

$$(\phi_{\text{in}}(\mu), \psi_{\text{out}}(\eta)) \in \mathscr{B}_2, \qquad (\mu, \eta) \in \mathscr{B}_1.$$

We denote by $\Phi \colon \mathscr{B}_1 \to \mathscr{B}_2$ the induced mapping satisfying

$$\Phi(\mu, \eta) = (\phi_{\text{in}}(\mu), \phi_{\text{out}}(\eta)). \qquad \bullet$$

Note that system transformations can be composed in the more or less obvious way. To wit, if $\Sigma_a = (\mathscr{U}_a, \mathscr{Y}_a, \mathscr{B}_a)$, $a \in \{1, 2, 3\}$, are general input/output systems and if $\Phi = (\phi_{\text{in}}, \phi_{\text{out}})$ and $\Psi = (\psi_{\text{in}}, \psi_{\text{out}})$ are system transformations from Σ_1 to Σ_2 and from Σ_2 to Σ_3, respectively, we define a system transformation $\Psi \circ \Phi$ from Σ_1 to Σ_2 by

$$\Psi \circ \Phi = (\psi_{\text{in}} \circ \phi_{\text{in}}, \psi_{\text{out}} \circ \phi_{\text{out}}).$$

Thus a system transformation send behaviours to behaviours. Let us see how this structure can address the questions above. We do this by making some related definitions.

2.3.10 Definition (Types of system transformations) Let $\Sigma_1 = (\mathscr{U}_1, \mathscr{Y}_1, \mathscr{B}_1)$ and $\Sigma_2 = (\mathscr{U}_2, \mathscr{Y}_2, \mathscr{B}_2)$ be general input/output systems and let $\Phi = (\phi_{\text{in}}, \phi_{\text{out}})$ be a system transformation from Σ_1 to Σ_2. The system transformation Φ is

(i) ***surjective*** if $\Phi(\mathscr{B}_1) = \mathscr{B}_2$,

(ii) ***injective*** if $\Phi(\mu, \eta) = \Phi(\mu', \eta')$ implies that $(\mu, \eta) = (\mu', \eta')$,

(iii) ***left invertible*** if there exists a system transformation $\Psi = (\psi_{\text{in}}, \psi_{\text{out}})$ from Σ_2 to Σ_1 such that $\Psi \circ \Phi = (\text{id}_{\mathscr{U}_1}, \text{id}_{\mathscr{Y}_1})$,

(iv) ***right invertible*** if there exists a system transformation $\Psi = (\psi_{\text{in}}, \psi_{\text{out}})$ from Σ_2 to Σ_1 such that $\Phi \circ \Psi = (\text{id}_{\mathscr{U}_2}, \text{id}_{\mathscr{Y}_2})$,

(v) an ***epimorphism*** if, for any general input/output system $\Sigma_3 = (\mathscr{U}_3, \mathscr{Y}_3, \mathscr{B}_3)$ and any system transformations Ψ, Ψ' from Σ_2 to Σ_3, we have

$$\Psi \circ \Phi = \Psi' \circ \Phi \implies \Psi = \Psi',$$

(vi) a ***monomorphism*** if, for any general input/output system $\Sigma_3 = (\mathscr{U}_3, \mathscr{Y}_3, \mathscr{B}_3)$ and any system transformations Ψ, Ψ' from Σ_3 to Σ_1, we have

$$\Phi \circ \Psi = \Phi \circ \Psi' \implies \Psi = \Psi',$$

(vii) an ***isomorphism*** if it is both left and right invertible. •

The reader will recognise some of these ideas from basic mappings between sets; see Definitions I-1.3.6 and I-1.3.8. However, the relationships between these notions of system transformations are not the same as those for sets that are given in Proposition I-1.3.9. To see this, let us give some fairly elementary results and fairly concocted examples that illustrate what is true and not true.

First we deal with the notions of "surjective," "right invertible," and "epimorphism."

2.3.11 Proposition (Relationship between surjective, right invertible, and epimorphism) *Let $\Sigma_1 = (\mathscr{U}_1, \mathscr{Y}_1, \mathscr{B}_1)$ and $\Sigma_2 = (\mathscr{U}_2, \mathscr{Y}_2, \mathscr{B}_2)$ be general input/output systems for which $\mathrm{dom}(\Sigma_a) = \mathscr{U}_a$ and $\mathrm{rng}(\Sigma_a) = \mathscr{Y}_a$, $a \in \{1, 2\}$, and let $\Phi = (\phi_{\mathrm{in}}, \phi_{\mathrm{out}})$ be a system transformation from Σ_1 to Σ_2. Then the following statements hold:*

(i) Φ is an epimorphism if and only if ϕ_{in} and ϕ_{out} are surjective;

(ii) if Φ is surjective, then it is an epimorphism;

(iii) if Φ is right invertible, then it is surjective.

Proof (i) Suppose first that ϕ_{in} and ϕ_{out} are surjective and let Ψ and Ψ' be a system transformation from Σ_2 to a general input/output system Σ_3. Then we have

$$\begin{aligned} &\Psi \circ \Phi = \Psi' \circ \Phi \\ &\implies (\psi_{\mathrm{in}} \circ \phi_{\mathrm{in}}(\mu), \psi_{\mathrm{out}} \circ \phi_{\mathrm{out}}(\eta)) = (\psi'_{\mathrm{in}} \circ \phi_{\mathrm{in}}(\mu), \psi'_{\mathrm{out}} \circ \phi_{\mathrm{out}}(\eta)), \qquad (\mu, \eta) \in \mathscr{B}_1. \end{aligned}$$

Since $\mathrm{dom}(\Sigma_1) = \mathscr{U}_1$, this implies that

$$\psi_{\mathrm{in}} \circ \phi_{\mathrm{in}} = \psi'_{\mathrm{in}} \circ \phi_{\mathrm{in}}.$$

Using Proposition I-1.3.9(ii), we compose on the right by the right inverse of ϕ_{in} and get $\psi_{\mathrm{in}} = \psi'_{\mathrm{in}}$. Now, since $\mathrm{rng}(\Sigma_1) = \mathscr{Y}_1$, we similarly have $\psi_{\mathrm{out}} = \psi'_{\mathrm{out}}$.

Now suppose that ϕ_{in} is not surjective. Thus $\phi_{\mathrm{in}}(\mathscr{U}_1) \subset \mathscr{U}_2$. Take $\mathscr{U}_3 = \{0, 1\}$ and define $\psi_{\mathrm{in}}, \psi'_{\mathrm{in}} : \mathscr{U}_2 \to \mathscr{U}_3$ by

$$\psi_{\mathrm{in}}(\mu) = \begin{cases} 1, & \mu \in \phi_{\mathrm{in}}(\mathscr{U}_1), \\ 0, & \text{otherwise}, \end{cases}, \quad \psi'_{\mathrm{in}}(\mu) = 1,$$

for $\mu \in \mathscr{U}_2$. Also take $\mathscr{Y}_3 = \{0\}$ and define $\psi_{\text{out}}, \psi'_{\text{out}} \colon \mathscr{Y}_2 \to \mathscr{Y}_3$ by

$$\psi_{\text{out}}(\eta) = \psi'_{\text{out}}(\eta), \qquad \eta \in \mathscr{Y}_2.$$

Then $\Psi \circ \Phi = \Psi' \circ \phi$ but $\Psi \neq \Psi'$. A similar argument can be fabricated for the case when ϕ_{out} is not constructed.

(ii) One can easily see that, if Φ is surjective, then ϕ_{in} and ϕ_{out} are surjective. This part of the result then follows from part (i).

(iii) This is simply one half of Proposition I-1.3.9(ii). ■

Let us show that the missing converses from the preceding result do not, in fact, hold.

2.3.12 Examples (Epimorphisms need not be surjective, surjections need not be right invertible)

1. Take $\mathscr{U} = \mathscr{Y}$ and

$$\mathscr{B}_1 = \{(\mu, \mu) \in \mathscr{U} \times \mathscr{Y} \mid \mu \in \mathscr{U}\}, \ \ \mathscr{B}_2 = \mathscr{U} \times \mathscr{Y}.$$

Then define Φ by $\phi_{\text{in}} = \phi_{\text{out}} = \text{id}_{\mathscr{U}}$. One can easily show that Φ is an epimorphism, but is not surjective.

2. We take

$$\mathscr{U} = \{\mu_1, \mu_2, \mu_3\}, \ \ \mathscr{Y} = \{\eta_1, \eta_2\}, \ \ \mathscr{B} = \{(\mu_1, \eta_1), (\mu_2, \eta_2), (\mu_3, \eta_2)\}$$

and

$$\mathscr{U}' = \{\mu'_1, \mu'_2, \mu'_3\}, \ \ \mathscr{Y}' = \{\eta'_1, \eta'_2\}, \ \ \mathscr{B}' = \{(\mu'_1, \eta'_1), (\mu'_2, \eta'_2), (\mu'_3, \eta'_2)\}.$$

Suppose that Φ is a system transformation from Σ to Σ' that satisfies

$$\phi_{\text{in}}(\mu_1) = \phi_{\text{in}}(\mu_2) = \mu'_1, \ \ \phi_{\text{in}}(\mu_3) = \mu'_2$$

and

$$\phi_{\text{out}}(\eta_1) = \eta'_1, \ \ \phi_{\text{out}}(\eta_2) = \eta'_2.$$

Suppose that Φ has a right inverse Ψ. The properties of Φ ensure that

$$\psi_{\text{in}}(\mu'_1) \in \{\mu_1, \mu_2\}, \ \ \psi_{\text{out}}(\eta'_1) = \eta_1, \ \psi_{\text{out}}(\eta'_2) = \eta_2.$$

If $\psi_{\text{in}}(\mu'_1) = \mu_1$ then

$$\Psi(\mu'_1, \eta'_2) = (\mu_1, \eta_2) \notin \mathscr{B}.$$

If $\psi_{\text{in}}(\mu'_1) = \mu_2$ then

$$\Psi(\mu'_1, \eta'_1) = (\mu_2, \eta_1) \notin \mathscr{B},$$

and so Ψ is not a system transformation. •

First we deal with the notions of "injective," "left invertible," and "monomorphism."

2.3.13 Proposition (Relationship between injective, left invertible, and monomorphism) *Let $\Sigma_1 = (\mathscr{U}_1, \mathscr{Y}_1, \mathscr{B}_1)$ and $\Sigma_2 = (\mathscr{U}_2, \mathscr{Y}_2, \mathscr{B}_2)$ be general input/output systems for which $\mathrm{dom}(\Sigma_a) = \mathscr{U}_a$, $a \in \{1, 2\}$, and let $\Phi = (\phi_{\mathrm{in}}, \phi_{\mathrm{out}})$ be a system transformation from Σ_1 to Σ_2. Then the following statements hold:*

(i) *Φ is a monomorphism if and only if Φ is injective;*

(ii) *if ϕ_{in} and ϕ_{out} are injective, then Φ is a monomorphism;*

(iii) *if Φ is left invertible, then ϕ_{in} and ϕ_{out} are injective.*

Proof (i) This is a consequence of Proposition I-1.3.9(i).

(ii) If ϕ_{in} and ϕ_{out} are injective, then Φ is also injective. By part (i), Φ is a monomorphism.

(iii) If Ψ is a left inverse of Φ then we have

$$\psi_{\mathrm{in}} \circ \phi_{\mathrm{in}} = \mathrm{id}_{\mathscr{U}}, \quad \psi_{\mathrm{out}} \circ \phi_{\mathrm{out}} = \mathrm{id}_{\mathscr{Y}},$$

and we conclude that ϕ_{in} and ϕ_{out} are injective by Proposition I-1.3.9(i). ■

Again, the missing converses are not generally true.

2.3.14 Examples (Monomorphisms need not have injective components, morphisms with injective components need not be left invertible)

1. We take

$$\mathscr{U} = \{\mu_1, \mu_2, \mu_3\}, \quad \mathscr{Y} = \{\eta_1, \eta_2, \eta_3\}, \quad \mathscr{B} = \{(\mu_2, \eta_1), (\mu_2, \eta_3), (\mu_3, \eta_2), (\mu_3, \eta_3)\}$$

and

$$\mathscr{U}' = \{\mu_1', \mu_2', \mu_3'\}, \quad \mathscr{Y}' = \{\eta_1', \eta_2'\}, \quad \mathscr{B}' = \{(\mu_1', \eta_1'), (\mu_1', \eta_2'), (\mu_2', \eta_1'), (\mu_3', \eta_2')\}.$$

Suppose that Φ is a system transformation from Σ to Σ' that satisfies

$$\phi_{\mathrm{in}}(\mu_1) = \phi_{\mathrm{in}}(\mu_2) = \mu_1', \quad \phi_{\mathrm{in}}(\mu_3) = \mu_2'$$

and

$$\phi_{\mathrm{out}}(\eta_1) = \phi_{\mathrm{out}}(\eta_2) = \eta_1', \quad \phi_{\mathrm{out}}(\eta_3) = \eta_2'.$$

It is then an easy matter to check that Φ is injective (and so a monomorphism), but that neither ϕ_{in} and ϕ_{out} are injective.

2. We take

$$\mathscr{U} = \{\mu_1, \mu_2\}, \quad \mathscr{Y} = \{\eta_1, \eta_2\}, \quad \mathscr{B} = \{(\mu_1, \eta_1), (\mu_2, \eta_2)\}$$

and

$$\mathscr{U}' = \{\mu_1', \mu_2', \mu_3'\}, \quad \mathscr{Y}' = \{\eta_1', \eta_2'\}, \quad \mathscr{B}' = \{(\mu_1', \eta_1'), (\mu_2', \eta_2'), (\mu_3', \eta_1'), (\mu_3', \eta_2')\}.$$

Suppose that Φ is a system transformation from Σ to Σ' that satisfies

$$\phi_{\mathrm{in}}(\mu_1) = \mu_1', \quad \phi_{\mathrm{in}}(\mu_2) = \mu_2'$$

and

$$\phi_{\text{out}}(\eta_1) = \eta_1', \quad \phi_{\text{out}}(\eta_2) = \eta_2'.$$

We see that ϕ_{in} and ϕ_{out} are injective. Suppose that Φ has a left inverse Ψ. The properties of Φ ensure that

$$\psi_{\text{in}}(\mu_3') \in \{\mu_1, \mu_2\}.$$

If $\psi_{\text{in}}(\mu_3') = \mu_1$ then

$$\Psi(\mu_3', \eta_2') = (\mu_1, \eta_2) \notin \mathscr{B}.$$

If $\psi_{\text{in}}(\mu_3') = \mu_2$ then

$$\Psi(\mu_3', \eta_1') = (\mu_2, \eta_1) \notin \mathscr{B},$$

and so Ψ is not a system transformation. Thus Φ is not left invertible. •

This version: 2022/03/07

Chapter 3

Differential and difference equations: General theory

Thus far we have considered a host of examples of systems from various areas of application (Chapter 1) and considered a general—too general—setting for system theory (Chapter 2). Having staked out the extreme positions of system theory from the very applied to the very abstract, let us now fill in some part of the middle by considering in detail a specific class of systems. In this volume, the classes of systems we work with are almost exclusively described by differential and difference equations. In this chapter we consider a fairly general setting for equations of this sort in order that we can understand in context the particular equations we study in detail subsequently. In Chapters 4 and 5 we shall study the particular classes and there we shall devote substantial effort towards solving differential and difference equations. However, it is generally impossible to solve a differential or difference equation chosen at random from the bag of differential or difference equations. For this reason, it is worth understanding what a differential or difference equation *is*, rather than how to solve one. Thus in this chapter we dedicate ourselves to the questions

1. What is a differential equation?
2. What is a difference equation?
3. What is a solution of a differential equation?
4. What is a solution of a difference equation?
5. Are there useful classes of differential equations?
6. Are there useful classes of difference equations?

The study of these questions is a little abstract and without context for someone new to the subjects of differential and difference equations. However, as one becomes more expert in these subjects, being able to be clear about answers to these questions is important, and ultimately less confusing than not addressing them.

Do I need to read this chapter? We present a point of view of differential and difference equations that is a little different than the usual view of differential and difference equations, and we use in chapters below the nonstandard language we

develop here. Therefore, just from the point of view of the notation we shall use, a reading of this chapter is essential. •

Contents

Section 3.1

Classification of differential equations

In Section 1.1 we saw many examples of differential equations, and there were many different types of differential equations represented in these examples. In this section we provide some procedures for separating differential equations into classes that are special. Such a process cannot be exhaustive, especially at the level which we are able to treat the subject. Nonetheless, the classifications we provide here give important first steps in any classification procedure, and allow us to clearly distinguish the very few differential equations that we can treat in detail by pointing these the special attributes of these equations.

Do I need to read this section? The language we present in this section will be often used below.

3.1.1 Variables in differential equations

In all of the examples in Section 1.1 we pointed out the independent and dependent variables. In this section we chat about this in a general sort of way.

The independent variables for a differential equation typically reside in an open subset $D \subseteq \mathbb{R}^n$ for some $n \in \mathbb{Z}_{>0}$. These are the variables upon which our objects of interest depend. In the case of $n = 1$, this variable is often thought of as time, although it is also common for this single variable to be a spatial variable.

The dependent variables in a differential equation represent the quantities whose behaviour, as functions of the independent variable, one wishes to understand. We typically regard dependent variables as being in an open subset $U \subseteq \mathbb{R}^m$ for some $m \in \mathbb{Z}_{>0}$. Very often, when one wishes to understand the behaviour of a solution of a differential equation, one plots graphs of the dependent variables as functions of the independent variables. For large numbers of variables, such graphical representations become difficult, and one is forced to think abstractly to understand the behaviour of solutions.

In cases where the number of independent variables is 1, as we mention above this variable typically represents time or space. We shall assume, in general situations, that this variable represents time which we denote by "t." In such cases we represent derivatives of the dependent variables with a dot, e.g., $\dot{x}$ for the first derivative, $\ddot{x}$ for the second derivative, and so on. Thus

$$\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t}, \quad \ddot{x} = \frac{\mathrm{d}^2x}{\mathrm{d}t^2}.$$

In the case of a single independent variable which is regarded as a spatial variable, we denote this spatial variable by "x." Derivatives of this spatial variable we

denote by a prime, e.g., y' is the first derivative and y'' is the second derivative. Thus

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x}, \quad y'' = \frac{\mathrm{d}^2y}{\mathrm{d}x^2}.$$

Higher-order derivatives will be denoted by $x^{(k)} = \frac{\mathrm{d}^k x}{\mathrm{d}t^k}$ (for time derivatives) or $y^{(k)} = \frac{\mathrm{d}^k y}{\mathrm{d}x^k}$ (for spatial derivatives) in the usual way.

When there is more than one independent variable, we will not use this notation, and indeed it is faulty to do so; stick to the partial derivative notation in this case. Some commonly encountered notation in this case is to use subscripts to connote the variable with which differentiation is occurring. For example, one sees

$$\frac{\partial^2 u}{\partial x^2} = u_{xx}, \quad \frac{\partial u}{\partial t} = u_t, \quad \frac{\partial^2 u}{\partial x \partial t} = u_{xt}.$$

Note that this notation is *never* to be used when dealing specifically with a single independent variable.

Let us adapt this subscript notation to give a general notation for derivatives. Let $D \subseteq \mathbb{R}^n$ be open and denote coordinates for D by $(x_1, \dots, x_n)$. As we have seen, the kth-order partial derivatives for a function $\boldsymbol{u}\colon D \to U$ are those partial derivatives

$$\frac{\partial u_a}{\partial x_{j_1} \cdots \partial x_{j_k}}, \qquad a \in \{1, \dots, m\},\ j_1, \dots, j_k \in \{1, \dots, n\}.$$

We can use this to motivate notation for coordinates for $\mathrm{L}^k_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m)$. Indeed, we shall use

$$u^a_{j_1 \cdots j_k}, \qquad a \in \{1, \dots, m\},\ j_1, \dots, j_k \in \{1, \dots, n\}, \tag{3.1}$$

for coordinates. Thus a k-multilinear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ can be denoted by

$$(v_1, \dots, v_k) \mapsto \left(\sum_{j_1, \dots, j_k = 1}^n u^1_{j_1 \cdots j_k} v_{1, j_1} \cdots v_{k, j_k}, \dots, \sum_{j_1, \dots, j_k = 1}^n u^m_{j_1 \cdots j_k} v_{1, j_1} \cdots v_{k, j_k} \right).$$

Of course, this is all just a notational encoding of the derivative as defined in Definitions II-1.4.2 and II-1.4.4, and using the symmetry of the derivative proved in Theorem II-1.4.5. We shall also be interested in the space that contains derivatives up to order k, and this is

$$\mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) = \oplus_{j=1}^k \mathrm{L}^j_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m).$$

3.1.2 Differential equations and solutions

In this section we give a *very* general definition of what is meant by a differential equation. While the definition we give is well suited to the objectives of classification in this section, we will not work deeply with this definition outside this section.

First let us give this definition.

3.1.1 Definition (Differential equation) A ***differential equation*** consists of a mapping

$$\boldsymbol{F}\colon D \times U \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l,$$

where $k, l, m, n \in \mathbb{Z}_{>0}$, and $D \subseteq \mathbb{R}^n$ and $U \subseteq \mathbb{R}^m$, with D open. We also have the following terminology:

(i) n is the number of ***independent variables***;
(ii) m is the number of ***unknowns*** or ***states***;
(iii) k is the ***order***;
(iv) l is the number of ***equations***;
(v) $D \subseteq \mathbb{R}^n$ is the ***domain*** for the differential equation;
(vi) $U \subseteq \mathbb{R}^m$ is the ***state space*** for the differential equation. •

To get an understanding of why the preceding definition might encode the notion of a differential equation, let us define what we mean by a solution to a differential equation.

3.1.2 Definition (Solution to a differential equation) Let

$$\boldsymbol{F}\colon D \times U \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l,$$

be a differential equation. A ***solution*** to the differential equation is a function $\boldsymbol{u}\colon D' \to U$ of class C^k defined on an open subset $D' \subseteq D$ such that

$$F(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{D}\boldsymbol{u}(\boldsymbol{x}), \dots, \boldsymbol{D}^k\boldsymbol{u}(\boldsymbol{x})) = \boldsymbol{0}, \qquad \boldsymbol{x} \in D'. \quad \bullet$$

This definitions seem quite abstract at this point, so let us illustrate how this works in all of our examples from Section 1.1. In doing this, we shall use the notation (3.1) to denote coordinates for derivatives. Some of the examples are a little tedious to write out in full detail, so we do not do so. However, we encourage the interested reader to undertake to carry out the procedure we describe for any of their favourite equations that we do not work out. For example, Star Wars nerds will probably *need* to work out how to write Einstein's field equations as a formal differential equation in the sense of Definition 3.1.1.

3.1.3 Examples (Differential equations and solutions)

1. For the mass-spring-damper equation we derived in (1.1), we have $n = 1$, $m = 1$, $l = 1$, and $k = 2$. We take $D = \mathbb{R}$ and $U = \mathbb{R}$ for concreteness. Thus we consider all possible times and vertical displacements in the description of the system; this is something that one generally chooses with the specific instantiation of the problem. We use the coordinate t for independent variable time, y for the unknown vertical displacement. Then we have coordinates y_t and y_{tt} for derivatives. We then have

$$F\colon \mathbb{R} \times \mathbb{R} \times \mathrm{L}^{\leq 2}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$

defined by

$$F(t, y, y_t, y_{tt}) = my_{tt} + dy_t + ky + ma_g.$$

A solution to this equation is then a mapping $y\colon \mathbb{T} \to \mathbb{R}$ defined on some interval $\mathbb{T}' \subseteq \mathbb{R}$ that satisfies

$$F\left(t, y(t), \frac{\mathrm{d}y}{\mathrm{d}t}(t), \frac{\mathrm{d}^2y}{\mathrm{d}t^2}(t)\right) = m\frac{\mathrm{d}^2y}{\mathrm{d}t^2}(t) + d\frac{\mathrm{d}y}{\mathrm{d}t}(t) + ky(t) + ma_g = 0.$$

This, of course, is exactly the equation (1.1).

2. For the coupled mass-spring-damper equation of (1.2), we have $n = 1$, $m = 2$, $k = 2$, and $l = 2$. We again take $D = \mathbb{R}$ and $U = \mathbb{R}$ for concreteness, and we use t as the independent variable time, x_1 and x_2 as the states, the displacements of the masses, and we denote the coordinates for the derivatives by

$$x_{1,t},\ x_{2,t},\ x_{1,tt},\ x_{2,tt}.$$

The map

$$\boldsymbol{F}\colon \mathbb{R} \times \mathbb{R} \times \mathrm{L}^{\leq 2}_{\mathrm{sym}}(\mathbb{R}^2; \mathbb{R}^2) \to \mathbb{R}^2$$

for this differential equation is then

$$\boldsymbol{F}(t, x_1, x_2, x_{1,t}, x_{2,t}, x_{1,tt}, x_{2,tt}) = (mx_{1,tt} + 2kx_1 - kx_2, mx_{2,tt} - kx_1 + 2kx_2),$$

and a solution $\boldsymbol{x}\colon \mathbb{T} \to \mathbb{R}^2$ satisfies the equation

$$\begin{aligned}\boldsymbol{F}&\left(t, x_1(t), x_2(t), \frac{\mathrm{d}x_1}{\mathrm{d}t}(t), \frac{\mathrm{d}x_2}{\mathrm{d}t}(t), \frac{\mathrm{d}^2x_1}{\mathrm{d}t^2}(t), \frac{\mathrm{d}^2x_2}{\mathrm{d}t^2}(t)\right)\\ &= \left(m\frac{\mathrm{d}^2x_1}{\mathrm{d}t^2}(t) + 2kx_1(t) - kx_2(t), m\frac{\mathrm{d}^2x_2}{\mathrm{d}t^2}(t) - kx_1(t) + 2kx_2(t)\right) = (0, 0).\end{aligned}$$

These equations are, of course, simply the equations (1.2) written in a different form. We can unify the two forms of the equations a little more by writing

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}_t, \boldsymbol{x}_{tt}) = \boldsymbol{M}\boldsymbol{x}_{tt} + \boldsymbol{K}\boldsymbol{x},$$

where $\boldsymbol{x}_t = (x_{1,t}, x_{2,t})$ and $\boldsymbol{x}_{tt} = (x_{1,tt}, x_{2,tt})$.

3. For the simple pendulum equation of (1.3), we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.1.
4. For Bessel's equation (1.5), we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.2.
5. For the equation (1.6) governing the current in a series RLC circuit, we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.3.

6. For the tank equations of (1.7), we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.4.
7. For the logistical model (1.8) of a population, we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.5.
8. For the Lotka–Volterra predator prey model of (1.9), we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.6.
9. For the Rapoport production and exchange model of (1.10), we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.7.
10. The Euler–Lagrange equations of (1.11) have $n = 1$, $m = 1$, $k = 2$, and $l = 1$. We take $D = [x_1, x_2]$ (let's overlook, for the moment, the fact that this D is not open) and $U = \mathbb{R}$, and use x as the independent variable, y as the unknown, and y_x and y_{xx} as variables for the required derivatives. The Lagrangian L is then a function of x, y, and y_x. The differential equation is then prescribed by the map

$$F\colon [x_1, x_2] \times \mathbb{R} \times \mathrm{L}^{\leq 2}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$

given by

$$F(x, y, y_x, y_{xx}) = \frac{\partial^2 L}{\partial y_x^2} y_{xx} + \frac{\partial^2 L}{\partial y_x \partial y} y_x - \frac{\partial L}{\partial y}.$$

A solution to these equations is then a function $y\colon [x_1, x_2] \to \mathbb{R}$ satisfying

$$F\left(x, y(x), \frac{\mathrm{d}y}{\mathrm{d}x}(x), \frac{\mathrm{d}^2 y}{\mathrm{d}x^2}(x)\right) = \frac{\partial^2 L}{\partial y_x^2} \frac{\mathrm{d}^2 y}{\mathrm{d}x^2}(x) + \frac{\partial^2 L}{\partial y_x \partial y} \frac{\mathrm{d}y}{\mathrm{d}x}(x) - \frac{\partial L}{\partial y} = 0,$$

which is exactly the Euler–Lagrange equation.

11. In Maxwell's equations (1.12), we have $n = 4$, $m = 10$, $k = 1$, and $l = 1+1+3+3 = 8$. To write the function $\boldsymbol{F}$ defining Maxwell's equations is tedious because of the largish number of variables. For example, if we include all required derivatives, the number of arguments for $\boldsymbol{F}$ in this case is $4 + 10 + 40 = 54$.
12. For the Navier–Stokes equations (1.14), along with the equations of continuity (1.13), we have $n = 4$, $m = 5$, $k = 1$, and $l = 3 + 1 = 4$. In this case, the number of variables is manageable, but the equations themselves are quite lengthy and complicated. Thus we do not go through the details of writing down $\boldsymbol{F}$ in this case.
13. For the heat equation (1.17), we have $n = 2$, $m = 1$, $k = 2$, and $l = 1$. For the domain D, we will suppose that we are working with a rod of length ℓ and that we consider positive times. Thus we take $D = [0, \ell] \times \mathbb{R}_{\geq 0}$ (sweeping under the rug the fact that D is not open). We also take $U = \mathbb{R}$. We denote the independent time/space variables as (x, t), the unknown temperature as u, and the required derivatives are

$$u_x,\ u_t,\ u_{xx},\ u_{xt},\ u_{tt},$$

keeping in mind that $u_{tx} = u_{xt}$ by symmetry of derivatives. The map

$$F\colon [0,\ell] \times \mathbb{R}_{\geq 0} \times \mathbb{R} \times \mathrm{L}^{\leq 2}_{\mathrm{sym}}(\mathbb{R}^2;\mathbb{R}) \to \mathbb{R}$$

is given by

$$F(x,t,u,u_x,u_t,u_{xx},u_{xt},u_{xx}) = u_t - ku_{xx}.$$

A solution is then a function $u\colon [0,\ell] \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfying

$$F\left(x,t,u(x,t),\frac{\partial u}{\partial x}(x,t),\frac{\partial u}{\partial t}(x,t),\frac{\partial^2 u}{\partial x^2}(x,t),\frac{\partial^2 u}{\partial x \partial t}(x,t),\frac{\partial^2 u}{\partial t^2}(x,t)\right) = \frac{\partial u}{\partial t}(x,t) - k\frac{\partial^2 u}{\partial x^2}(x,t) = 0,$$

which is just the heat equation, of course.

14. For the wave equation (1.18), we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.8.
15. For the potential equation (1.19), we leave the working out of this as a differential equation and the conditions for a solution as Exercise 3.1.9.
16. For the Einstein field equations (1.20), we have $n = 4$, $m = 10$, $k = 2$ (can you work out why?), and $l = 10$. These equations are extremely complicated to write as a differential equation as per Definition 3.1.1, and so we do not do this here. For example, the number of arguments of $\boldsymbol{F}$ in this case would be $4 + 10 + 40 + 100 = 154$!
17. Finally, we consider the Schrödinger equation (1.21). For this equation we have $n = 4$, $m = 2$, $k = 2$, and $l = 2$. Here, for simplicity, we take $D = \mathbb{R}^4$ and $U = \mathbb{C} \simeq \mathbb{R}^2$. We use coordinates (x_1, x_2, x_3, t) the independent variables, (ψ_1, ψ_2) for the unknown real and imaginary parts of the wave function, and the required derivatives are

$$\begin{aligned}&\psi_{1,x_1},\ \psi_{1,x_2},\ \psi_{1,x_3},\ \psi_{1,t},\ \psi_{2,x_1},\ \psi_{2,x_2},\ \psi_{2,x_3},\ \psi_{2,t},\\ &\quad \psi_{1,x_1x_1},\ \psi_{1,x_1x_2},\ \psi_{1,x_1x_3},\ \psi_{1,x_1t},\ \psi_{1,x_2x_2},\ \psi_{1,x_2x_3},\ \psi_{1,x_2t},\ \psi_{1,x_3x_3},\psi_{1,x_3t},\ \psi_{1,tt},\\ &\quad\quad \psi_{2,x_1x_1},\ \psi_{2,x_1x_2},\ \psi_{2,x_1x_3},\ \psi_{2,x_1t},\ \psi_{2,x_2x_2},\ \psi_{2,x_2x_3},\ \psi_{2,x_2t},\ \psi_{2,x_3x_3},\psi_{2,x_3t},\ \psi_{2,tt}.\end{aligned}$$

The map

$$\boldsymbol{F}\colon \mathbb{R}^4 \times \mathbb{R}^2 \times \mathrm{L}^{\leq 2}_{\mathrm{sym}}(\mathbb{R}^4;\mathbb{R}^2) \to \mathbb{R}$$

defining the Schrödinger equation is

$$\begin{aligned}&\boldsymbol{F}(x_1,x_2,x_3,t,\psi_1,\psi_2,\psi_{1,x_1},\ \psi_{1,x_2},\ \psi_{1,x_3},\ \psi_{1,t},\ \psi_{2,x_1},\ \psi_{2,x_2},\ \psi_{2,x_3},\ \psi_{2,t},\\ &\quad \psi_{1,x_1x_1},\ \psi_{1,x_1x_2},\ \psi_{1,x_1x_3},\ \psi_{1,x_1t},\ \psi_{1,x_2x_2},\ \psi_{1,x_2x_3},\ \psi_{1,x_2t},\ \psi_{1,x_3x_3},\psi_{1,x_3t},\ \psi_{1,tt},\\ &\quad \psi_{2,x_1x_1},\ \psi_{2,x_1x_2},\ \psi_{2,x_1x_3},\ \psi_{2,x_1t},\ \psi_{2,x_2x_2},\ \psi_{2,x_2x_3},\ \psi_{2,x_2t},\ \psi_{2,x_3x_3},\psi_{2,x_3t},\ \psi_{2,tt})\\ &= (\hbar\psi_{2,t}+\tfrac{\hbar^2}{2\mu}(\psi_{1,x_1x_1}+\psi_{1,x_2x_2}+\psi_{1,x_3,x_3})-V\psi_1, -\hbar\psi_{1,t}+\tfrac{\hbar^2}{2\mu}(\psi_{2,x_1x_1}+\psi_{2,x_2x_2}+\psi_{2,x_3,x_3})-V\psi_2).\end{aligned}$$

A solution is then a map $\boldsymbol{\psi}\colon D' \to \mathbb{R}^2$ defined on some open set $D' \subseteq \mathbb{R}^4$ that satisfies the equation (with the tedious arguments abbreviated)

$$\boldsymbol{F}\left(\boldsymbol{x}, t, \boldsymbol{\psi}(\boldsymbol{x}), \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{x}}, \frac{\partial^2 \boldsymbol{\psi}}{\partial \boldsymbol{x}^2}\right) = \left(\hbar \frac{\partial \psi_2}{\partial t} + \tfrac{\hbar^2}{2\mu}\left(\frac{\partial^2 \psi_1}{\partial x_1^2} + \frac{\partial^2 \psi_1}{\partial x_2^2} + \frac{\partial^2 \psi_1}{\partial x_3^2}\right) - V\psi_1, \right.$$
$$\left. -\hbar \frac{\partial \psi_1}{\partial t} + \tfrac{\hbar^2}{2\mu}\left(\frac{\partial^2 \psi_2}{\partial x_1^2} + \frac{\partial^2 \psi_2}{\partial x_2^2} + \frac{\partial^2 \psi_2}{\partial x_3^2}\right) - V\psi_2\right).$$

One can check that, indeed, these are the Schrödinger equations, broken into their real and imaginary parts. •

Having now introduced what we mean by a solution to a differential equation, let us point out that, in practice, one often has to be more careful about what a solution is.

3.1.4 Remark (Relaxing the properties of a solution) In our definition of a solution to a differential equation, we asked that the solution have the same number of continuous derivatives as appear in the differential equation. This seems like a natural thing to do. However, there are many instances where this idea of a solution is too strong. We shall not pursue this in any generality here; it will come up in specific instances and we will be sure to point this out when it happens. •

If this is a student's first encounter with the subject of differential equations, the preceding way of doing things may seem excessively complicated. Indeed, we went through a lot of trouble to just write down equations that were comparatively easy to write down in our modelling exercises of Section 1.1. The benefits of our work will now be seen. Since we know what a differential equation *is* (it is the map $\boldsymbol{F}$), we can speak intelligently about its attributes. And it is this that we now do.

3.1.3 Ordinary differential equations

We begin with a consideration of differential equations with a single independent variable, which we will think of as representing time. The states or unknowns we will represent by $\boldsymbol{x} \in X \subseteq \mathbb{R}^m$, hereby changing the notation for state spaces from U to X in the case of ordinary differential equations. Because of the simplicity of the single independent variable, we can make a more concrete representation for the derivatives. Specifically, we will denote the coordinates for the derivatives up to order k by

$$(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) \in \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^n).$$

Thus $\boldsymbol{x}^{(j)}$ represents the jth derivative with respect to time (this is not uncommon notation, the only difference here is we are thinking of this as being a coordinate rather than an actual derivative).

3.1.5 Remark (Simplification of derivatives with one independent variable) Now, we make a few observations to make things even more concrete:

1. because the domain is 1-dimensional, every multilinear map from $\mathbb{R}$ to $\mathbb{R}^m$ is symmetric;
2. we have a natural isomorphism of the vector spaces $\mathrm{L}^k(\mathbb{R};\mathbb{R}^m)$ with $\mathbb{R}^m$ by assigning to the k-multilinear map $\boldsymbol{T} \in \mathrm{L}^k(\mathbb{R};\mathbb{R}^m)$ the element $\boldsymbol{v}_{\boldsymbol{T}} \in \mathbb{R}^m$ given by

$$\boldsymbol{v}_{\boldsymbol{T}} = \boldsymbol{T}(1, \dots, 1).$$

The punchline of the preceding is that we can think of

$$\mathrm{L}^k_{\mathrm{sym}}(\mathbb{R};\mathbb{R}^m) \simeq \mathbb{R}^m \implies \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}^m) \simeq \underbrace{\mathbb{R}^m \oplus \cdots \oplus \mathbb{R}^m}_{k+1 \text{ times}}.$$

While we will continue to write things using the notation on the left of these isomorphisms, we shall, when convenient, use the isomorphisms to simplify things. •

3.1.3.1 General ordinary differential equations

With the preceding notation, we have the following definition.

3.1.6 Definition (Ordinary differential equation) An ***ordinary differential equation*** is a differential equation $\boldsymbol{F}$ subject to the following conditions:

(i) there is one independent variable, i.e., $n = 1$;
(ii) the independent variable takes values in an interval $\mathbb{T} \subseteq \mathbb{R}$ called the ***time-domain***;
(iii) the ***state space*** is an open subset $X \subseteq \mathbb{R}^m$;
(iv) there are the same number of equations as states, i.e., $l = m$;
(v) if the order of the differential equation is k, for each

$$(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}) \in \mathbb{T} \times X \times \mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}^m),$$

the equation

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}, \boldsymbol{x}^{(k)}) = \boldsymbol{0}$$

can be uniquely solved to give

$$\boldsymbol{x}^{(k)} = \widehat{\boldsymbol{F}}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}).$$

We call $\widehat{\boldsymbol{F}} \colon \mathbb{T} \times X \times \mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}^m) \to \mathbb{R}^m$ the ***right-hand side*** for the ordinary differential equation. •

We can give an alternative characterisation for solutions for ordinary differential equations.

3.1.7 Proposition (Solutions to ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with time-domain* $\mathbb{T}$*, state space* $X \subseteq \mathbb{R}^m$*, and right-hand side* $\widehat{\mathbf{F}}$*. Then the following statements are equivalent for a* C^k *map* $\xi\colon \mathbb{T}' \to X$ *defined on a subinterval* $\mathbb{T}' \subseteq \mathbb{T}$*:*

(i) ξ *is a solution for* $\mathbf{F}$*;*

(ii) ξ *satisfies the equation*

$$\frac{d^k\xi}{dt^k}(t) = \widehat{\mathbf{F}}\left(t, \xi(t), \frac{d\xi}{dt}(t), \dots, \frac{d^{k-1}\xi}{dt^k}(t)\right).$$

Proof First suppose that ξ is a solution for $\boldsymbol{F}$. Then

$$\boldsymbol{F}\left(t, \xi(t), \frac{d\xi}{dt}(t), \dots, \frac{d^k\xi}{dt^k}(t)\right) = \mathbf{0}.$$

The property (v) of Definition 3.1.6, we immediately have

$$\frac{d^k\xi}{dt^k}(t) = \widehat{\boldsymbol{F}}\left(t, \xi(t), \frac{d\xi}{dt}(t), \dots, \frac{d^{k-1}\xi}{dt^k}(t)\right).$$

Next suppose that ξ satisfies the preceding equation. Fix $t \in \mathbb{T}$ and consider the equation

$$\boldsymbol{F}\left(t, \xi(t), \frac{d\xi}{dt}(t), \dots, \frac{d^{k-1}\xi}{dt^k}(t), \boldsymbol{x}^{(k)}\right) = \mathbf{0}.$$

By property (v) of Definition 3.1.6, there exists a unique $\boldsymbol{x}^{(k)} \in \mathbb{R}^m$ that solves this equation and, moreover,

$$\boldsymbol{x}^{(k)} = \widehat{\boldsymbol{F}}\left(t, \xi(t), \frac{d\xi}{dt}(t), \dots, \frac{d^{k-1}\xi}{dt^k}(t)\right).$$

This means, however, that

$$\boldsymbol{x}^{(k)} = \frac{d^k\xi}{dt^k}(t).$$

Thus

$$\boldsymbol{F}\left(t, \xi(t), \frac{d\xi}{dt}(t), \dots, \frac{d^{k-1}\xi}{dt^k}(t), \frac{d^k\xi}{dt^k}(t)\right) = \mathbf{0},$$

i.e., ξ is a solution for $\boldsymbol{F}$. ■

This last condition in Definition 3.1.6 is one that very often arises naturally when looking at specific differential equations. To see how this arises, let us consider the examples of Section 1.1 with one independent variable, and see how their right-hand sides are naturally defined.

3.1.8 Examples (Ordinary differential equations)

1. For the mass-spring-damper equation we derived in (1.1), we can use our ordinary differential equation specific notation to write

$$F(t, y, y^{(1)}, y^{(2)}) = my^{(2)} + dy^{(1)} + ky + ma_g.$$

Note that this is indeed an ordinary differential equation since (1) $n = 1$, (2) $l = m = 1$, and (3) we can solve the equation

$$F(t, y, y^{(1)}, y^{(2)}) = 0$$

for $y^{(2)}$ as

$$y^{(2)} = \tfrac{1}{m}(-dy^{(1)} - ky - ma_g).$$

Thus the right-hand side is

$$\widehat{F}(t, y, y^{(1)}) = \tfrac{1}{m}(-dy^{(1)} - ky - ma_g).$$

As per Proposition 3.1.7, a solution to the differential equation then satisfies

$$\ddot{y}(t) = \tfrac{1}{m}(-d\dot{y}(t) - ky(t) - ma_g),$$

as expected.

2. For the coupled mass-spring-damper equation of (1.2), the differential equation can be conveniently expressed as

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \boldsymbol{M}\boldsymbol{x}^{(2)} + \boldsymbol{K}\boldsymbol{x}.$$

This is an ordinary differential equation since (1) $n = 1$, (2) $l = m = 2$, and (3) we can solve the equation

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \boldsymbol{0}$$

for $\boldsymbol{x}^{(2)}$ as

$$\boldsymbol{x}^{(2)} = -\boldsymbol{M}^{-1}\boldsymbol{K}\boldsymbol{x}.$$

Thus the right-hand side of this ordinary differential equation is

$$\widehat{\boldsymbol{F}}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}) = -\boldsymbol{M}^{-1}\boldsymbol{K}\boldsymbol{x}.$$

As per Proposition 3.1.7, a solution satisfies

$$\ddot{\boldsymbol{x}}(t) = -\boldsymbol{M}^{-1}\boldsymbol{K}\boldsymbol{x}(t),$$

which is simply our original equation, rewritten.

3. For the simple pendulum equation of (1.3), we leave the working out of the right-hand side and corresponding conditions for solutions as Exercise 3.1.10.

4. For Bessel's equation (1.5), we leave the working out of the right-hand side and corresponding conditions for solutions as Exercise 3.1.11.
5. For the current in a series RLC circuit of (1.6), we leave the working out of the right-hand side and corresponding conditions for solutions as Exercise 3.1.12.
6. For the tank flow model of (1.7), we leave the working out of the right-hand side and corresponding conditions for solutions as Exercise 3.1.13.
7. For the logistical model population of (1.8), we leave the working out of the right-hand side and corresponding conditions for solutions as Exercise 3.1.14.
8. For the Lotka–Volterra predator prey model of (1.9), we leave the working out of the right-hand side and corresponding conditions for solutions as Exercise 3.1.15.
9. For the Rapoport production and exchange model of (1.10), we leave the working out of the right-hand side and corresponding conditions for solutions as Exercise 3.1.16.
10. Our final example, that of the Euler–Lagrange equations, shows that one must sometimes take care with what is and is not an ordinary differential equation. We let x denote the single independent variable, y the unknown, and we follow our ordinary differential equation notation and denote derivatives by $y^{(1)}$ and $y^{(2)}$. The Lagrangian is then a function of x, y, and $y^{(1)}$, and the Euler–Lagrange equations are differential equations prescribed by

$$F(x, y, y^{(1)}, y^{(2)}) = \frac{\partial^2 L}{\partial y^{(1)} \partial y^{(1)}} y^{(2)} + \frac{\partial^2 L}{\partial y^{(1)} \partial y} y^{(1)} - \frac{\partial L}{\partial y}.$$

This differential equation is an ordinary differential equation if and only if

$$\frac{\partial^2 L}{\partial y^{(1)} \partial y^{(1)}}$$

is non-zero for every $(x, y, y^{(1)})$. This is true, for example, if

$$L(x, y, y^{(1)}) = (y^{(1)})^2.$$

It is not true, for example, when

$$L(x, y, y^{(1)}) = f(x, y)$$

for any function of (x, y) or when

$$L(x, y, y^{(1)}) = y^{(1)}.$$

Thus we cannot say that the Euler–Lagrange equations are ordinary differential equations, in general, but must examine particular Lagrangians. •

Note that an ordinary differential equation $\boldsymbol{F}$ determines uniquely its right-hand side $\widehat{\boldsymbol{F}}$, but that it is possible that two different ordinary differential equations can give rise to the same right-hand side. To resolve this ambiguity, we make the following definition.

3.1.9 Definition (Normalised ordinary differential equation) An ordinary differential equation

$$\boldsymbol{F}\colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

is ***normalised*** if

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) = \boldsymbol{x}^{(k)} - \widehat{\boldsymbol{F}}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)})$$

for all

$$(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) \in \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m). \qquad \bullet$$

If $\boldsymbol{F}$ is an ordinary differential equation that is *not* normalised, we can alway replace it with an ordinary differential equation $\boldsymbol{F}^*$ that *is* normalised, according to the formula

$$\boldsymbol{F}^*(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) = \boldsymbol{x}^{(k)} - \widehat{\boldsymbol{F}}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}).$$

Moreover, by Proposition 3.1.7, $t \mapsto \xi(t)$ is a solution for $\boldsymbol{F}$ if and only if it is a solution for $\boldsymbol{F}^*$. In short, we can without loss of generality assume that an ordinary differential equation is normalised. That being said, we will only rarely make this assumption.

Now that we have defined what we mean, in general terms, by an ordinary differential equation, let us examine certain special kinds of such equations.

We begin with a general and common sort of simplification that can be made with the general definition.

3.1.10 Definition (Autonomous ordinary differential equation) An ordinary differential equation

$$\boldsymbol{F}\colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

is ***autonomous*** if there exists $\boldsymbol{F}_0\colon X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$ so that

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) = \boldsymbol{F}_0(\boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)})$$

for every $(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) \in \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m)$. An ordinary differential equation that is not autonomous is ***nonautonomous***. $\bullet$

Simply put, an autonomous ordinary differential equation is independent of time.

One can equivalently characterise the notion of autonomous in terms of right-hand sides.

3.1.11 Proposition (Right-hand sides of autonomous ordinary differential equations) *If an ordinary differential equation*

$$\mathbf{F}\colon \mathbb{T} \times \mathrm{X} \times \mathrm{L}^{\leq \mathrm{k}}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

with right-hand side

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{X} \times \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

is autonomous, then there exists

$$\widehat{\mathbf{F}}_0\colon \mathrm{X} \times \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

such that

$$\widehat{\mathbf{F}}(\mathrm{t}, \mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}) = \widehat{\mathbf{F}}_0(\mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}).$$

for every $(\mathrm{t}, \mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}) \in \mathbb{T} \times \mathrm{X} \times \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}})$.

Proof Suppose that $\boldsymbol{F}$ is autonomous. Let

$$(\boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}) \in X \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{m})$$

and let $t_1, t_2 \in \mathbb{T}$. Then there exists a unique $\boldsymbol{x}_1^{(k)}, \boldsymbol{x}_2^{(k)} \in \mathrm{L}^{k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{m})$ such that

$$\boldsymbol{F}(t_a, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}, \boldsymbol{x}_a^{(k)}) = \boldsymbol{0}.$$

Moreover, since $\boldsymbol{F}$ is autonomous, we conclude that $\boldsymbol{x}_1^{(k)} = \boldsymbol{x}_2^{(k)}$. We also have

$$\boldsymbol{x}_a^{(k)} = \widehat{\boldsymbol{F}}(t_a, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}), \qquad a \in \{1, 2\},$$

and so

$$\widehat{\boldsymbol{F}}(t_1, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}) = \widehat{\boldsymbol{F}}(t_2, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}).$$

Thus $\widehat{\boldsymbol{F}}$ is independent of t, which is the assertion of the proposition. ■

It is easy to see that the converse of the preceding proposition is not generally true. This is because, while a differential equation uniquely determines its right-hand side, a right-hand side does not uniquely determine a differential equation. This is pursued in Exercise 3.1.20.

3.1.3.2 Linear ordinary differential equations Next we turn to a very important class of ordinary differential equations, namely those that are linear.

3.1.12 Definition (Linear ordinary differential equation) Let

$$\boldsymbol{F}\colon \mathbb{T} \times \mathbb{R}^m \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

be an ordinary differential equation with state space $X = \mathbb{R}^m$. The ordinary differential equation $\boldsymbol{F}$ is:

(i) ***linear*** if, for each $t \in \mathbb{T}$, the map

$$\begin{aligned} \boldsymbol{F}_t\colon \mathbb{R}^m \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) &\to \mathbb{R}^m \\ (x, x^{(1)}, \dots, x^{(k)}) &\mapsto \boldsymbol{F}(t, x, x^{(1)}, \dots, x^{(k)}) \end{aligned}$$

is affine;

(ii) ***linear homogeneous*** if, for each $t \in \mathbb{T}$, the map $\boldsymbol{F}_t$ is linear;

(iii) ***linear inhomogeneous*** if it is linear but not linear homogeneous. •

Before we get to examples, let us characterise linearity in terms of the right-hand side of the ordinary differential equation.

3.1.13 Proposition (Right-hand sides of linear ordinary differential equations) *Let*

$$\mathbf{F}\colon \mathbb{T} \times \mathbb{R}^{\mathrm{m}} \oplus \mathrm{L}^{\leq \mathrm{k}}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

be an ordinary differential equation with right-hand side

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathbb{R}^{\mathrm{m}} \oplus \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

The following statements hold:

(i) if $\mathbf{F}$ *is linear, then, for each* $\mathrm{t} \in \mathbb{T}$*, the map*

$$\begin{aligned} \widehat{\mathbf{F}}_{\mathrm{t}}\colon \mathbb{R}^{\mathrm{m}} \oplus \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) &\to \mathbb{R}^{\mathrm{m}} \\ (\mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}) &\mapsto \widehat{\mathbf{F}}(\mathrm{t}, \mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}) \end{aligned}$$

is affine;

(ii) if $\mathbf{F}$ *is linear homogeneous, then, for each* $\mathrm{t} \in \mathbb{T}$*, the map* $\widehat{\mathbf{F}}_{\mathrm{t}}$ *is linear;*

(iii) if $\mathbf{F}$ *is linear inhomogeneous, then, for each* $\mathrm{t} \in \mathbb{T}$*, the map* $\widehat{\mathbf{F}}_{\mathrm{t}}$ *is affine but not linear.*

Proof (i) Fix $t \in \mathbb{T}$. Since $\boldsymbol{F}_t$ is affine, there exists $\boldsymbol{L}_{0,t} \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$,

$$\boldsymbol{L}_{j,t} \in \mathrm{L}(\mathrm{L}^j_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m); \mathbb{R}^m), \qquad j \in \{1, \dots, k\},$$

and $\boldsymbol{b}_t \in \mathbb{R}^m$ such that

$$\boldsymbol{F}_t(x, x^{(1)}, \dots, x^{(k)}) = \boldsymbol{L}_{k,t}(x^{(k)}) + \cdots + \boldsymbol{L}_{1,t}(x^{(1)}) + \boldsymbol{L}_{0,t}(x) + \boldsymbol{b}_t. \tag{3.2}$$

Keeping in mind Remark 3.1.5, we have

$$\mathrm{L}(\mathrm{L}^j_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m); \mathbb{R}^m) \simeq \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m), \qquad j \in \{1, \dots, m\},$$

and so we can use this identification to think of $\boldsymbol{x}^{(j)}$, $j \in \{1, \dots, m\}$, as being in $\mathbb{R}^m$ and the linear maps $\boldsymbol{L}_{j,t}$ as being elements of $\mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$. We will denote by $\boldsymbol{A}_{j,t} \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ the corresponding linear maps, so equation (3.2) reads

$$\boldsymbol{F}_t(\boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) = \boldsymbol{A}_{k,t}(\boldsymbol{x}^{(k)}) + \cdots + \boldsymbol{A}_{1,t}(\boldsymbol{x}^{(1)}) + \boldsymbol{A}_{0,t}(\boldsymbol{x}) + \boldsymbol{b}_t.$$

Since $\boldsymbol{F}$ is an ordinary differential equation, $\boldsymbol{A}_{k,t}$ must be invertible, and we must also have

$$\widehat{\boldsymbol{F}}_t(\boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}) = -\boldsymbol{A}_{k,t}^{-1} \circ \boldsymbol{A}_{0,t}(\boldsymbol{x}) - \boldsymbol{A}_{k,t}^{-1} \circ \boldsymbol{A}_{1,t}(\boldsymbol{x}^{(1)}) - \cdots - \boldsymbol{A}_{k,t}^{-1} \circ \boldsymbol{A}_{k-1,t}(\boldsymbol{x}^{(k-1)}) - \boldsymbol{A}_{k,t}^{-1}(\boldsymbol{b}_t).$$

This gives the desired conclusion that $\widehat{\boldsymbol{F}}_t$ is affine.

(ii) This follows from the calculations of part (i), but with $\boldsymbol{b}_t = \boldsymbol{0}$.

(iii) This follows from parts (i) and (ii). ■

As with Proposition 3.1.11, the converses to the statements in the preceding result are generally false, and the reader can explore this in Exercise 3.1.21.

The proof of the proposition reveals the form for linear ordinary differential equations, and we reproduce this here outside the proof for emphasis. To wit, a differential equation

$$\boldsymbol{F} \colon \mathbb{T} \times \mathbb{R}^m \oplus \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

is linear if and only if there exist maps

$$\boldsymbol{A}_j \colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m), \qquad j \in \{0, 1, \dots, k\},$$

and $\boldsymbol{b} \colon \mathbb{T} \to \mathbb{R}^m$ such that

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) = \boldsymbol{A}_k(t)(\boldsymbol{x}^{(k)}) + \cdots + \boldsymbol{A}_1(t)(\boldsymbol{x}^{(1)}) + \boldsymbol{A}_0(t)(\boldsymbol{x}) + \boldsymbol{b}(t). \tag{3.3}$$

The right-hand side is then

$$-\boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_0(t)(\boldsymbol{x}) - \boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_1(t)(\boldsymbol{x}^{(1)}) - \cdots - \boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_{k-1}(t)(\boldsymbol{x}^{(k-1)}) - \boldsymbol{A}_k^{-1}(t)(\boldsymbol{b}(t)).$$

Solutions to this ordinary differential equation are then functions $t \mapsto \boldsymbol{x}(t)$ satisfying

$$\frac{\mathrm{d}^k \boldsymbol{x}}{\mathrm{d}t^k}(t) = -\boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_0(t)(\boldsymbol{x}(t)) - \boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_1(t)\left(\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}(t)\right) - \dots - \boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_{k-1}(t)\left(\frac{\mathrm{d}^{k-1}\boldsymbol{x}}{\mathrm{d}t^{k-1}}(t)\right) - \boldsymbol{A}_k^{-1}(t)(\boldsymbol{b}(t)).$$

We shall study equations like this in great detail subsequently, particularly in the case when the linear maps $\boldsymbol{A}_0, \boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ are independent of t. Indeed, equations like this have a particular name.

3.1.14 Definition (Constant coefficient linear ordinary differential equation) A linear ordinary differential equation given by (3.3) is a ***constant coefficient linear ordinary differential equation*** if the functions $\boldsymbol{A}_0, \boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ are independent of t. •

Let us consider the examples of Section 1.1 in terms of their linearity.

3.1.15 Examples (Linear ordinary differential equations (or not))

1. The mass-spring-damper equation we derived in (1.1) is an autonomous linear constant coefficient inhomogeneous ordinary differential equation. According to the notation of (3.3), we have
$$A_2 = m,\ A_1 = d,\ A_0 = k,\ b = -ma_g.$$
2. The coupled mass-spring-damper equation of (1.2) is an autonomous linear constant coefficient homogeneous ordinary differential equations. According to the notation of (3.3), we have
$$\boldsymbol{A}_2 = \begin{bmatrix} m & 0 \\ 0 & m \end{bmatrix},\ \boldsymbol{A}_1 = \boldsymbol{0},\ \boldsymbol{A}_0 = \begin{bmatrix} 2k & -k \\ -k & 2k \end{bmatrix},\ \boldsymbol{b} = \boldsymbol{0}.$$
3. For the simple pendulum equation of (1.3), we leave the working out of its attributes as Exercise 3.1.22.
4. For Bessel's equation (1.3), we leave the working out of its attributes as Exercise 3.1.22.
5. For the current in a series RLC circuit of simple pendulum equation of (1.6), we leave the working out of its attributes as Exercise 3.1.22.
6. For the tank flow model of (1.7), we leave the working out of its attributes as Exercise 3.1.22.
7. For the logistical population model of (1.8), we leave the working out of its attributes as Exercise 3.1.22.
8. For the Lotka–Volterra predator prey model of (1.9), we leave the working out of its attributes as Exercise 3.1.22.
9. For the Rapoport production and exchange model of (1.10), we leave the working out of its attributes as Exercise 3.1.22. •

3.1.3.3 Linear ordinary differential equations in vector spaces

In Chapter 5 we will work with linear ordinary differential equations and will, at times, delve quite deeply into the algebraic structure of such equations. This will be followed up on when we work with linear systems described by differential equations. In these cases, it is advantageous to consider state spaces that are abstract finite-dimensional vector spaces, rather than the specific $\mathbb{R}^n$. Indeed, the extra structure of $\mathbb{R}^n$ with its annoying standard basis, standard inner product, etc., can be a real distraction when ones goal is to understand other structure. In this short section we develop the tools required to talk about differential equations with a finite-dimensional $\mathbb{R}$-vector space as state space.

The following is the basic definition.

3.1.16 Definition (System of linear ordinary differential equations) Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let V be an n-dimensional $\mathbb{F}$-vector space.

(i) A ***system of linear ordinary differential equations*** in V is a map $F\colon \mathbb{T}\times\mathsf{V}\oplus\mathsf{V} \to \mathsf{V}$ of the form

$$F(t, x, x^{(1)}) = \mathsf{A}_1(t)(x^{(1)}) + \mathsf{A}_0(t)(x) - b_0(t)$$

for maps $\mathsf{A}_0, \mathsf{A}_1\colon \mathbb{T} \to \mathrm{L}(\mathsf{V};\mathsf{V})$ and $b_0\colon \mathbb{T} \to \mathsf{V}$, where $\mathsf{A}_1(t)$ is invertible for every $t \in \mathbb{T}$.

(ii) The ***right-hand side*** of a system of linear ordinary differential equations F is the map $\widehat{F}\colon \mathbb{T} \times \mathsf{V} \to \mathsf{V}$ is the map defined by

$$\widehat{F}(t, x) = -\mathsf{A}_1(t)^{-1} \circ \mathsf{A}_0(t)(x) + \mathsf{A}_1(t)^{-1}(b_0(t)).$$

We shall typically denote $\mathsf{A}(t) = -\mathsf{A}_1(t)^{-1} \circ \mathsf{A}_0(t)$ and $b(t) = \mathsf{A}_1(t)^{-1}(b_0(t))$.

(iii) The system of linear ordinary differential equations F

(a) is ***homogeneous*** if $b(t) = 0$ for every $t \in \mathbb{T}$,

(b) is ***inhomogeneous*** if $b(t) \neq 0$ for some $t \in \mathbb{T}$, and

(c) has ***constant coefficients*** if A is a constant map.

(iv) A ***solution*** for a system of linear ordinary differential equations F is a map $\xi \in \mathsf{C}^1(\mathbb{T}';\mathsf{V})$ defined on a subinterval $\mathbb{T}' \subseteq \mathbb{T}$ and satisfying

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) = \mathsf{A}(t)(\xi(t)) + b(t), \qquad t \in \mathbb{T}'. \qquad \bullet$$

3.1.4 Partial differential equations

In the preceding section we called differential equations with one independent variable, and satisfying a certain nondegeneracy condition, "ordinary differential equations." The other kind of differential equations are what we define next.

To do so, we introduce some useful general notation for the various variables and for the derivative coordinates. Independent variables will be denoted by $\boldsymbol{x}$ and states or unknowns by $\boldsymbol{u}$. Then the list of the coordinates representing the derivatives up to order k of the dependent variables with respect to the independent variables will be denoted by

$$(\boldsymbol{u}, \boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(k)}) \in U \times \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}^n;\mathbb{R}^m).$$

Note that, in the general case when $n > 1$, the simplifications of Remark 3.1.5 do not apply, and each of the derivative variables lives in a different space.

3.1.4.1 General partial differential equations

We begin with the definition.

3.1.17 Definition (Partial differential equation) A *partial differential equation* is a differential equation

$$\boldsymbol{F}\colon D \times U \times \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l$$

with the following properties:

(i) $n > 1$;

(ii) there exists $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(k-1)}) \in D \times U \times \mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m)$ such that the function

$$\boldsymbol{u}^{(k)} \mapsto \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(k-1)}, \boldsymbol{u}^{(k)})$$

is not constant. •

The second condition merits explanation. It serves a similar function to the nondegeneracy condition (v) of Definition 3.1.6 for ordinary differential equation. In the case of ordinary differential equations, we wished to be able to solve for the highest-order derivative. For partial differential equations, this is asking too much as it is typically *not* the case that the entire highest-order derivative can be solved for. However, the condition we give is that $\boldsymbol{F}$ should not be everywhere independent of the highest-order derivative. This is a condition that, while technically required for a sensible notion of order for a partial differential equation, is always met in practice.

There is not much to say about general partial differential equations. All of the examples of Section 1.1 that have more than one independent variable are partial differential equations as per Definition 3.1.17. The dichotomy into autonomous and nonautonomous equations is not so interesting for partial differential equations, so we do not give the definition here, although it is possible to do so. We also comment that there is no natural notion of a right-hand side for a partial differential equation as there is for an ordinary differential equation.

Thus we begin our specialisation of partial differential equations with various flavours of linearity.

3.1.4.2 Linear and quasilinear partial differential equations

Let us provide the appropriate definitions of linearity for partial differential equations.

3.1.18 Definition (Linear partial differential equation) Let

$$\boldsymbol{F}\colon D \times \mathbb{R}^m \oplus \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l$$

be a partial differential equation with state space $U = \mathbb{R}^m$. The partial differential equation $\boldsymbol{F}$ is:

(i) ***linear*** if, for each $\boldsymbol{x} \in D$, the map

$$\begin{aligned}\boldsymbol{F}_{\boldsymbol{x}}\colon \mathbb{R}^m \oplus \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) &\to \mathbb{R}^l\\ (\boldsymbol{u}, \boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(k)}) &\mapsto \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(k)})\end{aligned}$$

is affine;

(ii) ***linear homogeneous*** if, for each $x \in D$, the map F_x is linear;
(iii) ***linear inhomogeneous*** if it is linear but not linear homogeneous. •

3.1.19 Definition (Quasilinear partial differential equation) A partial differential equation

$$F\colon D \times U \times \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l$$

is ***quasilinear*** if, for each

$$(x, u, u^{(1)}, \dots, u^{(k-1)}) \in D \times U \times \mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m),$$

the map

$$u^{(k)} \mapsto F(x, u, u^{(1)}, \dots, u^{(k)})$$

is affine. •

We can immediately deduce from the definitions the following forms for the various flavours of linear and quasilinear partial differential equations.

3.1.20 Proposition (Linear partial differential equations) *Let*

$$\mathbf{F}\colon \mathrm{D} \times \mathbb{R}^{\mathrm{m}} \oplus \mathrm{L}^{\le \mathrm{k}}_{\mathrm{sym}}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{l}}$$

be a partial differential equation with state space $\mathrm{U} = \mathbb{R}^{\mathrm{m}}$. *Then the following statements hold:*

(i) $\mathbf{F}$ *is linear if and only if there exist maps*

$$\mathbf{A}_{\mathrm{j}}\colon \mathrm{D} \to \mathrm{L}(\mathrm{L}^{\mathrm{j}}_{\mathrm{sym}}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}}); \mathbb{R}^{\mathrm{l}}), \qquad \mathrm{j} \in \{0, 1, \dots, \mathrm{k}\},$$

and $\mathbf{b}\colon \mathrm{D} \to \mathbb{R}^{\mathrm{l}}$, *with* $\mathbf{A}_{\mathrm{k}}$ *not identically zero, such that*

$$\mathbf{F}(\mathbf{x}, \mathbf{u}, \mathbf{u}^{(1)}, \dots, \mathbf{u}^{(\mathrm{k})}) = \mathbf{A}_{\mathrm{k}}(\mathbf{x})(\mathbf{u}^{(\mathrm{k})}) + \cdots + \mathbf{A}_1(\mathbf{x})(\mathbf{u}^{(1)}) + \mathbf{A}_0(\mathbf{x})(\mathbf{u}) + \mathbf{b}(\mathbf{x}); \quad (3.4)$$

(ii) $\mathbf{F}$ *is linear homogeneous if and only if it has the form from part (i) with* $\mathbf{b}(\mathbf{x}) = \mathbf{0}$ *for every* $\mathbf{x} \in \mathrm{D}$;

(iii) $\mathbf{F}$ *is linear inhomogeneous if and only if it has the form from part (i) with* $\mathbf{b}(\mathbf{x}) \neq \mathbf{0}$ *for some* $\mathbf{x} \in \mathrm{D}$.

3.1.21 Proposition (Quasilinear partial differential equations) *A partial differential equation*

$$\mathbf{F}\colon \mathrm{D} \times \mathrm{U} \times \mathrm{L}^{\le \mathrm{k}}_{\mathrm{sym}}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{l}}$$

is quasilinear if and only if there exist maps

$$\mathbf{A}_1\colon \mathrm{D} \times \mathrm{U} \times \mathrm{L}^{\le \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}}) \to \mathrm{L}(\mathrm{L}^{\mathrm{k}}_{\mathrm{sym}}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}}); \mathbb{R}^{\mathrm{l}}),$$
$$\mathbf{A}_0\colon \mathrm{D} \times \mathrm{U} \times \mathrm{L}^{\le \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{l}},$$

with $\mathbf{A}_1$ *not identically zero, such that*

$$\mathrm{F}(\mathbf{x}, \mathbf{u}, \mathbf{u}^{(1)}, \dots, \mathbf{u}^{(\mathrm{k})}) = \mathbf{A}_1(\mathbf{x}, \mathbf{u}, \mathbf{u}^{(1)}, \dots, \mathbf{u}^{(\mathrm{k}-1)})(\mathbf{u}^{(\mathrm{k})}) + \mathbf{A}_0(\mathbf{x}, \mathbf{u}, \mathbf{u}^{(1)}, \dots, \mathbf{u}^{(\mathrm{k}-1)}).$$

The notion of having constant coefficients that we encountered for ordinary differential equations also makes sense for partial differential equations.

3.1.22 Definition (Constant coefficient linear partial differential equation) A linear partial differential equation given by (3.4) is a ***constant coefficient linear partial differential equation*** if the functions $A_0, A_1, \dots, A_k$ are constant. •

We leave to the reader in Exercise 3.1.25 the pleasure of classifying the example partial differential equations of Section 1.1.

3.1.4.3 Elliptic, hyperbolic, and parabolic second-order linear partial differential equations Many of the partial differential equations that arise from physics are linear second-order equations with a single unknown, and there are various classifications that can be applied to such equations that bear on the attributes of the solutions to these equations.

Let us write the general form of such a differential equation. In doing so, let us remind ourselves what our derivative notation means in this case. We will deal with derivatives of a single variable of at most second-order, so the first derivative $u^{(1)}$ represents a vector of partial derivatives

$$u^{(1)} = (u_{x_1}, \dots, u_{x_n})$$

and $u^{(2)}$ represents a matrix of partial derivatives

$$u^{(2)} = \begin{bmatrix} u_{x_1x_1} & u_{x_1x_2} & \cdots & u_{x_1x_n} \\ u_{x_2x_1} & u_{x_2x_2} & \cdots & u_{x_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{x_nx_1} & u_{x_nx_2} & \cdots & u_{x_nx_n} \end{bmatrix},$$

keeping in mind that this matrix will be symmetric. With this in mind, a general linear second-order partial differential equation will have the form

$$F(\boldsymbol{x}, u, u^{(1)}, u^{(2)}) = \sum_{j,k=1}^{n} A_{jk}(\boldsymbol{x}) u_{x_j x_k} + \sum_{j=1}^{n} a_j(\boldsymbol{x}) u_{x_j} + b(\boldsymbol{x}) \tag{3.5}$$

for functions

$$\boldsymbol{A} \colon D \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n), \quad \boldsymbol{a} \colon D \to \mathbb{R}^n, \quad b \colon D \to \mathbb{R}.$$

We can, without loss of generality, suppose that $\boldsymbol{A}(\boldsymbol{x})$ is a symmetric matrix for all $\boldsymbol{x} \in D$.[1] In this case, we know that the eigenvalues of $\boldsymbol{A}$ are real, allowing the following definition.

[1] Indeed, suppose that $\boldsymbol{A}$ is not symmetric. Then write $\boldsymbol{A}$ as a sum of a symmetric and skew-symmetric matrix:

$$\boldsymbol{A} = \underbrace{\tfrac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^T)}_{\boldsymbol{A}^+} + \underbrace{\tfrac{1}{2}(\boldsymbol{A} - \boldsymbol{A}^T)}_{\boldsymbol{A}^-},$$

3.1.23 Definition (Elliptic, hyperbolic, parabolic) Let

$$F\colon D \times \mathbb{R} \oplus \mathrm{L}^{\leq 2}_{\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}) \to \mathbb{R}$$

be a second-order linear partial differential equation, and so given by (3.5). Then $\boldsymbol{F}$ is:

(i) ***elliptic*** at $\boldsymbol{x} \in D$ if all eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$ are positive;

(ii) ***hyperbolic*** at $\boldsymbol{x} \in D$ if all eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$ are nonzero;

(iii) ***parabolic*** at $\boldsymbol{x} \in D$ if all eigenvalues of $\boldsymbol{A}(\boldsymbol{x})$ are nonnegative, and at least one of them is zero. •

Note that if $\boldsymbol{F}$ has constant coefficients, then the notion of being in one of the three cases of elliptic, hyperbolic, or parabolic does not depend on $\boldsymbol{x} \in D$. Generally, however, it will. Thus the notions are most frequently applied in the constant coefficient case. Let us consider examples that we have seen thus far, and see where they sit relative to the elliptic/hyperbolic/parabolic classification.

3.1.24 Examples (Elliptic, hyperbolic, and parabolic partial differential equations)

1. The standard example of an elliptic partial differential equation is the ***potential equation***, or ***Laplace's equation***. The domain $D \subseteq \mathbb{R}^n$ is normally thought of as being "space" in this case, so we denote coordinates for D by $(x_1, \dots, x_n)$. Then the differential equation is given by

$$\boldsymbol{F}(\boldsymbol{x}, u, u^{(1)}, u^{(2)}) = u_{x_1x_1} + \cdots + u_{x_nx_n}.$$

Thus $u\colon D' \to \mathbb{R}$ is a solution if it satisfies

$$\frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} = 0.$$

---

with $\boldsymbol{A}^+$ being symmetric and $\boldsymbol{A}^-$ being skew-symmetric. Then we have

$$\sum_{j,k=1}^n A^-_{jk} u_{x_jx_k} = -\sum_{j,k=1}^n A^-_{kj} u_{x_jx_k} = -\sum_{j,k=1}^n A^-_{kj} u_{x_kx_j} = -\sum_{j,k=1}^n A^-_{jk} u_{x_jx_k},$$

and so we conclude that

$$\sum_{j,k=1}^n A^-_{jk} u_{x_jx_k} = 0,$$

and so

$$\sum_{j,k=1}^n A_{jk} u_{x_jx_k} = \sum_{j,k=1}^n A^+_{jk} u_{x_jx_k},$$

giving our claim that we can assume that $\boldsymbol{A}$ is symmetric.

We saw examples of how this equation arises in applications in Section 1.1.13. Note that, in this case,

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

so all eigenvalues are 1, i.e., are positive. This ensures that $\boldsymbol{F}$ in this case is indeed elliptic.

2. The standard example of an hyperbolic partial differential equation is the ***wave equation***. In this case, the domain D is normally thought of as encoding time and space, and so we denote coordinates by $(t, x_1, \ldots, x_n)$. The differential equation is given by

$$F((t, \boldsymbol{x}), u, u^{(1)}, u^{(2)}) = -u_{tt} + u_{x_1x_1} + \cdots + u_{x_nx_n}.$$

Solutions u thus satisfy the equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}.$$

We saw that in Section 1.1.12 that the wave equation arises in the model of the transverse vibrations of a taut string. In this case we have

$$\boldsymbol{A} = \begin{bmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

and so the eigenvalues are $-1, 1, \ldots, 1$, showing that this is indeed an hyperbolic equation.

3. The usual example of a parabolic equation is the ***heat equation***, which we saw modelled the temperature distribution in a rod in Section 1.1.11. In this case, like the wave equation, the domain D is coordinatised by time and space: $(t, x_1, \ldots, x_n)$. The differential equation is

$$F((t, \boldsymbol{x}), u, u^{(1)}, u^{(2)}) = -u_t + u_{x_1x_1} + \cdots + u_{x_nx_n}.$$

Solutions $u \colon D' \to \mathbb{R}$ satisfy

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}.$$

In this case

$$A = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

and so the eigenvalues are $0, 1, \ldots, 1$, showing that this is indeed a parabolic equation. •

3.1.5 How to think about differential equations

A reader having read and understood the content of this section will have an excellent understanding of what a differential equation is, and some of the special classes of differential equations. The reader will subsequently embark on a mission to actually *solve* some ordinary differential equations. Before doing so, it is worth putting this process of solving differential equations into a general context.

First of all, let us state very clearly: *if you reach into the bag of differential equations and pull one out, it is extremely unlikely you will be able to solve it*. This is rather like what a student has already encountered in their study of differentiation and integration; one has at hand a small but important collection of functions that one can actually differentiate or integrate, and these are to be regarded as isolated and valuable gems. But this does raise the question of what one can *do* with a differential equation pulled at random from the bag of differential equations.

Let us explore this a little.

1. *Analysis:* Even if one cannot explicitly solve a given differential equation, there are still sometimes things that can be done to get some insight into its behaviour. Let us consider some of the things one might try to do.
   (a) *Understand steady-state behaviour:* In some equations one has time t as the, or one of the, independent variables. In such cases, it is often of interest to understand the behaviour of solutions as $t \to \infty$. This behaviour is known as ***steady-state*** behaviour. Sometimes the steady-state behaviour is not interesting, as in "blows up to infinity." But sometimes this behaviour is all one really wants, and sometimes it can even be determined. We shall see some instances of this sort of investigation in the text.
   (b) *Approximating solutions:* Sometimes in a differential equation there are effects that are dominant, and the remaining effects can be regarded as "perturbations" of these dominant effects. If the dominant part of the equations are something that one can understand, one can hope (pray, really) that the perturbations do not materially affect the dominant behaviour. In practice, methods like this should be used with great care, since the "perturbations," while small, may have significant impact on the character of solutions, particularly for long times in cases where time is

one of the independent variables. However, there are cases where "perturbation theory" can be applied to give useful conclusions. However, this is not something we will get deeply into in any sort of general way.

(c) *Equilibria and their stability:* A special case of the preceding idea of approximation involves the study of equilibria. This is most easily discussed by reference to ordinary differential equations, but the basic ideas can be adapted by a flexible mind to partial differential equations. Suppose that we have an ordinary differential equation

$$\boldsymbol{F}\colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m.$$

An ***equilibrium*** is a point $\boldsymbol{x}_0 \in X$ for which

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{0}, \dots, \boldsymbol{0}) = \boldsymbol{0}, \qquad t \in \mathbb{T}.$$

Note that the constant function $t \mapsto \boldsymbol{x}_0$ is then a solution of this differential equation. The fact that it is constant is what leads to its being called an "equilibrium." One can then consider the ***stability*** of this equilibrium, which loosely means the matter of whether solutions starting near $\boldsymbol{x}_0$ (i) remain near $\boldsymbol{x}_0$, (ii) approach $\boldsymbol{x}_0$ as $t \to \infty$, or (iii) diverge away from $\boldsymbol{x}_0$. We shall be precise about this in the text in various situations.

2. *Numerical solution:* One can attempt to use a computer to solve the differential equation. For most ordinary differential equations, there are reliable methods for solving them numerically. The situation with partial differential equations is quite different, and significant science has been, is, and will be dedicated to numerical techniques for solving partial differential equations. In the text we will talk a little about using numerical methods to solve ordinary differential equations, and will give the reader some opportunity to use the standard package MATLAB® for plotting numerical solutions to differential equations.

   While this is definitely *not* a text on numerical methods, it is worth understanding a little bit of what is under the hood when one is using a computer package to obtain numerical solutions to ordinary differential equations.

   The basic step in converting an ordinary differential equation into something that can be worked with numerically is to replace derivatives with algebraic approximations. Suppose that one has a function $t \mapsto \xi(t)$. The obvious thing to do to approximate the derivative of ξ is to work with the standard difference quotient:

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) \approx \frac{\xi(t+h) - \xi(t)}{h}.$$

   Here, $h \in \mathbb{R}_{>0}$ is to be thought of as small (in the limit as $h \to 0$ we get the actual derivative, if it exists), and is known as the ***time step***. Even here, there are multiple ways in which one might work with such a difference quotient; for

example, here are two:

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) \approx \frac{\xi(t) - \xi(t-h)}{h}, \qquad \frac{\mathrm{d}\xi}{\mathrm{d}t}(t) \approx \frac{\xi(t+\frac{h}{2}) - \xi(t-\frac{h}{2})}{h}.$$

The first rule is call the "forward difference," the second the "backward difference," and the third the "midpoint rule." If one knows the value of ξ at time t_0, one can then get an approximation for the value of ξ at time $t_0 + h$ by

$$\xi(t_0 + h) = h\frac{\mathrm{d}\xi}{\mathrm{d}t}(t_0) + \xi(t + 0),$$

then the value at time $t_0 + 2h$ by

$$\xi(t_0 + 2h) = h\frac{\mathrm{d}\xi}{\mathrm{d}t}(t_0 + h) + \xi(t_0 + h).$$

Then can, of course, be repeated, provided one has values for the derivatives. However, if ξ is the solution to a first-order scalar ordinary differential equation F with right-hand side $\widehat{F}$,

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)),$$

then one indeed does have the values for the derivatives. Indeed, one have

$$\begin{aligned} \xi(t_0 + h) &= h\widehat{F}(t_0, \xi(t_0)) + \xi(t_0), \\ \xi(t_0 + 2h) &= h\widehat{F}(t_0 + h, \xi(t_0 + h)) + \xi(t_0 + h), \\ &\vdots \end{aligned}$$

Thus we have determined a simple means of numerically generating an approximation for a solution for F given an initial condition!

We note, however, that any numerical computation package will use a much more sophisticated method for approximating derivatives than the forward difference method we have used above. Nonetheless, the basic principle is as we have outlined it in our simple illustration above. This is explored in the simple setting, where one can say a few precise facts, in Section 5.10.

A matter related to what one can *do* with a differential equation is the manner in which one can think of a solution, since it is solutions in which we are interested. No matter what else you do, here is how you should *not* think about solutions:

> ***Be a grown up about what a solution is:*** *A solution to a differential equation, or any equation for that matter, is not a formula that you write on the page as the byproduct of some algorithmic procedure. This way of thinking about "solution" should remain in high school, which is where it was unfortunately taught to you.*

So. . . how *should* you think about what a solution is?

For ordinary differential equations, a profitable way to think about it is to think about curves, since a solution is indeed a curve $t \mapsto \boldsymbol{x}(t)$. Let us focus on first-order ordinary differential equations.[2] In this case, $\dot{\boldsymbol{x}}(t)$ is the tangent vector to this curve, and so the equation

$$\dot{\boldsymbol{x}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{x}(t))$$

should be thought of as prescribing the tangent vectors to solution curves. What becomes important, then is the vector $\widehat{\boldsymbol{F}}(t, \boldsymbol{x})$ one assigns to the point $(t, \boldsymbol{x})$.

Let us be explicit about this in an example.

3.1.25 Example (Ordinary differential equations and vector fields) We consider the autonomous first-order ordinary differential equation in two unknowns defined by

$$\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = \widehat{\boldsymbol{F}}_0(x_1, x_2) = (x_2, -x_1 + \tfrac{1}{2}x_2(1 - x_1^2)).$$

Thus solutions are defined by the equations

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= -x_1(t) + \tfrac{1}{2}x_2(t)(1 - x_1(t)^2). \end{aligned}$$

In Figure 3.1 we plot the vector field. Thus, at each point $(x_1, x_2) \in \mathbb{R}^2$ we draw an arrow in the direction of

$$\boldsymbol{F}_0(x_1, x_2) = (x_2, -x_1 + \tfrac{1}{2}x_2(1 - x_1^2)).$$

A solution to the differential equation will then be a curve $t \mapsto (x_1(t), x_2(t))$ whose tangent vector at $(x_1(t), x_2(t))$ points in the direction of $\boldsymbol{F}_0(x_1(t), x_2(t))$. In Figure 3.2 we show a few such solution curves; these are known in the business as ***integral curves***.

It is also not uncommon to look at plots of $x_1(t)$ and $x_2(t)$ as functions of t. In Figure 3.3 we show such plots starting at a fixed point $(x_1(0), x_2(0))$ at $t = 0$.[3]

We hope that a reader will find looking at pictures like this, particularly Figure 3.2, more insightful than looking at some formula for the solution, produced as a byproduct of some algorithmic procedure. Also, for this equation, there is no algorithmic procedure for determining the solutions. . . but the pictures can still be produced and offer insight. •

For partial differential equations, solutions are no longer curves, i.e., vector functions of a single independent variable, but it is still worthwhile to think about, and represent where possible, a solution as a graph of a function of the independent variables.

[2] We shall see that a kth-order ordinary differential equation can always be converted into a first-order ordinary differential equation, so the assumption of the equation being first-order is made without loss of generality.

[3] As one varies $(x_1(0), x_2(0))$, one also varies these plots, and this is something we will consider in Section 3.2.

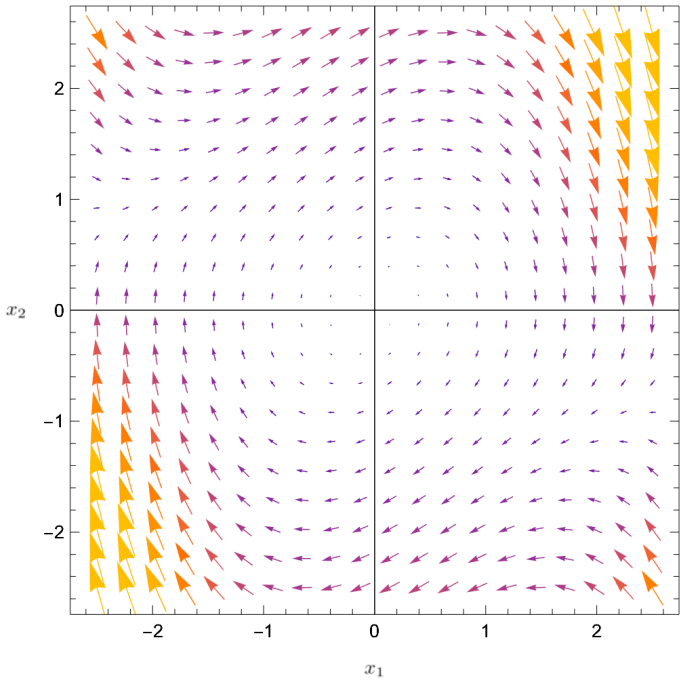


Figure 3.1 A vector field in $\mathbb{R}^2$

Exercises

3.1.1 Work out Example 3.1.3–3. Thus:
- (a) identify n, m, k, and l;
- (b) name the independent variables;
- (c) name the states;
- (d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;
- (e) write the equation that must be satisfied by a solution.

3.1.2 Work out Example 3.1.3–4. Thus:
- (a) identify n, m, k, and l;
- (b) name the independent variables;
- (c) name the states;
- (d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;
- (e) write the equation that must be satisfied by a solution.

3.1.3 Work out Example 3.1.3–5. Thus:
- (a) identify n, m, k, and l;
- (b) name the independent variables;
- (c) name the states;
- (d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;

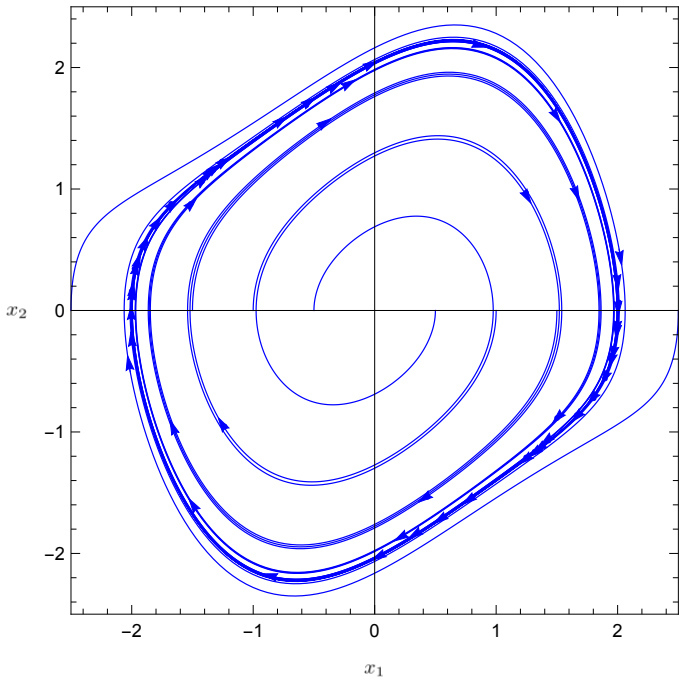


Figure 3.2 A few solution curves for the vector field of Figure 3.1

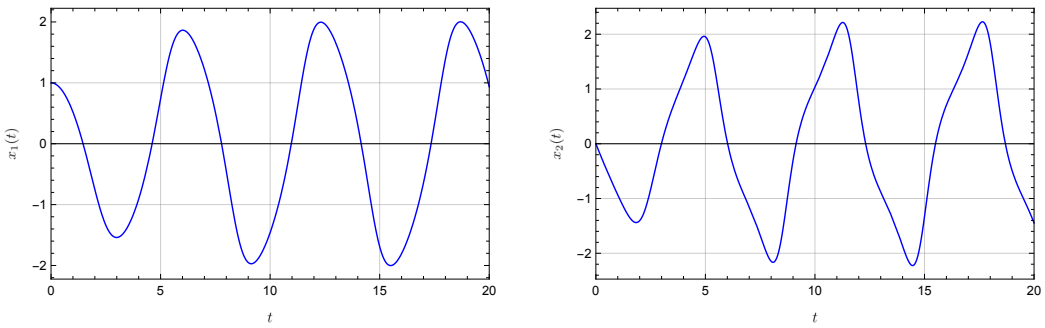


Figure 3.3 Plots of the solutions as functions of time

(e) write the equation that must be satisfied by a solution.

3.1.4 Work out Example 3.1.3–6. Thus:

(a) identify n, m, k, and l;
(b) name the independent variables;
(c) name the states;
(d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;
(e) write the equation that must be satisfied by a solution.

3.1.5 Work out Example 3.1.3–7. Thus:
- (a) identify n, m, k, and l;
- (b) name the independent variables;
- (c) name the states;
- (d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;
- (e) write the equation that must be satisfied by a solution.

3.1.6 Work out Example 3.1.3–8. Thus:
- (a) identify n, m, k, and l;
- (b) name the independent variables;
- (c) name the states;
- (d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;
- (e) write the equation that must be satisfied by a solution.

3.1.7 Work out Example 3.1.3–9. Thus:
- (a) identify n, m, k, and l;
- (b) name the independent variables;
- (c) name the states;
- (d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;
- (e) write the equation that must be satisfied by a solution.

3.1.8 Work out Example 3.1.3–14. Thus:
- (a) identify n, m, k, and l;
- (b) name the independent variables;
- (c) name the states;
- (d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;
- (e) write the equation that must be satisfied by a solution.

3.1.9 Work out Example 3.1.3–15. Thus:
- (a) identify n, m, k, and l;
- (b) name the independent variables;
- (c) name the states;
- (d) write $\boldsymbol{F}$ as a map, explicitly denoting its domain and codomain;
- (e) write the equation that must be satisfied by a solution.

3.1.10 Work out Example 3.1.8–3. Thus:
- (a) write $\boldsymbol{F}$ using the ordinary differential equation notation for derivatives;
- (b) show that $\boldsymbol{F}$ is an ordinary differential equation;
- (c) write down the right-hand side;
- (d) write the condition for a solution using Proposition 3.1.7.

3.1.11 Work out Example 3.1.8–4. Thus:
- (a) write $\boldsymbol{F}$ using the ordinary differential equation notation for derivatives;
- (b) show that $\boldsymbol{F}$ is an ordinary differential equation;

(c) write down the right-hand side;
(d) write the condition for a solution using Proposition 3.1.7.

3.1.12 Work out Example 3.1.8–5. Thus:
(a) write $\boldsymbol{F}$ using the ordinary differential equation notation for derivatives;
(b) show that $\boldsymbol{F}$ is an ordinary differential equation;
(c) write down the right-hand side;
(d) write the condition for a solution using Proposition 3.1.7.

3.1.13 Work out Example 3.1.8–6. Thus:
(a) write $\boldsymbol{F}$ using the ordinary differential equation notation for derivatives;
(b) show that $\boldsymbol{F}$ is an ordinary differential equation;
(c) write down the right-hand side;
(d) write the condition for a solution using Proposition 3.1.7.

3.1.14 Work out Example 3.1.8–7. Thus:
(a) write $\boldsymbol{F}$ using the ordinary differential equation notation for derivatives;
(b) show that $\boldsymbol{F}$ is an ordinary differential equation;
(c) write down the right-hand side;
(d) write the condition for a solution using Proposition 3.1.7.

3.1.15 Work out Example 3.1.8–8. Thus:
(a) write $\boldsymbol{F}$ using the ordinary differential equation notation for derivatives;
(b) show that $\boldsymbol{F}$ is an ordinary differential equation;
(c) write down the right-hand side;
(d) write the condition for a solution using Proposition 3.1.7.

3.1.16 Work out Example 3.1.8–9. Thus:
(a) write $\boldsymbol{F}$ using the ordinary differential equation notation for derivatives;
(b) show that $\boldsymbol{F}$ is an ordinary differential equation;
(c) write down the right-hand side;
(d) write the condition for a solution using Proposition 3.1.7.

3.1.17 For each of the following ordinary differential equations $\boldsymbol{F}$, determine their right-hand sides:
(a) $F(t, x, x^{(1)}, x^{(2)}) = 3(1 + t^2)x^{(2)}$;
(b) $\boldsymbol{F}(t, (x_1, x_2), (x_1^{(1)}, x_2^{(1)})) = (x_2^{(1)} + 2x_1 - x_2, -x_1^{(1)} - x_1^2)$;
(c) $F(t, x, x^{(1)}, x^{(2)}, x^{(3)}) = -x^{(3)} + t(x^{(1)})^2 + \sin(x)$;
(d) $\boldsymbol{F}(t, (x_1, x_2), (x_1^{(1)}, x_2^{(1)})) = (-x_1^{(1)} + x_2^{(1)} + x_1^2 - x_2, 2x_1^{(1)} + 2x_2^{(1)} + \cos(x_2) - x_1)$;
(e) $F(t, x, x^{(1)}) = x^{(1)} + a(t)x$.

3.1.18 For each of the following right-hand sides $\widehat{F}$, determine the associated normalised ordinary differential equation $\boldsymbol{F}$:
(a) $\widehat{F}(t, x, x^{(1)}) = 0$;
(b) $\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = (-x_1^2, -2x_2 + x_2)$;

(c) $\widehat{F}(t, x, x^{(1)}, x^{(2)}) = t(x^{(1)})^2 + \sin(x)$;
(d) $\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = (\frac{1}{4}(x_1 + 2x_1^2 - 2x_2 - \cos(x_2)), \frac{1}{4}(x_1 - 2x_1^2 + 2x_2 - \cos(x_2)))$;
(e) $\widehat{F}(t, x) = -a(t)x$.

In the next exercise we shall show how autonomous ordinary differential equations are special in terms of their solutions. In order for the exercise to make sense, we require the existence and uniqueness theorem we state below, Theorem 3.2.8.

3.1.19 Let

$$\boldsymbol{F} \colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

be an autonomous ordinary differential equation satisfying the conditions of Theorem 3.2.8(ii), let

$$(\boldsymbol{x}_0, \boldsymbol{x}_0^{(1)}, \dots, \boldsymbol{x}_0^{(k-1)}) \in X \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m),$$

and let $t_1, t_2 \in \mathbb{T}$. Let $\xi_1 \colon \mathbb{T} \to X$ and $\xi_2 \colon \mathbb{T} \to X$ be solutions for $\boldsymbol{F}$ satisfying

$$\xi_1(t_1) = \xi_2(t_2) = \boldsymbol{x}_0, \quad \frac{\mathrm{d}^j \xi_1}{\mathrm{d}t^j}(t_1) = \frac{\mathrm{d}^j \xi_2}{\mathrm{d}t^j}(t_2) = \boldsymbol{x}_0^{(j)}, \qquad j \in \{1, \dots, k-1\}.$$

Answer the following questions.

(a) Show that $\xi_2(t) = \xi_1(t + t_1 - t_2)$ for all $t \in \mathbb{T}$ for which $\xi_2(t)$ is defined and for which $t + t_1 - t_2 \in \mathbb{T}$.
(b) Assuming that $\mathbb{T} = \mathbb{R}$ and that all solutions are defined for all time for simplicity, express your conclusion from part (a) as a condition on the flow $\Phi^{\boldsymbol{F}}$.

3.1.20 Let us consider the following two differential equations:

$$\begin{aligned} F_1 \colon \mathbb{R} \times \mathbb{R} \times \mathrm{L}^{\leq 1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) &\to \mathbb{R} & F_2 \colon \mathbb{R} \times \mathbb{R} \times \mathrm{L}^{\leq 1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) &\to \mathbb{R} \\ (t, x, x^{(1)}) &\mapsto x^{(1)}, & (t, x, x^{(1)}) &\mapsto (1 + t^2)x^{(1)}. \end{aligned}$$

Answer the following questions.

(a) Show that both F_1 and F_1 are ordinary differential equations, and determine the right-hand sides $\widehat{F_1}$ and $\widehat{F_2}$.
(b) Show that both $\widehat{F_1}$ and $\widehat{F_2}$ are independent of t.
(c) Which of F_1 and F_2 is autonomous?

3.1.21 Let us consider the following two differential equations:

$$\begin{aligned} F_1 \colon \mathbb{R} \times \mathbb{R} \times \mathrm{L}^{\leq 1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) &\to \mathbb{R} & F_2 \colon \mathbb{R} \times \mathbb{R} \times \mathrm{L}^{\leq 1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) &\to \mathbb{R} \\ (t, x, x^{(1)}) &\mapsto x^{(1)}, & (t, x, x^{(1)}) &\mapsto (1 + x^2)x^{(1)}. \end{aligned}$$

Answer the following questions.

(a) Show that both F_1 and F_1 are ordinary differential equations, and determine the right-hand sides $\widehat{F_1}$ and $\widehat{F_2}$.
(b) Show that both $\widehat{F_1}$ and $\widehat{F_2}$ are linear.
(c) Which of F_1 and F_2 is linear?

3.1.22 Consider the ordinary differential equations of Examples 3.1.3–3 to 9.
(a) Which of the equations is autonomous?
(b) Which of the equations is linear?
(c) Which of the equations is linear and homogeneous?
(d) Which of the equations is linear and inhomogeneous?
(e) Which of the equations is a linear constant coefficient equation?

3.1.23 Let

$$\boldsymbol{F}\colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

be an ordinary differential equation with right-hand side $\widehat{\boldsymbol{F}}$. As usual, let t be the independent variable and $\boldsymbol{x}$ the state, with $\boldsymbol{x}^{(j)} \in \mathrm{L}^{j}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m)$ being the coordinate for the jth derivative. As per Remark 3.1.5, we can think of $\boldsymbol{x}^{(j)}$ as being an element of $\mathbb{R}^m$.

We will associate to $\boldsymbol{F}$ a first-order ordinary differential equation $\boldsymbol{F}_1$ with time domain $\mathbb{T}$ and state space

$$X_1 = X \times \underbrace{\mathbb{R}^m \times \cdots \times \mathbb{R}^m}_{k-1 \text{ times}}.$$

To do so, answer the following questions.

(a) Denote coordinates for the state space X_1 by $\boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_{k-1}$, and relate these to $(\boldsymbol{x}, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k-1)})$ by

$$\boldsymbol{y}_0 = \boldsymbol{x}, \ \boldsymbol{y}_j = \boldsymbol{x}^{(j)}, \qquad j \in \{1, \ldots, k-1\}.$$

If $t \mapsto \boldsymbol{x}(t)$ is a solution for $\boldsymbol{F}$, write down the corresponding differential equations that must be satisfied by $(\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1})$.
Hint: *For each* $\mathrm{j} \in \{0, 1, \ldots, \mathrm{k}-1\}$, *write down* $\dot{\mathbf{y}}_{\mathrm{j}}(\mathrm{t})$, *and express the result in terms of the coordinates for* X_1.
(b) What is the right-hand side $\widehat{\boldsymbol{F}_1}$ corresponding to the equations you derived in part (a)?
(c) Write down a first-order ordinary differential equation $\boldsymbol{F}_1$ with time domain $\mathbb{T}$ and state space X_1 whose right-hand side is the function $\widehat{\boldsymbol{F}_1}$ you determined in part (b).
(d) State *precisely* the relationship between solutions for $\boldsymbol{F}$ and solutions for $\boldsymbol{F}_1$, and show that if solutions for $\boldsymbol{F}_1$ are of class C^1, then solutions for $\boldsymbol{F}$ are of class C^k.

(e) Show that $\boldsymbol{F}_1$ can be taken to be linear if $\boldsymbol{F}$ is linear, and show that $\boldsymbol{F}_1$ is homogeneous if and only if $\boldsymbol{F}$ is, in this case.

3.1.24 Consider the motion of a projectile fired with initial velocity V_0 at an angle to the ground of θ_0. After firing, the projectile is subject to the forces of gravity and of drag. The gravitational force is directed "downwards" and has magnitude proportional to the mass of the projectile. The drag force is proportional to the square of the velocity of the projectile and is directed opposite to the direction of the velocity of the projectile.

Answer the following questions.

(a) What are m and k from Definition 3.1.6?

(b) Determine $\boldsymbol{F}$?

(c) Is the equation autonomous?

(d) Put the equation into first-order form as in Exercise 3.1.23?

3.1.25 For the partial differential equations of Examples 3.1.3–11 to 17, determine whether they are (a) linear homogeneous, (b) linear inhomogeneous, (c) quasilinear, and/or (d) has constant coefficients.

The next exercise concerns itself with the so-called method of characteristics for simple second-order linear partial differential equations. Although the presentation is for a simple class of equations, the language and methodology we introduce is readily generalised. The class of differential equations we consider are given by

$$\begin{aligned} F\colon D \times \mathbb{R} \oplus \mathrm{L}^{\leq 2}_{\mathrm{sym}}(\mathbb{R}^2;\mathbb{R}) &\to \mathbb{R} \\ (x, y, u, u^{(1)}, u^{(2)}) &\mapsto a u_{xx} + 2b u_{x,y} + d u_{yy} + d u_x + e u_y + f u \end{aligned} \tag{3.6}$$

for functions $a, b, c, d, e, f, g\colon D \to \mathbb{R}$ defined on an open subset D of $\mathbb{R}^2$. The ***symbol*** for the equation is the $\mathbb{C}$-valued function

$$\begin{aligned} \sigma(F)\colon D \times \mathbb{R}^2 &\to \mathbb{C} \\ (x, y, \xi, \eta) &\mapsto -a\xi^2 - 2b\xi\eta - c\eta^2 + \mathrm{i}d\xi + \mathrm{i}e\eta + f, \end{aligned}$$

defined by "substituting" $\mathrm{i}\xi$ for $\frac{\partial u}{\partial x}$ and $\mathrm{i}\eta$ for $\frac{\partial u}{\partial y}$. The ***principal symbol*** $\sigma_0(F)$ is the quadratic part of the symbol

$$\sigma_0(F)(x, y, \xi, \eta) = -a\xi^2 - 2b\xi\eta - c\eta^2.$$

Consider a curve in D defined by $\phi(x, y) = 0$. The curve is a ***characteristic*** if

$$\sigma_0(F)\left(x, y, \frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}\right) = 0.$$

It turns out that it is possible for a solution of a partial differential equation to have points of discontinuity, but one may determine that these are necessarily located along characteristic curves.

The above development outlines why the symmetric matrix

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

is useful in determining some properties of a partial differential equation of the form (3.6).

3.1.26 In the preceding, suppose that a, b, and c are constant, and define the function $f_{a,b,c}\colon \mathbb{R}^2 \to \mathbb{R}$ by

$$f_{a,b,c}(\xi, \eta) = a\xi^2 + 2b\xi\eta + c\eta^2,$$

and answer the following questions.

(a) Show that when $b^2 - ac = 0$ the following statements hold:
   (a) the curve $f_{a,b,c}(x, y) = 1$ is a parabola for $a > 0$;
   (b) through each point in $\mathbb{R}^2$ there passes a single characteristic for (3.6).

   Show that the heat equation falls into this category.

(c) Show that when $b^2 - ac > 0$ the following statements hold:
   (a) the curve $f_{a,b,c}(x, y) = 1$ is an hyperbola for $a > 0$;
   (b) through each point in $\mathbb{R}^2$ there passes two characteristics for (3.6).

   Show that the wave equation falls into this category.

(c) Show that when $b^2 - ac < 0$ the following statements hold:
   (a) the curve $f_{a,b,c}(x, y) = 1$ is an ellipse for $a > 0$;
   (b) the differential equation (3.6) possesses no characteristics curves.

   Show that the potential equation falls into this category.

3.1.27 *(Mini-project)* We consider a model of an Hopfield neural network with n neurons, where in Figure 3.4 we depict the jth neuron taking inputs from the

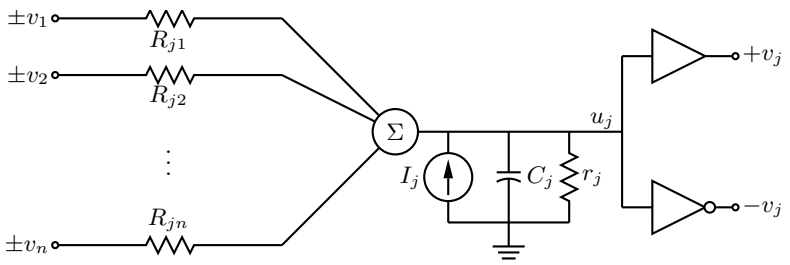


Figure 3.4 A neuron in an Hopfield neural network

other neurons. A crucial ingredient in the network is the characteristic of the amplifier which takes the input voltage u_j and produces the output voltage v_j according to a nonlinear function $g_j\colon \mathbb{R} \to [-V_{\max}, V_{\max}]$ that saturates the input to a maximum amplitude of $V_{\max}$. For example, one might take

$$g_j(u_j) = \frac{2V_{\max}}{\pi} \tan^{-1}\left(\frac{\sigma\pi u_j}{2V_{\max}}\right).$$

The inverting amplifier allows a selection of $+v_j$ or $-v_j$ to be given as input to the other neurons. The state for the system is the voltages $(v_1, \dots, v_n)$. The model here was introduced by Hopfield [1982].

We wish to assemble all of this into an ordinary differential equation.

(a) What is the state space X for the system?

(b) What is the time-domain $\mathbb{T}$ for the system?

(c) What are the dynamics $\boldsymbol{f}$?
Hint: *For each neuron, apply Kirchhoff's current law at the input to the amplifier.*

Do some explorations as follows.

(d) Do some research to describe what the model is used for and how the ordinary differential equation model should behave to be useful.

Assume that the matrix of resistances R_{jk}, $j,k \in \{1, \dots, n\}$, is symmetric with zero diagonal.

(e) Using a computer package for simulating ordinary differential equations, setup the system for simulation, and run some interesting simulations for various initial conditions.

Section 3.2

Existence and uniqueness of solutions for differential equations

The preceding section concerning differential equations was of a taxonomic nature. In this section we produce a few important results, especially for ordinary differential equations. The results are concerned with two important questions: (1) does a given differential equation possess solutions; (2) how many solutions does a differential equation possess? In mathematics, questions like this are known as questions of "existence and uniqueness" (think about similar sorts of questions for linear algebraic equations, as discussed in Section I-5.4.8.)

Do I need to read this section? If you are of the frame of mind where all equations obviously have unique solutions, then this section will appear to be pointless. However, for those capable of more subtle thought, the results in this section are essential for the understanding of the our subsequent and extensive use of differential equations, particular ordinary differential equations.

The theoretical results notwithstanding, the notion of a flow introduced in Definition 3.2.11 will be one of which we will frequently make use. •

3.2.1 Results for ordinary differential equations

We begin our discussion by looking in detail at ordinary differential equations, where a fairly complete story can be told. We shall begin by framing the sort of questions and answers we might expect by looking at some examples. Then we state the principal existence and uniqueness theorems for solutions of ordinary differential equations. We close the section by considering how all solutions of an ordinary differential equation "fit together."

Note that, by Exercise 3.1.23, it suffices to consider first-order ordinary differential equations, and so this is what we shall consider in this section.

3.2.1.1 Examples motivating existence and uniqueness of solutions for ordinary differential equations

Our first three examples make use of the fact that, when a differential has a right-hand side that is independent of the unknown, then solutions are obtained by integration.

3.2.1 Example (An ordinary differential equation with no solutions (sometimes)) We consider the scalar nonautonomous first-order differential equation with time-domain $\mathbb{R}$ and with right-hand side

$$\widehat{F}(t,x) = \begin{cases} t^{-1}, & t \neq 0, \\ 0, & t = 0. \end{cases}$$

A solution to this differential equation satisfies

$$\dot{x}(t) = f(t),$$

where

$$f(t) = \begin{cases} t^{-1}, & t \neq 0, \\ 0, & t = 0. \end{cases}$$

Since we ask that a solution be a C^1-function, the Fundamental Theorem of Calculus gives that a solution should satisfy

$$x(t) = x(t_0) + \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau.$$

We claim that, if $t_0 t \leq 0$ and if $t \neq t_0$, then the integral does not exist. Indeed, if $t_0 t \leq 0$, then one of the following four instances must hold: (1) $t = 0$; (2) $t_0 = 0$; (3) $t < 0 < t_0$; (4) $t_0 < 0 < t$. In all four of these instances, the integral will not exist since the function $f(t) = t^{-1}$ is not integrable about 0. Thus this differential equation only can be solved when t and t_0 are both on the same side of 0. •

3.2.2 Example (An ordinary differential equation with no solutions (all the time)) This example is beyond the abilities of a typical student taking a first course in differential equations, but we present it because it shows something interesting.

We let $f : \mathbb{R} \to \mathbb{R}$ be a function with the properties that (1) f takes values in $[0, 1]$ and (2) the integral of the restriction of f to any interval does not exist, cf. Example III-2.4.3 and Example III-2.6.8–2. Given such an f, we define a scalar nonautonomous ordinary differential equation with right-hand side $\widehat{F}(t, x) = f(t)$. As in Example 3.2.1, a solution of this differential equation is given by

$$x(t) = x(t_0) + \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau.$$

In this case, because no matter how we choose t and t_0, the integral of $f|[t_0, t]$ (or $f|[t, t_0]$ if $t < t_0$) does not exist, and so a solution cannot exist for any choice of t and t_0. •

3.2.3 Example (Solutions, when they exist, may not be continuously differentiable) We made the comment in Remark 3.1.4 that our definition of solution in Definition 3.1.2 is sometimes too strong. Here we examine this idea in a simple case, and then make some general observations about how to rectify this situation.

Let us define $f : [0, 1] \to \mathbb{R}$ by

$$f(t) = \begin{cases} 1, & t \in [0, \frac{1}{2}], \\ 0, & t \in (\frac{1}{2}, 1], \end{cases}$$

and then define a scalar nonautonomous ordinary differential equation with right-hand side $\widehat{F}(t,x) = f(t)$. Following Example 3.2.1, we believe that a solution of this differential equation is given by

$$x(t) = \int_0^t f(\tau)\,\mathrm{d}\tau = \begin{cases} t, & t \in [0, \frac{1}{2}], \\ \frac{1}{2}, & t \in (\frac{1}{2}, 1]. \end{cases}$$

There is a problem with this "solution," however, namely that it is not differentiable at $t = \frac{1}{2}$, and so the equation $\dot{x}(t) = \widehat{F}(t, x(t))$ does not strictly hold.

It does hold, however, almost everywhere. Thus perhaps we should alter our notion of solution to be that the equation is satisfied almost everywhere. This, however, is not enough since it does not preserve uniqueness since there are non-constant functions with almost everywhere zero derivative (Example I-3.2.27). A moments though indicates that the requirement of local absolute continuity is the right one for solutions of ordinary differential equations since it is for this class of functions that the Fundamental Theorem of Calculus holds. The reader can consider this a little more deeply in Example 3.2.1. •

3.2.4 Example (Uniqueness of solutions is not the right thing to ask for) Let us now let $f\colon \mathbb{R} \to \mathbb{R}$ be a continuous function, which implies that the integral of $f|[a,b]$ exists for any $a < b$. As in our preceding two examples, we consider a differential equation with right-hand side $F(t,x) = f(t)$. And, as with the preceding two examples, solutions to this differential equation satisfy

$$x(t) = x(t_0) + \int_{t_0}^t f(\tau)\,\mathrm{d}\tau.$$

In this case, the integral exists for any t_0 and t, and this shows that this differential equation has *many* solutions. But what we notice is that, once we fix an initial time t_0 and an initial value $x(t_0)$ at this time, then the solution does become unique. •

3.2.5 Example (Solutions, when they exist, may have a limited domain of definition) The next example we consider shows that, even for seemingly well-behaved right-hand sides, solutions to differential equations will not be defined for all time. We consider a scalar autonomous ordinary differential equation with right-hand side $\widehat{F}(t,x) = x^2$. Thus solutions satisfy the equation

$$\dot{x}(t) = x(t)^2.$$

This equation can be easily solved (we shall see how to solve a class of equations including this one in Section 4.1.1) to give

$$x(t) = \begin{cases} 0, & x(t_0) = 0, \\ \frac{x(t_0)}{x(t_0)(t_0 - t) + 1}, & x(t_0) \neq 0. \end{cases}$$

(Alternatively, one can just verify by substitution that this is a solution of the differential equation and satisfies "$x(t_0) = x(t_0)$.") Let us assume that $x(t_0) \neq 0$. One can see that the solution in this case is only defined for

$$x(t_0)(t_0 - t) + 1 \neq 0 \iff t \neq t_0 + \frac{1}{x(t_0)} \triangleq t_*.$$

From this we conclude the following about solutions:

1. if $x(t_0) > 0$, then $\lim_{t \downarrow -\infty} x(t) = 0$ and $\lim_{t \uparrow t_*} x(t) = \infty$;
2. if $x(t_0) < 0$, then $\lim_{t \downarrow t_*} x(t) = -\infty$ and $\lim_{t \uparrow \infty} x(t) = 0$.

The essential point is that although (1) solutions exist for any initial time t_0 and any initial value $x(t_0)$ at that time and (2) the differential equation is defined for all times (and indeed is independent of time), solutions with initial values different from 0 will not exist for all times. •

3.2.6 Example (Solutions may not be unique even when things seem nice) We consider the scalar autonomous differential equation with right-hand side $\widehat{F}(t, x) = x^{1/3}$. We will show that there are infinitely many solutions $t \mapsto x(t)$ satisfying the equation

$$\dot{x}(t) = x(t)^{1/3}$$

with $x(0) = 0$. One can use the techniques of Section 4.1.1 to obtain the solution $t \mapsto x_0(t)$ given by

$$x_0(t) = \left(\frac{2}{3}t\right)^{3/2}.$$

However, $x_1(t) = 0$ is also clearly a solution. Indeed, there is a family of solutions of the form

$$x(t) = \begin{cases} x_0(t + t_-), & t \in (-\infty, -t_-], \\ 0, & t \in (-t_-, t_+), \\ x_0(t - t_+), & t \in [t_+, \infty), \end{cases}$$

where $t_-, t_+ \in \mathbb{R}_{>0}$. •

From the preceding examples, we draw the following conclusions about the questions of existence and uniqueness of solutions to ordinary differential equations.

1. From Examples 3.2.1 and 3.2.2 we conclude that we must prescribe some conditions on the right-hand side of $\widehat{\boldsymbol{F}}$ of an ordinary differential equation if we are to expect solutions to exist. This is hardly a surprise, of course. However, just what are the right conditions is something that took smart people some time to figure out, cf. the proof of Theorem 3.2.8 below.
2. From Example 3.2.3 we see that it is natural to relax the requirement that solutions be continuously differentiable, and that local absolute continuity is actually a more natural requirement.

3. Example 3.2.4 shows that in the case when we have solutions, we will have lots of them, so ordinary differential equations should not be expected to have unique solutions. However, in the example we saw that perhaps the matter of uniqueness can be resolved by asking that the unknown x take on a prescribed value at a prescribed time t_0. This is altogether akin to constants of integration disappearing when fixed upper and lower limits for the integral are chosen.
4. Example 3.2.5 shows that, even when solutions exist for all initial times and values of the unknown at these times, and even when the differential equation is autonomous, it can arise that solutions only exist locally in time, i.e., solutions cannot be defined for all times. It turns out that this is just a fact of life when dealing with differential equations.
5. Finally, Example 3.2.6 shows that, even when the differential equation is autonomous with a continuous right-hand side, it can happen that multiple, indeed infinitely many, solutions pass through the same initial value for the unknown at the same time. This is a quite undesirable state of affairs, and can be hypothesised away easily by conditions that are nearly always met in practice.

With an excellent understanding of the context of the existence and uniqueness problem bestowed upon us by these motivational examples, we can now state precisely with the problem is, and provide some notation for stating the main theorem.

Let us first state precisely the problem for whose solutions we consider existence and uniqueness.

3.2.7 Definition (Initial value problem for ordinary differential equations) Let F be an ordinary differential equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^m.$$

Let $t_0 \in \mathbb{T}$ and $x_0 \in X$. A map $\xi\colon \mathbb{T}' \to X$ is a ***solution*** for F with ***initial value*** x_0 at t_0 if it satisfies the following conditions:

(i) $\mathbb{T}' \subseteq \mathbb{T}$ is an interval;

(ii) ξ is locally absolutely continuous;

(iii) $\xi(t) = x_0 + \displaystyle\int_{t_0}^{t} \widehat{F}(\tau, \xi(\tau))\,\mathrm{d}\tau$ for all $t \in \mathbb{T}'$;

(iv) $\xi(t_0) = x_0$.

In this case, we say that ξ is a solution to the ***initial value problem***

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x_0. \qquad \bullet$$

We note that the differential equation is not necessarily satisfied by a solution, just since ξ need not be continuously differentiable. However, since ξ is locally

absolutely continuous, we have

$$\xi(t) = x_0 + \int_{t_0}^{t} \widehat{F}(\tau, \xi(\tau))\,\mathrm{d}\tau \quad \iff \quad \dot{\xi}(t) = \widehat{F}(t, \xi(t)), \text{ a.e. } t \in \mathbb{T}', \quad \xi(t_0) = x_0.$$

3.2.1.2 Principal existence and uniqueness theorems for ordinary differential equations The following is the principal existence and uniqueness result for ordinary differential equations. In the statement of the result we make use of the property of Lipschitzness, for which we refer to Section II-1.10.8.

3.2.8 Theorem (Existence and uniqueness of solutions for ordinary differential equations) *Let $X \subseteq \mathbb{R}^m$ be open, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $\mathbf{F}$ be a first-order ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times X \to \mathbb{R}^m.$$

We have the following two statements.

(i) Existence for continuous ordinary differential equations. *Suppose that $\mathbf{F}$ satisfies the following conditions:*

(a) the map $t \mapsto \widehat{\mathbf{F}}(t, \mathbf{x})$ is measurable for each $\mathbf{x} \in X$;

(b) the map $\mathbf{x} \mapsto \widehat{\mathbf{F}}(t, \mathbf{x})$ is continuous for each $t \in \mathbb{T}$;

(c) for each $(t, \mathbf{x}) \in \mathbb{T} \times X$, there exists $r, \rho \in \mathbb{R}_{>0}$ and

$$g \in L^1([t - \rho, t + \rho]; \mathbb{R}_{\geq 0})$$

such that

$$\|\widehat{\mathbf{F}}(s, \mathbf{y})\| \leq g(s), \qquad (s, \mathbf{y}) \in ([t - \rho, t + \rho] \cap \mathbb{T}) \times B(r, \mathbf{x}).$$

Then, for each $(t_0, \mathbf{x}_0) \in \mathbb{T} \times X$, there exists a subinterval $\mathbb{T}' \subseteq \mathbb{T}$, relatively open in $\mathbb{T}$ and with $t_0 \in \operatorname{int}_{\mathbb{T}}(\mathbb{T}')$, and a solution $\xi\colon \mathbb{T}' \to X$ for $\mathbf{F}$ such that $\xi(t_0) = \mathbf{x}_0$.

(ii) Uniqueness for Lipschitz ordinary differential equations. *Suppose that $\mathbf{F}$ satisfies the following conditions:*

(a) the map $t \mapsto \widehat{\mathbf{F}}(t, \mathbf{x})$ is locally integrable for each $\mathbf{x} \in X$;

(b) the map $\mathbf{x} \mapsto \widehat{\mathbf{F}}(t, \mathbf{x})$ is locally Lipschitz for each $t \in \mathbb{T}$;

(c) for each $(t, \mathbf{x}) \in \mathbb{T} \times X$, there exist $r, \rho \in \mathbb{R}_{>0}$ and

$$g, L \in L^1([t - \rho, t + \rho]; \mathbb{R}_{\geq 0})$$

such that

$$\|\widehat{\mathbf{F}}(s, \mathbf{y})\| \leq g(s), \qquad (s, \mathbf{y}) \in ([t - \rho, t + \rho] \cap \mathbb{T}) \times B(r, \mathbf{x}), \tag{3.7}$$

and

$$\|\widehat{\mathbf{F}}(s, \mathbf{y}_1) - \widehat{\mathbf{F}}(s, \mathbf{y}_2)\| \leq L(s)\|\mathbf{y}_1 - \mathbf{y}_2\|, \qquad s \in ([t-\rho, t+\rho] \cap \mathbb{T}),\ \mathbf{y}_1, \mathbf{y}_2 \in B(r, \mathbf{x}). \tag{3.8}$$

Then, for each $(\mathrm{t}_0, \mathbf{x}_0) \in \mathbb{T} \times \mathrm{X}$, *there exists a subinterval* $\mathbb{T}' \subseteq \mathbb{T}$, *relatively open in* $\mathbb{T}$ *and with* $\mathrm{t}_0 \in \mathrm{int}_{\mathbb{T}}(\mathbb{T}')$, *and a solution* $\xi\colon \mathbb{T}' \to \mathrm{X}$ *for* $\mathbf{F}$ *such that* $\xi(\mathrm{t}_0) = \mathbf{x}_0$. *Moreover, if* $\mathbb{T}''$ *is another such interval and* $\eta\colon \mathbb{T}'' \to \mathrm{X}$ *is another such solution, then* $\eta(\mathrm{t}) = \xi(\mathrm{t})$ *for all* $\mathrm{t} \in \mathbb{T}'' \cap \mathbb{T}'$.

Proof (i) Let us first prove a lemma.

1 Lemma *For a continuous map* $\xi\colon \mathbb{T} \to \mathrm{X}$, *the function* $\mathrm{t} \mapsto \widehat{\mathbf{F}}(\mathrm{t}, \xi(\mathrm{t}))$ *is locally integrable.*

Proof First of all, let us show that $t \mapsto \widehat{F}(t, \xi(t))$ is measurable. It suffices to prove this when $\mathbb{T}$ is compact, so we make this assumption. Since ξ is continuous, by Theorem III-2.9.2 there exists a sequence $(\xi_j)_{j\in\mathbb{Z}_{>0}}$ of piecewise constant functions converging uniformly to ξ. That is, for each $j \in \mathbb{Z}_{>0}$ there exists a partition $(\mathbb{T}_{j,1}, \dots, \mathbb{T}_{j,k_j})$ of $\mathbb{T}$ such that $\xi_j(t) = \boldsymbol{x}_{j,l}$ for some $\boldsymbol{x}_{j,l} \in \mathbb{R}^m$ when $t \in \mathbb{T}_{j,l}$ for $l \in \{1, \dots, k_j\}$. Then

$$\widehat{F}(t\, \xi_j(t)) = \sum_{l=1}^{k_j} \widehat{F}(t, \boldsymbol{x}_{j,l}) \chi_{\mathbb{T}_{j,l}},$$

where χ_A denotes the characteristic function of a subset A of a set S, and so $t \mapsto \widehat{F}(t, \xi_j(t))$ is measurable. Now, by continuity of $\boldsymbol{x} \mapsto \widehat{F}(t, \boldsymbol{x})$,

$$\lim_{j\to\infty} \widehat{F}(t, \xi_j(t)) = \widehat{F}(t, \xi(t))$$

and measurability of $t \mapsto \widehat{F}(t, \xi(t))$ follows since the pointwise limit of measurable functions is measurable by Proposition III-2.6.18(v).

Now let $t, t_0 \in \mathbb{T}$ and suppose that $t > t_0$. Then, by continuity of ξ, there exists a compact set $K \subseteq X$ such that $\xi(s) \in K$ for every $s \in [t_0, t_0 + t]$. By assumption, there exists a locally integrable function $g\colon \mathbb{T} \to \mathbb{R}_{\geq 0}$ such that $\|\widehat{F}(s, \boldsymbol{x})\| \leq g(s)$ for every $(s, \boldsymbol{x}) \in \mathbb{T} \times K$. Therefore,

$$\int_{t_0}^{t} \|\widehat{F}(s, \xi(s))\| \, \mathrm{d}s \leq \int_{t_0}^{t} g(s) \, \mathrm{d}s < \infty.$$

The same statement holds if $t < t_0$, flipping the limits of integration, and this gives the desired local integrability. ▼

Let $r \in \mathbb{R}_{>0}$ be chosen so that $\overline{\mathsf{B}}(r, x_0) \subseteq X$ and let $\rho \in \mathbb{R}_{>0}$. By choosing r and ρ sufficiently small, there exists

$$g \in \mathsf{L}^1([t_0 - \rho, t_0 + \rho]; \mathbb{R}_{\geq 0})$$

such that

$$\|\widehat{F}(t, \boldsymbol{x})\| \leq g(t), \qquad (t, \boldsymbol{x}) \in ([t_0 - \rho, t_0 + \rho] \cap \mathbb{T}) \times \overline{\mathsf{B}}(r, \boldsymbol{x}_0).$$

Then, since g is integrable, the function $G_+\colon [t_0, t_0 + \rho] \cap \mathbb{T} \to \mathbb{R}$ defined by

$$G_+(t) = \int_{t_0}^{t} g(s) \, \mathrm{d}s \tag{3.9}$$

is continuous.

Let us suppose that $t_0 \neq \sup \mathbb{T}$ so that we can choose ρ so that $[t_0, t_0+\rho] \subseteq \mathbb{T}$. Thus, since g is nonnegative, there exists $T_+ \in \mathbb{R}_{>0}$ such that $[t_0, t_0+T_+] \subseteq \mathbb{T}$ and such that

$$G_+(t) = \int_{t_0}^{t} g(s)\,\mathrm{d}s < r, \qquad t \in [t_0, t_0+T_+].$$

For the remainder of the proof, we consider r and T_+ to be chosen as above.

Let $\mathsf{C}^0([t_0, t_0+T_+]; \mathbb{R}^m)$ be the Banach space of continuous $\mathbb{R}^m$-valued functions on $[t_0, t_0+T_+]$ equipped with the norm

$$\|\xi\|_\infty = \sup\{\|\xi(t)\| \mid t \in [t_0, t_0+T_+]\}$$

(see Definition III-3.8.28). Let $\xi_0 \in \mathsf{C}^0([t_0, t_0+T_+]; \mathbb{R}^m)$ be defined by $\xi_0(t) = x_0$. Let $\overline{\mathsf{B}}_+(r, \xi_0)$ be the closed ball of radius r and centre ξ_0 in $\mathsf{C}^0([t_0, t_0+T_+]; \mathbb{R}^m)$. For $\alpha \in (0, T_+]$, let us define $\xi_\alpha \in \overline{\mathsf{B}}_+(r, \xi_0)$ by

$$\xi_\alpha(t) = \begin{cases} x_0, & t \in [t_0, t_0+\alpha], \\ x_0 + \int_{t_0}^{t} \widehat{F}(s, \xi_\alpha(s-\alpha))\,\mathrm{d}s, & t \in (t_0+\alpha, t_0+T_+]. \end{cases}$$

It is not clear that this definition makes sense, so let us verify how it does. We fix $\alpha \in (0, T_+]$. If $t \in [t_0, t_0+\alpha]$, then the meaning of $\xi_\alpha(t)$ is unambiguous. If $t \in (t_0+\alpha, t_0+2\alpha] \cap [t_0, t_0+T_+]$, then $\xi_\alpha(t)$ is determined from the already known value of ξ_α on $[t_0, t_0+\alpha]$. Similarly, if $t \in (t_0+2\alpha, t_0+3\alpha] \cap [t_0, t_0+T_+]$, then $\xi_\alpha(t)$ is determined from the already known value of ξ_α on $[t_0, t_0+2\alpha]$. In a finite number of such steps, one determines ξ_α on $[t_0, t_0+T_+]$. We now show that $\xi_\alpha \in \overline{\mathsf{B}}_+(r, \xi_0)$. If $t \in [t_0, t_0+\alpha]$, then $\|\xi_\alpha(t) - x_0\| = 0$. If $t \in (t_0+\alpha, t_0+2\alpha]$, then

$$\begin{aligned} \|\xi_\alpha(t) - x_0\| &= \left\| \int_{t_0}^{t_0+\alpha} \mathbf{0}\,\mathrm{d}s + \int_{t_0+\alpha}^{t} \widehat{F}(s, x_0)\,\mathrm{d}s \right\| \\ &\leq \int_{t_0}^{t_0+\alpha} \mathbf{0}\,\mathrm{d}s + \int_{t_0+\alpha}^{t} \|\widehat{F}(s, x_0)\|\,\mathrm{d}s \leq \int_{t_0}^{t} g(s)\,\mathrm{d}s < r. \end{aligned}$$

By induction, if $t \in (t_0+(k-1)\alpha, t_0+k\alpha]$, then

$$\|\xi_\alpha(t) - x_0\| \leq \sum_{j=0}^{k-2} \int_{t_0+j\alpha}^{t_0+(j+1)\alpha} g(s)\,\mathrm{d}s + \int_{t_0+(k-1)\alpha}^{t} g(s)\,\mathrm{d}s \leq r,$$

giving $\xi_\alpha \in \overline{\mathsf{B}}_+(r, \xi_0)$, as desired.

We claim that the family $(\xi_\alpha)_{\alpha \in (0, T_+]}$ is equicontinuous, i.e., for each $\epsilon \in \mathbb{R}_{>0}$ there exists $\delta \in \mathbb{R}_{>0}$ such that

$$|t_1 - t_2| < \delta \quad \Longrightarrow \quad \|\xi_\alpha(t_1) - \xi_\alpha(t_2)\| < \epsilon$$

for all $\alpha \in (0, T_+]$. So let $\epsilon \in \mathbb{R}_{>0}$ and note that the function $G_+ \colon [t_0, t_0+T_+] \to \mathbb{R}$ defined by (3.9) is continuous, and so uniformly continuous, its domain being compact. Therefore, there exists $\delta \in \mathbb{R}_{>0}$ such that

$$|t_1 - t_2| < \delta \quad \Longrightarrow \quad |G_+(t_1) - G_+(t_2)| < \epsilon.$$

Let δ be so chosen. Then, if $|t_1 - t_2| < \delta$ with $t_1 > t_2$,

$$\begin{aligned}\|\xi_\alpha(t_1) - \xi_\alpha(t_2)\| &= \left\| \int_{t_0}^{t_1} \widehat{\boldsymbol{F}}(s, \xi_\alpha(t-\alpha))\,\mathrm{d}s - \int_{t_0}^{t_2} \widehat{\boldsymbol{F}}(s, \xi_\alpha(t-\alpha))\,\mathrm{d}s \right\| \\ &\leq \int_{t_2}^{t_1} \|\widehat{\boldsymbol{F}}(s, \xi_\alpha(t-\alpha))\|\,\mathrm{d}s \leq \int_{t_2}^{t_1} g(s)\,\mathrm{d}s = G_+(t_1) - G_+(t_2) < \epsilon,\end{aligned}$$

as desired.

Thus we have an equicontinuous family $(\xi_\alpha)_{\alpha\in(0,T_+]}$ contained in the bounded set $\overline{\mathsf{B}}_+(r, \xi_0)$. Consider then the sequence $(\xi_{T_+/j})_{j\in\mathbb{Z}_{>0}}$ contained in this family. By the Arzelà–Ascoli Theorem and the Bolzano–Weierstrass Theorem there exists ref an increasing sequence $(j_k)_{k\in\mathbb{Z}_{>0}}$ such that the sequence $(\xi_{T_+/j_k})_{k\in\mathbb{Z}_{>0}}$ converges in ref $\mathsf{C}^0([t_0, t_0+T_+]; \mathbb{R}^m)$, i.e., converges uniformly. Let us denote the limit by $\xi_+ \in \overline{\mathsf{B}}_+(r, \xi_0)$. It remains to show that the ξ_+ is a solution for $\boldsymbol{F}$ satisfying $\xi_+(t_0) = \boldsymbol{x}_0$. For this, an application of Theorem III-2.7.28, continuity of $\widehat{\boldsymbol{F}}$ in the second argument, and equicontinuity of $(\xi_\alpha)_{\alpha\in(0,T_+]}$ gives

$$\begin{aligned}\xi_+(t) &= \lim_{k\to\infty} \xi_{T_+/j_k}(t) = \boldsymbol{x}_0 + \lim_{j_k\to\infty} \int_{t_0}^{t} \widehat{\boldsymbol{F}}(s, \xi_{T_+/j_k}(s - T_+/j_k))\,\mathrm{d}s \\ &= \boldsymbol{x}_0 + \int_{t_0}^{t} \widehat{\boldsymbol{F}}(s, \lim_{\alpha\to 0} \xi_\alpha(s-\alpha))\,\mathrm{d}s = \boldsymbol{x}_0 + \int_{t_0}^{t} \widehat{\boldsymbol{F}}(s, \xi_+(s))\,\mathrm{d}s.\end{aligned}$$

Therefore, by the lemma above, ξ_+ is absolutely continuous and

$$\dot{\xi}_+(t) = \widehat{\boldsymbol{F}}(t, \xi_+(t))$$

for almost every $t \in [t_0, t_0 + T_+]$. Thus ξ_+ is a solution for $\boldsymbol{F}$. Obviously $\xi_+(t_0) = \boldsymbol{x}_0$.

Next suppose that $t_0 \neq \inf \mathbb{T}$. Then there exists $a \in \mathbb{R}_{>0}$ such that $[t_0 - a, t_0] \subseteq \mathbb{T}$. As above, we let $r \in \mathbb{R}_{>0}$ be such that $\overline{\mathsf{B}}(r, \boldsymbol{x}_0) \subseteq X$. Define $G_-\colon (-\infty, t_0] \cap \mathbb{T} \to \mathbb{R}$ by

$$G_-(t) = \int_t^{t_0} g(s)\,\mathrm{d}s$$

so that G_- is continuous. Since g is nonnegative, there exists $T_- \in \mathbb{R}_{>0}$ such that $[t_0, t_0 - T_-] \subseteq \mathbb{T}$ and such that

$$G_-(t) = \int_t^{t_0} g(s)\,\mathrm{d}s < r, \qquad t \in [t_0 - T_-, t_0].$$

Now, with r and T_- thusly defined, we can proceed as above to show the existence of a solution $\xi_-\colon [t_0 - T_-, t_0] \to X$ for $\boldsymbol{F}$ such that $\xi_-(t_0) = \boldsymbol{x}_0$.

The proof of this part of the theorem is complete if we define $\mathbb{T}'$ and ξ as follows.

1. $\operatorname{int}(\mathbb{T}) = \emptyset$: The interval $\mathbb{T}' = \{t_0\}$ and the trivial solution $\xi_0(t) = \boldsymbol{x}_0$ satisfies the conclusions of the theorem.
2. $t_0 \neq \sup \mathbb{T}$ and $t_0 = \inf \mathbb{T}$: The interval $\mathbb{T}' = [t_0, t_0 + T_+)$ and the solution $\xi = \xi_+$ as defined above satisfy the conclusions of the theorem.

3. $t_0 = \sup \mathbb{T}$ and $t_0 \neq \inf \mathbb{T}$: The interval $\mathbb{T}' = [t_0 - T_-, t_0)$ and the solution $\xi = \xi_-$ as defined above satisfy the conclusions of the theorem.
4. $t_0 \neq \sup \mathbb{T}$ and $t_0 \neq \inf \mathbb{T}$: The interval $\mathbb{T}' = (t_0 - T_-, t_0 + T_+)$ and the solution

$$\xi(t) = \begin{cases} \xi_-(t), & t \in (t_0 - T_-, t_0], \\ \xi_+(t), & t \in (t_0, t_0 + T_+] \end{cases}$$

satisfy the conclusions of the theorem.

(ii) Note that the existence statement follows from part (i) since the hypotheses of part (ii) imply those of part (i). However, we shall reprove this via an argument that also ensures uniqueness.

Let $r \in \mathbb{R}_{>0}$ be such that $\overline{\mathsf{B}}(r, x_0) \subseteq X$. Let $\rho \in \mathbb{R}_{>0}$. By choosing r and ρ sufficiently small, there exist

$$g, L \in \mathsf{L}^1([t_0 - \rho, t_0 + \rho]; \mathbb{R}_{\geq 0})$$

such that

$$\|\widehat{F}(t, x)\| \leq g(t), \qquad (t, x) \in ([t_0 - \rho, t_0 + \rho] \cap \mathbb{T}) \times \overline{\mathsf{B}}(r, x_0)$$

and

$$\|\widehat{F}(t, x) - \widehat{F}(t, y)\| \leq L(t)\|x - y\|$$

for all $t \in [t_0 - \rho, t_0 + \rho] \cap \mathbb{T}$ and $x, y \in \overline{\mathsf{B}}(r, x_0)$. Let us choose $\lambda \in (0, 1)$.

We first consider the case where $t_0 \neq \sup \mathbb{T}$ so that we can choose ρ sufficiently small that $[t_0, t_0 + \rho] \subseteq \mathbb{T}$. Define $G_+, \ell_+ \colon [t_0, t_0 + \rho] \to \mathbb{R}$ by

$$G_+(t) = \int_{t_0}^{t} g(s)\,\mathrm{d}s, \quad \ell_+(t) = \int_{t_0}^{t} L(s)\,\mathrm{d}s.$$

Since g and L are nonnegative, we can choose $T_+ \in \mathbb{R}_{>0}$ such that

$$G_+(t) = \int_{t_0}^{t} g(s)\,\mathrm{d}s \leq r, \quad \ell_+(t) = \int_{t_0}^{t} L(s)\,\mathrm{d}s < \lambda$$

for $t \in [t_0, t_0 + T_+]$.

As in the proof of part (i), let ξ_0 be the trivial function $t \mapsto x_0$, $t \in [t_0, t_0 + T_+]$, and let $\overline{\mathsf{B}}_+(r, \xi_0)$ be the ball of radius r and centre ξ_0 in $\mathsf{C}^0([t_0, t_0 + T_+]; \mathbb{R}^m)$. Define $F_+ \colon \overline{\mathsf{B}}_+(r, \xi_0) \to \mathsf{C}^0([t_0, t_0 + T_+]; \mathbb{R}^m)$ by

$$F_+(\xi)(t) = x_0 + \int_{t_0}^{t} \widehat{F}(s, \xi(s))\,\mathrm{d}s.$$

By the lemma from the proof of part (i), $s \mapsto \widehat{F}(s, \xi(s))$ is locally integrable, showing that F_+ is well-defined and that $F_+(\xi)$ is absolutely continuous.

We claim that $F_+(\overline{\mathsf{B}}_+(r, \xi_0)) \subseteq \overline{\mathsf{B}}_+(r, \xi_0)$. Suppose that $\xi \in \overline{\mathsf{B}}_+(r, \xi_0)$ so that

$$\|\xi(t) - x_0\| \leq r, \qquad t \in [t_0, t_0 + T_+].$$

Then, for $t \in [t_0, t_0 + T_+]$,

$$\|F_+(\xi)(t) - x_0\| = \left\| \int_{t_0}^t \widehat{\boldsymbol{F}}(s, \xi(s)) \, \mathrm{d}s \right\| \leq \int_{t_0}^t \|\widehat{\boldsymbol{F}}(s, \xi(s))\| \, \mathrm{d}s \leq \int_{t_0}^t g(s) \, \mathrm{d}s \leq r,$$

as desired.

We claim that $F_+|\overline{\mathsf{B}}_+(r, \xi_0)$ is a contraction mapping. That is, we claim that there exists $\rho \in [0, 1)$ such that

$$\|F_+(\xi) - F_+(\eta)\|_\infty \leq \rho \|\xi - \eta\|_\infty$$

for every $\xi, \eta \in \overline{\mathsf{B}}_+(r, \xi_0)$. Indeed, let $\xi, \eta \in \overline{\mathsf{B}}_+(r, \xi_0)$ and compute, for $t \in [t_0, t_0 + T_+]$,

$$\begin{aligned} \|F_+(\xi)(t) - F_+(\eta)(t)\| &= \left\| \int_{t_0}^t \widehat{\boldsymbol{F}}(s, \xi(s)) \, \mathrm{d}s - \int_{t_0}^t \widehat{\boldsymbol{F}}(s, \eta(s)) \right\| \\ &\leq \int_{t_0}^t \|\widehat{\boldsymbol{F}}(s, \xi(s)) - \widehat{\boldsymbol{F}}(s, \eta(s))\| \, \mathrm{d}s \\ &\leq \int_{t_0}^t L(s) \|\xi(s) - \eta(s)\| \, \mathrm{d}s \leq \ell_+(t) \|\xi - \eta\|_\infty \leq \lambda \|\xi - \eta\|_\infty, \end{aligned}$$

since $\xi(s), \eta(s) \in \mathsf{B}(r, x_0)$ for every $s \in [t_0, t_0 + T_+]$. This proves that $F_+|\overline{\mathsf{B}}_+(r, \xi_0)$ is a contraction mapping.

By the Contraction Mapping Theorem, Theorem III-1.1.23, there exists a unique fixed point for F_+ which we denote by ξ_+. Thus

$$\xi_+(t) = F_+(\xi_+)(t) = x_0 + \int_{t_0}^t \widehat{\boldsymbol{F}}(s, \xi_+(s)) \, \mathrm{d}s.$$

Differentiating the first and last expressions with respect to t shows that ξ_+ is a solution for $\boldsymbol{F}$.

Now we consider the case when $t_0 \neq \inf \mathbb{T}$ so there exists $a \in \mathbb{R}_{>0}$ such that $[t_0 - a, t_0] \subseteq \mathbb{T}$. We proceed as above, cf. the corresponding part of the proof of part (i), to provide $T_- \in \mathbb{R}_{>0}$ such that

$$G_-(t) \triangleq \int_t^{t_0} g(s) \, \mathrm{d}s < r, \quad \ell_-(t) \triangleq \int_t^{t_0} L(s) \, \mathrm{d}s < \lambda$$

for $t \in [t_0 - T_-, t_0]$. We then define $\overline{\mathsf{B}}_-(r, \xi_0)$ as the ball of radius r and centre ξ_0 in $\mathsf{C}^0([t_0 - T_-, t_0]; \mathbb{R}^m)$ and define $F_- \colon \overline{\mathsf{B}}_-(r, x_0) \to \mathsf{C}^0([t_0 - T_-, t_0]; \mathbb{R}^m)$ by

$$F_-(\xi)(t) = x_0 + \int_{t_0}^t \widehat{\boldsymbol{F}}(s, \xi(s)) \, \mathrm{d}s.$$

We show, as above, that $F_-(\overline{\mathsf{B}}_-(r, \xi_0)) \subseteq \overline{\mathsf{B}}_-(r, \xi_0)$ and that $F_-|\overline{\mathsf{B}}_-(r, \xi_0)$ is a contraction mapping, so possessing a unique fixed point ξ_-. This fixed point is a solution for $\boldsymbol{F}$, as above.

We can then define an interval $\mathbb{T}'$ and a solution ξ for F as at the end of the proof of part (i). We now prove uniqueness of this solution on $\mathbb{T}'$. Suppose that $\eta\colon \mathbb{T}' \to X$ is another solution satisfying $\eta(t_0) = x_0$. Then

$$\dot{\eta}(t) = \widehat{F}(t, \eta(t)), \qquad t \in \mathbb{T}'.$$

Therefore, by the Fundamental Theorem of Calculus,

$$\eta(t) = \eta(t_0) + \int_{t_0}^{t} \dot{\eta}(s)\,\mathrm{d}s = x_0 + \int_{t_0}^{t} \widehat{F}(s, \eta(s))$$

for $t \geq t_0$ and

$$\eta(t) = \eta(t_0) + \int_{t_0}^{t} \dot{\eta}(s)\,\mathrm{d}s = x_0 + \int_{t_0}^{t} \widehat{F}(s, \eta(s))$$

for $t \leq t_0$. It then follows that $\eta|[t_0, \infty) \cap \mathbb{T}'$ is a fixed point for F_+ and $\eta|(-\infty, t_0] \cap \mathbb{T}'$ is a fixed point for F_-. Therefore, η agrees with ξ on $\mathbb{T}'$ by the uniqueness part of the Contraction Mapping Theorem.

Now suppose that $\mathbb{T}'' \subseteq \mathbb{R}$ is some other interval containing t_0 and that $\eta\colon \mathbb{T}'' \to X$ is a solution for F satisfying $\eta(t_0) = x_0$. Suppose that $\xi(t) \neq \eta(t)$ for some $t \in \mathbb{T}'' \cap \mathbb{T}'$. Suppose that $t < t_0$. Let

$$t_1 = \inf\{t \in [t_0, \infty) \cap \mathbb{T}'' \cap \mathbb{T}' \mid \xi(t) \neq \eta(t)\}.$$

Then $\xi(t) = \eta(t)$ for $t \in [t_0, t_1)$. Continuity of solutions implies that $\xi(t_1) = \eta(t_1)$. Denote $x_1 = \xi(t_1)$. Note that both ξ and η are solutions for F satisfying $\xi(t_1) = \eta(t_1) = x_1$. By our above arguments for existence and uniqueness, there exists $T_+ \in \mathbb{R}_{>0}$ and a unique solution ζ on $[t_1, t_1 + T_+]$ satisfying $\zeta(t_1) = x_1$. Thus ξ and η must agree with ζ on $[t_1, t_1 + T_+]$ contradicting the definition of t_1. A similar argument leads to a similar contradiction when we assume that ξ and η disagree at some $t \in \mathbb{T}'' \cap \mathbb{T}'$ with $t < t_0$. ■

The matter of checking the conditions of Theorem 3.2.8 is normally quite straightforward, particularly since if we know that a function is differentiable, then it is locally Lipschitz. Indeed, let us encode in the following result a situation where the hypotheses of Theorem 3.2.8 are easily verified.

3.2.9 Corollary (An existence and uniqueness result that is easy to apply) *Let $X \subseteq \mathbb{R}^m$ be open, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $\mathbf{F}$ be a first-order ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times X \to \mathbb{R}^m.$$

If $\widehat{\mathbf{F}}$ is of class C^1 on $\mathbb{T} \times X$, then, for each $(t_0, x_0) \in \mathbb{T} \times X$, there exists a subinterval $\mathbb{T}' \subseteq \mathbb{T}$, relatively open in $\mathbb{T}$ and with $t_0 \in \mathrm{int}_{\mathbb{T}}(\mathbb{T}')$, and a solution $\xi\colon \mathbb{T}' \to X$ for $\mathbf{F}$ such that $\xi(t_0) = x_0$. Moreover, if $\mathbb{T}''$ is another such interval and $\eta\colon \mathbb{T}'' \to X$ is another such solution, then $\eta(t) = \xi(t)$ for all $t \in \mathbb{T}'' \cap \mathbb{T}'$.

We ask the reader to check that the hypotheses of Theorem 3.2.8 are satisfied for the examples of Section 1.1 as Exercise 3.2.4. In Exercise 3.2.5 we ask the reader to show which hypotheses of Theorem 3.2.8 are violated for the examples we gave at the beginning of this section.

3.2.1.3 Flows for ordinary differential equations With the above notions of existence and uniqueness of solutions for initial value problems, in this section we give some notation that ties together *all* solutions to *all* initial value problems. In doing this, we naturally run up against the question of how solutions to initial value problems depend on initial conditions. We shall at various points in the text run into situations where this sort of dependence is important, so the results in this section, while a bit technical, are certainly an essential part of any deep understanding of ordinary differential equations.

First we introduce the notation.

3.2.10 Definition (Interval of existence, domain of solutions) Let F be an ordinary differential equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^m,$$

and assume that F satisfies the conditions of Theorem 3.2.8(ii) for existence and uniqueness of solutions for initial value problems.

(i) For $(t_0, x_0) \in \mathbb{T} \times X$, denote

$$J_F(t_0, x_0) = \cup\{J \subseteq \mathbb{T} \mid \begin{array}{r} J \text{ is an interval and there is a solution} \\ \xi\colon J \to X \text{ for } F \text{ satisfying } \xi(t_0) = x_0\}. \end{array}$$

The interval $J_F(t_0, x_0)$ is the ***interval of existence*** for the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x_0.$$

(ii) The ***domain of solutions*** for F is

$$D_F = \{(t, t_0, x_0) \in \mathbb{T} \times \mathbb{T} \times X \mid t \in J_F(t_0, x_0)\}. \qquad \bullet$$

We shall carefully enumerate various properties of intervals of existence and domains of solutions, but to do this let us first introduce a very useful bit of notation.

3.2.11 Definition (Flow of an ordinary differential equation) Let F be an ordinary differential equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^m,$$

and assume that F satisfies the conditions of Theorem 3.2.8(ii) for existence and uniqueness of solutions for initial value problems. The ***flow*** of F is the map $\Phi^F\colon D_F \to X$ defined by asking that $\Phi^F(t, t_0, x_0)$ is the solution, evaluated at t, of the initial value problem

$$\dot{\xi}(\tau) = \widehat{F}(\tau, \xi(\tau)), \quad \xi(t_0) = x_0. \qquad \bullet$$

The definition, phrased differently, says that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi^F(t,t_0,x_0) = f(t,\Phi^F(t,t_0,x_0)), \quad \Phi^F(t_0,t_0,x_0) = x_0.$$

For $t, t_0 \in \mathbb{T}$, it is sometimes convenient to denote

$$D_F(t,t_0) = \{x \in X \mid (t,t_0,x) \in D_F\},$$

and then

$$\begin{aligned}\Phi^F_{t,t_0}\colon D_F(t,t_0) &\to X\\ x &\mapsto \Phi^F(t,t_0,x).\end{aligned}$$

Along similar lines, for $t_0 \in \mathbb{T}$, we denote

$$D_F(t_0) = \{(t,x) \in \mathbb{T}\times X \mid (t,t_0,x) \in D_F\},$$

and then

$$\begin{aligned}\Phi^F(t_0)\colon D_F(t_0) &\to X\\ (t,x) &\mapsto \Phi^F(t,t_0,x).\end{aligned}$$

Let us enumerate some of the more elementary properties of the flow.

3.2.12 Proposition (Elementary properties of flows of ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times X \to \mathbb{R}^m,$$

and assume that $\mathbf{F}$ *satisfies the conditions of Theorem 3.2.8(ii) for existence and uniqueness of solutions for initial value problems. Then the following statements hold:*

(i) for each $(t_0,x_0) \in \mathbb{T}\times X$, $(t_0,t_0,x_0) \in D_{\mathbf{F}}$ *and* $\Phi^{\mathbf{F}}(t_0,t_0,x_0) = x_0$*;*

(ii) if $(t_2,t_1,x) \in D_{\mathbf{F}}$, *then* $(t_3,t_2,\Phi^{\mathbf{F}}(t_2,t_1,x)) \in D_{\mathbf{F}}$ *if and only if* $(t_3,t_1,x) \in D_{\mathbf{F}}$ *and, if this holds, then*

$$\Phi^{\mathbf{F}}(t_3,t_1,x) = \Phi^{\mathbf{F}}(t_3,t_2,\Phi^{\mathbf{F}}(t_2,t_1,x)).$$

(iii) if $(t_2,t_1,x) \in D_{\mathbf{F}}$, *then* $(t_1,t_2,\Phi^{\mathbf{F}}(t_2,t_1,x)) \in D_{\mathbf{F}}$ *and* $\Phi^{\mathbf{F}}(t_1,t_2,\Phi^{\mathbf{F}}(t_2,t_1,x)) = x$.

Proof (i) This is part of the definition of the flow.

(ii) Suppose that $t_2 \geq t_1$ and $t_3 \geq t_2$.

First suppose that $(t_2,t_1,x) \in D_F$ and $(t_3,t_2,\Phi^F(t_2,t_1,x)) \in D_F$. We then have solutions $\xi_1\colon [t_1,t_2] \to X$ and $\xi_2\colon [t_2,t_3] \to X$ to the initial value problems

$$\dot\xi_1(t) = \widehat{F}(t,\xi_1(t)), \quad \xi_1(t_1) = x,$$

and

$$\dot\xi_2(t) = \widehat{F}(t,\xi_2(t)), \quad \xi_2(t_2) = \Phi^F(t_2,t_1,x),$$

respectively. Then define $\xi\colon [t_1, t_3] \to X$ by

$$\xi(t) = \begin{cases} \xi_1(t), & t \in [t_1, t_2], \\ \xi_2(t), & t \in [t_2, t_3]. \end{cases}$$

It is clear, then, that

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_1) = x.$$

It is then also clear that

$$\xi(t_3) = \Phi^F(t_3, t_2, \Phi^F(t_2, t_1, x))$$

and that $\xi(t_3) = \Phi^F(t_3, t_1, x)$. This gives $(t_3, t_1, x) \in D_F$.

Now suppose that $(t_2, t_1, x) \in D_F$ and $(t_3, t_1, x) \in D_F$. Let $\xi_1\colon [t_1, t_2] \to X$ and $\xi_3\colon [t_1, t_3] \to X$ be the solutions to the initial value problems

$$\dot{\xi}_1(t) = \widehat{F}(t, \xi_1(t)), \quad \xi_1(t_1) = x,$$

and

$$\dot{\xi}_3(t) = \widehat{F}(t, \xi_3(t)), \quad \xi_3(t_1) = x,$$

respectively. Then, by uniqueness of solutions, the curve $\xi_2\colon [t_2, t_3] \to X$ give by

$$\xi_2(t) = \xi_1(t) = \xi_3(t)$$

satisfies the initial value problem

$$\dot{\xi}_2(t) = \widehat{F}(t, \xi_2(t)), \quad \xi_2(t_2) = \xi_1(t_2) = \Phi^F(t_2, t_1, x),$$

and so $(t_3, t_2, \Phi^F(t_2, t_1, x)) \in D_F$.

The assertion that

$$\Phi^F(t_3, t_1, x) = \Phi^F(t_3, t_2, \Phi^F(t_2, t_1, x))$$

follows from uniqueness of solutions.

In the cases that (1) $t_1 \geq t_2$ and $t_3 \leq t_2$, (2) $t_2 \leq t_1$ and $t_3 \geq t_2$, and (3) $t_3 \leq t_2$ and $t_2 \leq t_1$, similarly styled arguments can be made, appropriately fussing with going in "different directions" in cases (1) and (2).

(iii) This is a special case of (ii), using (i). ∎

Useful mnemonics associated with parts (i)–(iii) are:

$$\Phi^F_{t_0,t_0} = \mathrm{id}_X, \quad (\Phi^F_{t_2,t_1})^{-1} = \Phi^F_{t_1,t_2}, \quad \Phi^F_{t_3,t_2} \circ \Phi^F_{t_2,t_1} = \Phi^F_{t_3,t_1}.$$

However, these really are just mnemonics, since they do not account carefully for the domains of the mappings being used.

The following result encodes some less elementary properties of the flow of an ordinary differential equation, including the regularity of the dependence on time and state.

3.2.13 Theorem (Properties of flows of ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times X \to \mathbb{R}^m,$$

and assume that $\mathbf{F}$ *satisfies the conditions of Theorem 3.2.8(ii) for existence and uniqueness of solutions for initial value problems. Then the following statements hold:*

(i) *for* $(t_0, \mathbf{x}_0) \in \mathbb{T} \times X$, $J_{\mathbf{F}}(t_0, \mathbf{x}_0)$ *is an interval that is a relatively open subset of* $\mathbb{T}$*;*

(ii) *for* $(t_0, \mathbf{x}_0) \in \mathbb{T} \times X$, *the curve*

$$\begin{aligned} \gamma_{(t_0,\mathbf{x}_0)}\colon J_{\mathbf{F}}(t_0, \mathbf{x}_0) &\to X \\ t &\mapsto \Phi^{\mathbf{F}}(t, t_0, \mathbf{x}_0) \end{aligned}$$

is well-defined and absolutely continuous;

(iii) *for* $t, t_0 \in \mathbb{T}$, $D_{\mathbf{F}}(t, t_0)$ *is open in* X*;*

(iv) *for* $t, t_0 \in \mathbb{T}$ *for which* $D_{\mathbf{F}}(t, t_0) \neq \emptyset$, $\Phi^{\mathbf{F}}_{t,t_0}$ *is a locally bi-Lipschitz homeomorphism onto its image;*

(v) *for* $t_0 \in \mathbb{T}$, $D_{\mathbf{F}}(t_0)$ *is relatively open in* $\mathbb{T} \times X$*;*

(vi) *for* $t_0 \in \mathbb{T}$, *the map*

$$\begin{aligned} \Phi^{\mathbf{F}}(t_0)\colon D_{\mathbf{F}}(t_0) &\to X \\ (t, \mathbf{x}) &\mapsto \Phi^{\mathbf{F}}(t, t_0, \mathbf{x}) \end{aligned}$$

is well-defined and continuous;

(vii) $D_{\mathbf{F}}$ *is relatively open in* $\mathbb{T} \times \mathbb{T} \times X$*;*

(viii) *the map*

$$\Phi^{\mathbf{F}}\colon D_{\mathbf{F}} \to X$$

is continuous;

(ix) *for* $(t_0, \mathbf{x}_0) \in \mathbb{T} \times X$ *and for* $\epsilon \in \mathbb{R}_{>0}$, *there exists* $r, \alpha \in \mathbb{R}_{>0}$ *such that*

$$\sup J_{\mathbf{F}}(t, \mathbf{x}) > \sup J_{\mathbf{F}}(t_0, \mathbf{x}_0) - \epsilon, \qquad \inf J_{\mathbf{F}}(t, \mathbf{x}) < \inf J_{\mathbf{F}}(t_0, \mathbf{x}_0) + \epsilon$$

for all $(t, \mathbf{x}) \in (t_0 - \alpha, t_0 + \alpha) \cap \mathbb{T} \times \mathsf{B}(r, \mathbf{x}_0)$.

Proof (i) Since $J_F(t_0, x_0)$ is a union of intervals, each of which contains t_0, it follows that it is itself an interval. To show that it is an open subset of $\mathbb{T}$, we show that, if $t \in J_F(t_0, x_0)$, there exists $\epsilon \in \mathbb{R}_{>0}$ such that

$$(-\epsilon, \epsilon) \cap \mathbb{T} \subseteq J_F(t_0, x_0).$$

First let us consider the case when t is not an endpoint of $\mathbb{T}$, in the event that $\mathbb{T}$ contains one or both of its endpoints. In this case, by definition of $J_F(t_0, x_0)$, there is an open interval $J \subseteq \mathbb{T}$ containing t_0 and t, and a solution $\xi\colon J \to X$ of the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x_0.$$

In particular, there exists $\epsilon \in \mathbb{R}_{>0}$ such that $(-\epsilon, \epsilon) \subseteq J \subseteq J_F(t_0, x_0)$, which gives the desired conclusion in this case.

Next suppose that t is the right endpoint of $\mathbb{T}$, which we assume is contained in $\mathbb{T}$, of course. In this case, by definition of $J_F(t_0, x_0)$, there is an interval $J \subseteq \mathbb{T}$ containing t_0 and t, and a solution $\xi\colon J \to X$ of the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x_0.$$

In this case, there exists $\epsilon \in \mathbb{R}_{>0}$ such that

$$(-\epsilon, \epsilon) \cap \mathbb{T} = (-\epsilon, t] \subseteq J_F(t_0, x_0),$$

which gives the desired conclusion in this case.

A similar argument gives the desired conclusion when t is the left endpoint of $\mathbb{T}$.

(ii) That $\gamma_{(t_0,x_0)}$ is defined in $J_F(t_0, x_0)$ was proved as part of the preceding part of the proof. The assertion that $\gamma_{(t_0,x_0)}$ is locally absolutely continuous follows from Theorem 3.2.8(ii).

We shall prove the assertions (iii)–(vi) of the theorem together, first by proving that these conditions hold locally, and then giving an extension argument to give the global version of the statement.

Let us first prove a few technical lemmata that will be useful to us.

1 Lemma *Let $\mathbb{T}$ be an interval, and let $\alpha, \beta, \xi\colon \mathbb{T} \to \mathbb{R}$, and $t_0 \in \mathbb{T}$ be such that*

(i) *α is continuous,*

(ii) *β is nonnegative-valued and locally integrable,*

(iii) *ξ is nonnegative-valued and continuous, and*

(iv) $\xi(t) \leq \alpha(t) + \int_{t_0}^{t} \beta(s)\xi(s)\,ds$ *for all* $t \in \mathbb{T} \cap [t_0, \infty)$.

Then

$$\xi(t) \leq \alpha(t) + \int_{t_0}^{t} \alpha(s)\beta(s)e^{\int_s^t \beta(\tau)\,d\tau}\,ds, \qquad t \in \mathbb{T} \cap [t_0, \infty).$$

Moreover, if α is additionally nondecreasing, then we have

$$\xi(t) \leq \alpha(t)e^{\int_{t_0}^{t} \beta(s)\,ds}, \qquad t \in \mathbb{T} \cap [t_0, \infty).$$

Proof Define

$$\eta(s) = e^{-\int_{t_0}^{s} \beta(\tau)\,d\tau} \int_{t_0}^{s} \beta(\tau)\xi(\tau)\,d\tau$$

and calculate, for almost every $s \in [t_0, t]$,

$$\begin{aligned}
\frac{d\eta}{ds}(s) &= -\beta(s)e^{-\int_{t_0}^{s} \beta(\tau)\,d\tau} \int_{t_0}^{s} \beta(\tau)\xi(\tau)\,d\tau + \beta(s)\xi(s)e^{-\int_{t_0}^{s} \beta(\tau)\,d\tau} \\
&= \beta(s)e^{-\int_{t_0}^{s} \beta(\tau)\,d\tau} \left(\xi(s) - \int_{t_0}^{s} \beta(\tau)\xi(\tau)\,d\tau\right) \\
&\leq \alpha(s)\beta(s)e^{-\int_{t_0}^{s} \beta(\tau)\,d\tau},
\end{aligned}$$

using the hypotheses of the lemma. Therefore,

$$\eta(t) \leq \int_{t_0}^{t} \alpha(s)\beta(s)\mathrm{e}^{-\int_{t_0}^{s}\beta(\tau)\,\mathrm{d}\tau}\mathrm{d}s.$$

Using the definition of η we then have

$$\begin{aligned}\int_{t_0}^{t} \beta(s)\xi(s)\,\mathrm{d}s &\leq \int_{t_0}^{t} \alpha(s)\beta(s)\mathrm{e}^{\int_{t_0}^{t}\beta(s)\,\mathrm{d}s}\mathrm{e}^{-\int_{t_0}^{s}\beta(\tau)\,\mathrm{d}\tau}\,\mathrm{d}s \\ &= \int_{t_0}^{t} \alpha(s)\beta(s)\mathrm{e}^{\int_{s}^{t}\beta(\tau)\,\mathrm{d}\tau}\mathrm{d}s,\end{aligned}$$

which immediately gives the first conclusion of the lemma.

For the second, we first note that, for almost every $s \in [t_0, t]$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathrm{e}^{\int_{s}^{t}\beta(\tau)\,\mathrm{d}\tau} = -\beta(s)\mathrm{e}^{\int_{s}^{t}\beta(\tau)\,\mathrm{d}\tau}.$$

Then

$$\int_{t_0}^{t} \beta(s)\mathrm{e}^{\int_{s}^{t}\beta(\tau)\,\mathrm{d}\tau}\mathrm{d}s = -\mathrm{e}^{\int_{s}^{t}\beta(\tau)\,\mathrm{d}\tau}\Big|_{s=t_0}^{s=t} = \mathrm{e}^{\int_{t_0}^{t}\beta(\tau)\,\mathrm{d}\tau} - 1.$$

Then we use the first part of the lemma and the additional assumption on α:

$$\begin{aligned}\xi(t) &\leq \alpha(t) + \int_{t_0}^{t} \alpha(s)\beta(s)\mathrm{e}^{\int_{s}^{t}\beta(\tau)\,\mathrm{d}\tau}\mathrm{d}s \\ &\leq \alpha(t) + \alpha(t)\left(\int_{t_0}^{t} \mathrm{e}^{\beta(s)\mathrm{d}s} - 1\right),\end{aligned}$$

and the lemma follows. ▼

Now we give the initial part of the local version of the theorem.

2 Lemma *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{X} \to \mathbb{R}^{\mathrm{m}},$$

and assume that $\mathbf{F}$ *satisfies the conditions of Theorem 3.2.8(ii) for existence and uniqueness of solutions for initial value problems. Then, for each* $(\mathrm{t}_0, \mathbf{x}_0) \in \mathbb{T} \times \mathrm{X}$*, there exists* $\mathrm{r}, \alpha \in \mathbb{R}_{>0}$ *such that* $(\mathrm{t}, \mathrm{t}_0, \mathbf{x}) \in \mathrm{D}_{\mathbf{F}}$ *for each* $\mathbf{x} \in \mathsf{B}(\mathrm{r}, \mathbf{x}_0)$ *and* $\mathrm{t} \in (\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha) \cap \mathbb{T}$*. Moreover,*

(i) the map

$$\mathsf{B}(\mathrm{r}, \mathbf{x}_0) \ni \mathbf{x} \mapsto \Phi^{\mathbf{F}}_{\mathrm{t},\mathrm{t}_0}(\mathbf{x}) \in \mathbb{R}^{\mathrm{m}}$$

is Lipschitz for every $\mathrm{t} \in (\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha) \cap \mathbb{T}$*;*

(ii) the map

$$(\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha) \cap \mathbb{T} \times \mathsf{B}(\mathrm{r}, \mathbf{x}_0) \ni (\mathrm{t}, \mathbf{x}) \mapsto \Phi^{\mathbf{F}}(\mathrm{t}, \mathrm{t}_0, \mathbf{x})$$

is continuous.

Proof First let $r' \in \mathbb{R}_{>0}$ be such that $\mathsf{B}(r', x_0) \subseteq X$ and let $r = \frac{r'}{2}$. As in the proof of Theorem 3.2.8(ii), there exist locally integrable $g, L \colon \mathbb{T} \to \mathbb{R}_{\geq 0}$ such that

$$\|\widehat{F}(t,x)\| \leq g(t), \qquad (t,x) \in \mathbb{T} \times \overline{\mathsf{B}}(r', x_0).$$

and

$$\|\widehat{F}(t,x_1) - \widehat{F}(t,x_2)\| \leq L(t)\|x_1 - x_2\|$$

for all $t \in \mathbb{T}$ and $x_1, x_2 \in \overline{\mathsf{B}}(r', x_0)$. Let us choose $\lambda \in (0,1)$. As in the proof of Theorem 3.2.8(ii), there exists $\alpha \in \mathbb{R}_{>0}$ such that

$$\left| \int_{t_0}^{t} g(s)\,\mathrm{d}s \right| \leq r, \quad \left| \int_{t_0}^{t} L(s)\,\mathrm{d}s \right| < \lambda, \qquad t \in [t_0 - \alpha, t_0 + \alpha].$$

If $x \in \mathsf{B}(r, x_0)$, then $\mathsf{B}(r, x) \subseteq \mathsf{B}(r', x_0)$. Therefore,

$$\|\widehat{F}(t,y)\| \leq g(t), \qquad (t,y) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x).$$

and

$$\|\widehat{F}(t,y_1) - \widehat{F}(t,y_2)\| \leq L(t)\|y_1 - y_2\|$$

for all $t \in \mathbb{T}$ and $y_1, y_2 \in \overline{\mathsf{B}}(r, x)$. If $\xi_1 \in \mathsf{C}^0([t_0 - \alpha, t_0 + \alpha]; \mathbb{R}^m)$ is the constant function $\xi_0(t) = x_0$, then the arguments from the proof of Theorem 3.2.8(ii) allow us to conclude that there is a solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

in $\overline{\mathsf{B}}(r, \xi_0) \subseteq \mathsf{C}^0([t_0 - \alpha, t_0 + \alpha]; \mathbb{R}^m)$. This is the existence assertion of the lemma.

(i) Let $x_1, x_2 \in \mathsf{B}(r, x_0)$ and let $t \in [t_0 - \alpha, t_0 + \alpha]$. Then

$$\Phi^F(t, t_0, x_1) = x_1 + \int_{t_0}^{t} \widehat{F}(s, \Phi^F(s, t_0, x_1))\,\mathrm{d}s, \quad \Phi^F(t, t_0, x_2) = x_2 + \int_{t_0}^{t} \widehat{F}(s, \Phi^F(s, t_0, x_2))\,\mathrm{d}s,$$

for all $t \in [t_0 - \alpha, t_0 + \alpha]$. Therefore,

$$\begin{aligned}
\|\Phi^F(t, t_0, x_1) - \Phi^F(t, t_0, x_2)\| &\leq \|x_1 - x_2\| + \int_{t_0}^{t} \|\widehat{F}(s, \Phi^F(s, t_0, x_1)) - \widehat{F}(s, \Phi^F(s, t_0, x_2))\|\,\mathrm{d}s \\
&\leq \|x_1 - x_2\| + \int_{t_0}^{t} L(s)\|\Phi^F(s, t_0, x_1) - \Phi^F(s, t_0, x_2)\|\,\mathrm{d}s \\
&\leq \|x_1 - x_2\| \mathrm{e}^{\int_{t_0}^{t} L(s)\,\mathrm{d}s} \leq \|x_1 - x_2\| \mathrm{e}^{\lambda}.
\end{aligned}$$

This shows that $\Phi^F_{t,t_0}|\mathsf{B}(r, x_0)$ is Lipschitz, as claimed, when $t \geq t_0$. A similar computation gives the analogous conclusion when $t \leq t_0$.

(ii) Let $t_1, t_2 \in (t_0 - \alpha, t_0 + \alpha) \cap \mathbb{T}$ be such that $t_1 \leq t_2$. Just as above, we have

$$\begin{aligned}
\|\Phi^F(t_1, t_0, x_1) - \Phi^F(t_2, t_0, x)\| &\leq \|x_1 - x_2\| \\
&\quad + \int_{t_0}^{t_1} \|\widehat{F}(s, \Phi^F(s, t_0, x_1)) - \widehat{F}(s, \Phi^F(s, t_0, x_2))\|\,\mathrm{d}s + \int_{t_1}^{t_2} \|\widehat{F}(s, t_0, \Phi^F(s, t_0, x_2))\|\,\mathrm{d}s.
\end{aligned}$$

Let $\epsilon \in \mathbb{R}_{>0}$. By Lemma 1 from the proof of Theorem 3.2.8, there exists $\delta_1 \in \mathbb{R}_{>0}$ sufficiently small that, if $|t_2 - t_1| < \delta_1$, then

$$\int_{t_1}^{t_2} \|\widehat{\boldsymbol{F}}(s, t_0, \Phi^{\boldsymbol{F}}(s, t_0, \boldsymbol{x}_2))\| \, \mathrm{d}s < \frac{\epsilon}{2}.$$

Since $\Phi^{\boldsymbol{F}}_{t_1,t_0}$ is continuous, let $\delta_2 \in \mathbb{R}_{>0}$ be sufficiently small that, if $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| < \delta_1$, then

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \int_{t_0}^{t_1} \|\widehat{\boldsymbol{F}}(s, \Phi^{\boldsymbol{F}}(s, t_0, \boldsymbol{x}_1)) - \widehat{\boldsymbol{F}}(s, \Phi^{\boldsymbol{F}}(s, t_0, \boldsymbol{x}_2))\| \, \mathrm{d}s < \frac{\epsilon}{2}.$$

Then, if $|t_1 - t_2| < \delta_1$ and $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| < \delta_2$,

$$\|\Phi^{\boldsymbol{F}}(t_1, t_0, \boldsymbol{x}_1) - \Phi^{\boldsymbol{F}}(t_2, t_0, \boldsymbol{x}_2)\| < \epsilon,$$

giving the desired conclusion. ▼

The next lemma is a refinement of the preceding one, giving the local version of the theorem statement.

3 Lemma *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times \mathrm{X} \to \mathbb{R}^{\mathrm{m}},$$

and assume that $\mathbf{F}$ *satisfies the conditions of Theorem 3.2.8(ii) for existence and uniqueness of solutions for initial value problems. Then, for each* $(\mathrm{t}_0, \mathbf{x}_0) \in \mathbb{T} \times \mathrm{X}$, *there exists* $\mathrm{r}, \alpha \in \mathbb{R}_{>0}$ *such that*

(i) $(\mathrm{t}, \mathrm{t}_0, \mathbf{x}) \in \mathrm{D}_{\mathbf{F}}$ *for each* $\mathbf{x} \in \mathsf{B}(\mathrm{r}, \mathbf{x}_0)$ *and* $\mathrm{t} \in (\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha) \cap \mathbb{T}$,

(ii) the map

$$(\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha) \times \mathsf{B}(\mathrm{r}, \mathbf{x}_0) \ni (\mathrm{t}, \mathbf{x}) \mapsto \Phi^{\mathbf{F}}(\mathrm{t}, \mathrm{t}_0, \mathbf{x})$$

is continuous, and

(iii) the map

$$\mathsf{B}(\mathrm{r}, \mathbf{x}_0) \ni \mathbf{x} \mapsto \Phi^{\mathbf{F}}_{\mathrm{t},\mathrm{t}_0}(\mathbf{x}) \in \mathbb{R}^{\mathrm{m}}$$

is a bi-Lipschitz homeomorphism onto its image for every $\mathrm{t} \in (\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha) \cap \mathbb{T}$.

Proof Let r', α' be as in Lemma 2 and let $r \in (0, r']$ and $\alpha \in (0, \alpha']$ be such that

$$\Phi^{\boldsymbol{F}}_{t,t_0}(\boldsymbol{x}) \in \mathsf{B}(r', \boldsymbol{x}_0), \qquad \boldsymbol{x} \in \mathsf{B}(r, \boldsymbol{x}_0), \ t \in [t_0 - \alpha, t_0 + \alpha],$$

this being possible by Lemma 2(ii). Let $t \in (t_0 - \alpha, t_0 + \alpha) \cap \mathbb{T}$ and denote

$$V = \Phi^{\boldsymbol{F}}_{t,t_0}(\mathsf{B}(r, \boldsymbol{x}_0)) \subseteq \mathsf{B}(r', \boldsymbol{x}_0).$$

Let $\boldsymbol{x} \in \mathsf{B}(r, \boldsymbol{x}_0)$. Since $\boldsymbol{y} \triangleq \Phi^{\boldsymbol{F}}_{t,t_0}(\boldsymbol{x}) \in \mathsf{B}(r', \boldsymbol{x}_0)$ and $t \in [t_0 - \alpha', t_0 + \alpha'] \cap \mathbb{T}$, there exists $\rho \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{y}' \in \mathsf{B}(\rho, \boldsymbol{y})$, then $(t_0, t, \boldsymbol{y}') \in D_{\boldsymbol{F}}$. Moreover, since $\Phi^{\boldsymbol{F}}_{t_0,t}$ is continuous (indeed, Lipschitz, with Lipschitz constant e^{λ}, with λ as in the proof of Lemma 2) and $\Phi^{\boldsymbol{F}}_{t_0,t}(\boldsymbol{y}) = \boldsymbol{x}$, we may choose ρ sufficiently small that $\Phi^{\boldsymbol{F}}_{t_0,t}(\boldsymbol{y}') \in \mathsf{B}(r, \boldsymbol{x}_0)$ if $\boldsymbol{y}' \in \mathsf{B}(\rho, \boldsymbol{y})$. By Lemma 2, $\Phi^{\boldsymbol{F}}_{t_0,t}|\mathsf{B}(\rho, \boldsymbol{y})$ is Lipschitz with Lipschitz constant e^{λ}. Thus there is a neighbourhood of $\boldsymbol{x}$ on which the restriction of $\Phi^{\boldsymbol{F}}_t t_0$ is invertible, Lipschitz, and with a Lipschitz inverse. ▼

We now need to show that the theorem holds globally. To this end, let $(t_0, x_0) \in \mathbb{T} \times X$ and denote by $J_+(t_0, x_0) \subseteq \mathbb{T}$ the set of $b > t_0$ such that, for each $b' \in [t_0, b)$, there exists a relatively open interval $J \subseteq \mathbb{T}$ and a $r \in \mathbb{R}_{>0}$ such that

1. $b' \in J$,
2. $J \times \{t_0\} \times \mathsf{B}(r, x_0) \subseteq D_F$,
3. $J \times \mathsf{B}(r, x_0) \ni (t, x) \mapsto \Phi^F(t, t_0, x) \in X$ is continuous, and
4. for each $t \in J$, $\mathsf{B}(r, x_0) \ni x \mapsto \Phi^F(t, t_0, x)$ is a locally bi-Lipschitz homeomorphism onto its image.

By Lemma 3, $J_+(t_0, x_0) \neq \emptyset$. We then consider two cases.

The first case is $J_+(t_0, x_0) \cap [t_0, \infty) = \mathbb{T} \cap [t_0, \infty)$. In this case, for each $t \in \mathbb{T}$ with $t \geq t_0$, there exists a relatively open interval $J \subseteq \mathbb{T}$ and $r \in \mathbb{R}_{>0}$ such that

1. $t \in J$,
2. $J \times \{t_0\} \times \mathsf{B}(r, x_0) \subseteq D_F$,
3. $J \times \mathsf{B}(r, x_0) \ni (\tau, x) \mapsto \Phi^F(\tau, t_0, x) \in X$ is continuous, and
4. for each $\tau \in J$, $\mathsf{B}(r, x_0) \ni x \mapsto \Phi^F(\tau, t_0, x)$ is a locally bi-Lipschitz homeomorphism onto its image.

The second case is $J_+(t_0, x_0) \cap [t_0, \infty) \subset \mathbb{T} \cap [t_0, \infty)$. In this case we let $t_1 = \sup J_+(t_0, x_0)$ and note that $t_1 \neq \sup \mathbb{T}$. We claim that $t_1 \in J_F(t_0, x_0)$. Were this not the case, then we must have $b \triangleq \sup J_F(t_0, x_0) < t_1$. Since $b \in J_+(t_0, x_0)$, there must be a relatively open interval $J \subseteq \mathbb{T}$ containing b such that $t \in J_F(t_0, x_0)$ for all $t \in J$. But, since there are t's in J larger than b, this contradicts the definition of $J_F(t_0, x_0)$, and so we conclude that $t_1 \in J_F(t_0, x_0)$. Let us denote $x_1 = \Phi^F(t_1, t_0, x_0)$. By Lemma 3, there exists $\alpha_1, r_1 \in \mathbb{R}_{>0}$ such that $(t, t_1, x) \in D_F$ for every $t \in (t_1 - \alpha_1, t_1 + \alpha_1)$ and $x \in \mathsf{B}(r_1, x_1)$, and such that the map

$$(t_1 - \alpha_1, t_1 + \alpha_1) \times \mathsf{B}(r_1, x_1) \ni (t, x) \mapsto \Phi^F(t, t_1, x)$$

is continuous, and the map

$$\mathsf{B}(r_1, x_1) \ni x \mapsto \Phi^F(t, t_1, x)$$

is a locally bi-Lipschitz homeomorphism onto its image for every $t \in (t_1 - \alpha_1, t_1 + \alpha_1)$. Since $t \mapsto \Phi^F(t, t_0, x_0)$ is continuous and $\Phi^F(t_1, t_0, x_0) = x_1$, let $\delta \in \mathbb{R}_{>0}$ be such that $\delta < \frac{\alpha_1}{2}$ and $\Phi^F(t, t_0, x_0) \in \mathsf{B}(r_1/4, x_1)$ for $t \in (t_1 - \delta, t_1)$. Now let $\tau_1 \in (t_1 - \delta, t_1)$ and, by our hypotheses on t_1, there exists an open interval J and $r_1' \in \mathbb{R}_{>0}$ such that

1. $\tau_1 \in J$,
2. $J \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F$,
3. $J \times \mathsf{B}(r_1', x_0) \ni (\tau, x) \mapsto \Phi^F(\tau, t_0, x) \in X$ is continuous, and
4. for each $\tau \in J$, $\mathsf{B}(r_1', x_0) \ni x \mapsto \Phi^F(\tau, t_0, x)$ is a locally bi-Lipschitz homeomorphism onto its image.

We also choose J and r_1' sufficiently small that

$$\{\Phi^F(t, t_0, x) \mid t \in J,\ x \in \mathsf{B}(r_1', x_0)\} \subseteq \mathsf{B}(r_1/2, x_1).$$

Now we claim that

$$(\tau_1 - \alpha_1, \tau_1 + \alpha_1) \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F.$$

We first show that

$$[\tau_1, \tau_1 + \alpha_1) \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F. \tag{3.10}$$

Indeed, we have $(\tau_1, t_0, x) \in D_F$ for every $x \in \mathsf{B}(r_1', x_0)$ since $\tau_1 \in J$. By definition of J, $\Phi^F(\tau_1, t_0, x) \in \mathsf{B}(r_1/2, x_1)$. By definition of τ_1, $t_1 - \tau_1 < \delta < \frac{\alpha_1}{2}$. Then, by definition of α_1 and r_1,

$$(t_1, \tau_1, \Phi^F(\tau_1, t_0, x)) \in D_F$$

for every $x \in \mathsf{B}(r_1', x_0)$. From this we conclude that $(t_1, t_0, x) \in D_F$ for every $x \in \mathsf{B}(r_1', x_0)$. Now, since

$$t \in [\tau_1, \tau_1 + \alpha_1) \quad \Longrightarrow \quad t \in (t_1 - \alpha_1, t_1 + \alpha_1),$$

we have $(t, t_1, \Phi^F(t, t_1, x)) \in D_F$ for every $t \in (\tau_1 - \alpha_1, \tau_1 + \alpha_1)$ and $x \in \mathsf{B}(r_1', x_0)$. Since

$$\Phi^F(t, t_1, \Phi^F(t_1, t_0, x)) = \Phi^F(t, t_0, x),$$

we conclude (3.10). A similar but less complicated argument gives

$$(\tau_1 - \alpha_1, \tau_1) \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F.$$

Now we claim that the map

$$(\tau_1 - \alpha_1, \tau_1 + \alpha_1) \times \mathsf{B}(r_1', x_0) \ni (t, x) \mapsto \Phi^F(t, t_0, x)$$

is continuous. This map is continuous at

$$(t, x) \in (\tau_1 - \alpha_1, \tau_1] \times \mathsf{B}(r_1', x_0)$$

by definition of τ_1. For $t \in (\tau_1, \tau_1 + \alpha_1)$ we have

$$\Phi^F(t, t_0, x) = \Phi^F(t, \tau_1, \Phi^F(\tau_1, t_0, x)),$$

and continuity in this case follows since compositions of continuous maps are continuous.

Next we claim that the map

$$\mathsf{B}(r_1', x_0) \ni x \mapsto \Phi^F(t, t_0, x)$$

is a locally bi-Lipschitz homeomorphism onto its image for every $t \in (\tau_1 - \alpha_1, \tau_1 + \alpha_1)$. By definition of τ_1, the map

$$\Phi^F_{t,t_0} \colon \mathsf{B}(r_1', x_0) \to \mathsf{B}(r_1/2, x_1)$$

is a locally bi-Lipschitz homeomorphism onto its image for $t \in (\tau_1 - \alpha_1, \tau_1]$. We also have that

$$\Phi^F_{t,\tau_1} \colon \mathsf{B}(r_1, x_1) \to X$$

is a locally bi-Lipschitz homeomorphism onto its image for $t \in (\tau_1, \tau_1 + \alpha_1)$. Since the composition of locally bi-Lipschitz homeomorphisms onto their image is a locally bi-Lipschitz homeomorphisms onto its image, our assertion follows.

By our above arguments, we have an open interval J' and $r_1' \in \mathbb{R}_{>0}$ such that

1. $t_1 \in J'$,
2. $J' \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F$,
3. $J' \times \mathsf{B}(r_1', x_0) \ni (t, x) \mapsto \Phi^F(t, t_0, x) \in X$ is continuous, and
4. for each $t \in J'$, $\mathsf{B}(r_1', x_0) \ni x \mapsto \Phi^F(t, t_0, x)$ is a locally bi-Lipschitz homeomorphism onto its image.

This contradicts the fact that $t_1 = \sup J_+(t_0, x_0)$ and so the condition

$$J_+(t_0, x_0) \cap [t_0, \infty) \subset \mathbb{T} \cap [t_0, \infty)$$

cannot obtain.

One similarly shows that it must be the case that $J_-(t_0, x_0) \cap (-\infty, t_0] = \mathbb{T} \cap (-\infty, t_0]$¡ where $J_-(t_0, x_0)$ has the obvious definition.

Finally, we note that Φ^F_{t,t_0} injective by uniqueness of solutions for $\boldsymbol{F}$. Now, assertions (i)–(vi) of the theorem now follow since the notions of "continuous" and "locally bi-Lipschitz homeomorphism" can be tested locally, i.e., in a neighbourhood of any point.

We shall prove assertions (vii) and (viii) together. We let $(t_1, t_0, x_0) \in D_F$. As above, there exists $r_1, \alpha_1 \in \mathbb{R}_{>0}$ such that

$$(t_1 - \alpha_1, t_1 + \alpha_1) \cap \mathbb{T} \times \{t_0\} \times \mathsf{B}(r_1, x_0) \subseteq D_F,$$

and the map $(t, x) \mapsto \Phi^F(t, t_0, x_0)$ is continuous on this domain. We claim that the map

$$(t, x) \mapsto \Phi^F(t_0, t, x) \tag{3.11}$$

is continuous for (t, x) nearby (t_0, x_0). To see this, we proceed rather as in the proof of Theorem 3.2.8, using the Contraction Mapping Theorem.

Let $r \in \mathbb{R}_{>0}$ be such that there exists a locally integrable $g \colon \mathbb{T} \to \mathbb{R}_{\geq 0}$ such that

$$\|\widehat{\boldsymbol{F}}(t, x)\| \leq g(t), \qquad (t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0),$$

and also there exists a locally integrable $L \colon \mathbb{T} \to \mathbb{R}_{\geq 0}$ such that

$$\|\widehat{\boldsymbol{F}}(t, x) - \widehat{\boldsymbol{F}}(t, y)\| \leq L(t)\|x - y\|$$

for all $t \in \mathbb{T}$ and $x, y \in \overline{\mathsf{B}}(r, x_0)$. Let us choose $\lambda \in (0, 1)$. Let us suppose that $t \leq t_0$. Define $G_-, \ell_- \colon (-\infty, t_0] \cap \mathbb{T} \to \mathbb{R}$ by

$$G_-(t) = \int_t^{t_0} g(s) \, \mathrm{d}s, \quad \ell_+(t) = \int_t^{t_0} L(s) \, \mathrm{d}s.$$

Since g and L are nonnegative, we can choose $T_- \in \mathbb{R}_{>0}$ such that

$$G_-(t) = \int_t^{t_0} g(s) \, \mathrm{d}s \leq \frac{r}{2}, \quad \ell_-(t) = \int_t^{t_0} L(s) \, \mathrm{d}s < \lambda$$

for $t \in [t_0 - T_-, t_0]$. For $x \in \mathsf{B}(r/2, x_0)$, let ξ_0 be the trivial function $t \mapsto x$, $t \in [t_0 - T_-, t_0]$, and let $\overline{\mathsf{B}}_-(r, \xi_0)$ be the ball of radius r and centre ξ_0 in $\mathsf{C}^0([t_0 - T_-, t_0]; \mathbb{R}^m)$. Define $F_- \colon \overline{\mathsf{B}}_-(r, \xi_0) \to \mathsf{C}^0([t_0 - T_-, t_0]; \mathbb{R}^m)$ by

$$F_-(\xi)(t) = x + \int_t^{t_0} \widehat{\boldsymbol{F}}(s, \xi(s)) \, \mathrm{d}s.$$

By the lemma from the proof of Theorem 3.2.8, $s \mapsto \widehat{\boldsymbol{F}}(s, \xi(s))$ is locally integrable, showing that F_- is well-defined and that $F_-(\xi)$ is absolutely continuous.

We claim that $F_-(\overline{\mathsf{B}}_-(r, \xi_0)) \subseteq \overline{\mathsf{B}}_-(r, \xi_0)$. Suppose that $\xi \in \overline{\mathsf{B}}_-(r, \xi_0)$ so that

$$\|\xi(t) - \boldsymbol{x}_0\| \leq r, \qquad t \in [t_0 - T_-, t_0].$$

Then, for $t \in [t_0 - T_-, t_0]$,

$$\begin{aligned}\|F_-(\xi)(t) - \boldsymbol{x}_0\| &\leq \|\boldsymbol{x} - \boldsymbol{x}_0\| + \left\| \int_t^{t_0} \widehat{\boldsymbol{F}}(s, \xi(s)) \, \mathrm{d}s \right\| \\ &\leq \frac{r}{2} + \int_t^{t_0} \|\widehat{\boldsymbol{F}}(s, \xi(s))\| \, \mathrm{d}s \leq \frac{r}{2} + \int_t^{t_0} g(s) \, \mathrm{d}s \leq r,\end{aligned}$$

as desired.

We claim that $F_-|\overline{\mathsf{B}}_-(r, \xi_0)$ is a contraction mapping. That is, we claim that there exists $\rho \in [0, 1)$ such that

$$\|F_-(\xi) - F_-(\eta)\|_\infty \leq \rho \|\xi - \eta\|_\infty$$

for every $\xi, \eta \in \overline{\mathsf{B}}_-(r, \xi_0)$. Indeed, let $\xi, \eta \in \overline{\mathsf{B}}_-(r, \xi_0)$ and compute, for $t \in [t_0 - T_-, t_0]$,

$$\begin{aligned}\|F_-(\xi)(t) - F_-(\eta)(t)\| &= \left\| \int_t^{t_0} \widehat{\boldsymbol{F}}(s, \xi(s)) \, \mathrm{d}s - \int_t^{t_0} \widehat{\boldsymbol{F}}(s, \eta(s)) \right\| \\ &\leq \int_t^{t_0} \|\widehat{\boldsymbol{F}}(s, \xi(s)) - \widehat{\boldsymbol{F}}(s, \eta(s))\| \, \mathrm{d}s \\ &\leq \int_t^{t_0} L(s) \|\xi(s) - \eta(s)\| \, \mathrm{d}s \leq \ell_-(t) \|\xi - \eta\|_\infty \leq \lambda \|\xi - \eta\|_\infty,\end{aligned}$$

since $\xi(s), \eta(s) \in \mathsf{B}(r, \boldsymbol{x}_0)$ for every $s \in [t_0, t_0 + T_+]$. This proves that $F_-|\overline{\mathsf{B}}_-(r, \xi_0)$ is a contraction mapping.

By the Contraction Mapping Theorem, Theorem III-1.1.23 there exists a unique fixed point for F_- which we denote by ξ_-. Thus

$$\xi_-(t) = F_-(\xi_+)(t) = \boldsymbol{x} + \int_t^{t_0} \widehat{\boldsymbol{F}}(s, \xi_-(s)) \, \mathrm{d}s.$$

Differentiating the first and last expressions with respect to t shows that ξ_+ is a solution for $\boldsymbol{F}$, and we moreover have $\xi(t_0) = \boldsymbol{x}$. This show that, if $\boldsymbol{x} \in \mathsf{B}(r/2, \boldsymbol{x}_0)$ and $t \in [t_0 - T_-, t_0]$, then we have $\Phi^{\boldsymbol{F}}(t_0, t, \boldsymbol{x}) \in \mathsf{B}(r, \boldsymbol{x}_0)$ and

$$\Phi^{\boldsymbol{F}}(t_0, t, \boldsymbol{x}) = \boldsymbol{x} + \int_t^{t_0} \widehat{\boldsymbol{F}}(s, \Phi^{\boldsymbol{F}}(t_0, s, \boldsymbol{x}) \, \mathrm{d}s.$$

A similar argument, of course, can be fabricated for $t \geq t_0$, and we conclude that there exists $\alpha_0 \in \mathbb{R}_{>0}$ and $r_0 \in \mathbb{R}_{>0}$ such that

$$\Phi^{\boldsymbol{F}}(t_0, t, \boldsymbol{x}) \in \mathsf{B}(r_1, \boldsymbol{x}_0), \qquad (t, \boldsymbol{x}) \in (t_0 - \alpha_0, t_0 + \alpha_0) \cap \mathbb{T} \times \mathsf{B}(r_0, \boldsymbol{x}_0).$$

Finally, we show that the map (3.11) is continuous on $(t_0-\alpha_0, t_0+\alpha_0)\cap\mathbb{T}\times\mathsf{B}(r_0,\boldsymbol{x}_0)$. Note that, as in the proof of Lemma 2 above and assuming that $\tau_2 \geq \tau_1$,

$$\begin{aligned}\|\Phi^{\boldsymbol{F}}(t_0,\tau_1,\boldsymbol{x}_1) - \Phi^{\boldsymbol{F}}(t_0,\tau_2,\boldsymbol{x}_2)\| &\leq \|\boldsymbol{x}_1-\boldsymbol{x}_2\| \\ &+ \int_{\tau_2}^{t_0} \|\widehat{\boldsymbol{F}}(s,\Phi^{\boldsymbol{F}}(t_0,s,\boldsymbol{x}_1)) - \widehat{\boldsymbol{F}}(s,\Phi^{\boldsymbol{F}}(t_0,s,\boldsymbol{x}_2))\|\,\mathrm{d}s + \int_{\tau_1}^{\tau_2}\|\widehat{\boldsymbol{F}}(s,\Phi^{\boldsymbol{F}}(t_0,s,\boldsymbol{x}_1))\|\,\mathrm{d}s.\end{aligned}$$

Let $\epsilon \in \mathbb{R}_{>0}$. By Lemma 1 from the proof of Theorem 3.2.8, there exists $\delta_1 \in \mathbb{R}_{>0}$ sufficiently small that, if $|\tau_2-\tau_1|<\delta_1$, then

$$\int_{\tau_1}^{\tau_2}\|\widehat{\boldsymbol{F}}(s,\Phi^{\boldsymbol{F}}(t_0,s,\boldsymbol{x}_2))\|\,\mathrm{d}s < \frac{\epsilon}{2}.$$

Since $\Phi^{\boldsymbol{F}}_{t_0,\tau_2}$ is continuous, let $\delta_2 \in \mathbb{R}_{>0}$ be sufficiently small that, if $\|\boldsymbol{x}_1-\boldsymbol{x}_2\|<\delta_2$, then

$$\|\boldsymbol{x}_1-\boldsymbol{x}_2\| + \int_{\tau_2}^{t_0}\|\widehat{\boldsymbol{F}}(s,\Phi^{\boldsymbol{F}}(t_0,s,\boldsymbol{x}_1)) - \widehat{\boldsymbol{F}}(s,\Phi^{\boldsymbol{F}}(t_0,s,\boldsymbol{x}_2))\|\,\mathrm{d}s < \frac{\epsilon}{2}.$$

Then, if $|t_1-t_2|<\delta_1$ and $\|\boldsymbol{x}_1-\boldsymbol{x}_2\|<\delta_2$,

$$\|\Phi^{\boldsymbol{F}}(t_0,\tau_1,\boldsymbol{x}_1) - \Phi^{\boldsymbol{F}}(t_0,\tau_2,\boldsymbol{x}_2)\| < \epsilon,$$

given the desired continuity.

Finally, if $(t',t_0',\boldsymbol{x}) \in (t-\alpha,t+\alpha)\cap\mathbb{T}\times(t_0-\alpha_0,t_0+\alpha_0)\cap\mathbb{T}\times\mathsf{B}(r_0,\boldsymbol{x}_0)$, then

$$\Phi^{\boldsymbol{F}}(t',t_0,\Phi^{\boldsymbol{F}}(t_0,t_0',\boldsymbol{x})) = \Phi^{\boldsymbol{F}}(t',t_0',\boldsymbol{x}),$$

which shows both that $D_{\boldsymbol{F}}$ is open and that $\Phi^{\boldsymbol{F}}$ is continuous, since the composition of continuous mappings is continuous.

(ix) Let $T_+ = \sup J_{\boldsymbol{F}}(t_0,\boldsymbol{x}_0)$. Then $(T_+-\epsilon,t_0,\boldsymbol{x}_0)\in D_{\boldsymbol{F}}$. Since $D_{\boldsymbol{F}}$ is open, there exists $r\in\mathbb{R}_{>0}$ such that

$$\{T_+-\tfrac{\epsilon}{2}\}\times(t_0-\alpha,t_0+\alpha)\cap\mathbb{T}\times\mathsf{B}(r,\boldsymbol{x}_0)\subseteq D_{\boldsymbol{F}}.$$

In other words, $[t_0,T_+-\frac{\epsilon}{2}]\subseteq J_{\boldsymbol{F}}(t,\boldsymbol{x})$ for every $(t,\boldsymbol{x})\in(t_0-\alpha,t_0+\alpha)\cap\mathbb{T}\times\mathsf{B}(r,\boldsymbol{x}_0)$. Thus, for such $(t,\boldsymbol{x})$,

$$\sup J_{\boldsymbol{F}}(t,\boldsymbol{x}) \geq T_- - \tfrac{\epsilon}{2} > T_- - \epsilon = \sup J_{\boldsymbol{F}}(t_0,\boldsymbol{x}_0) - \epsilon,$$

as claimed. A similar argument holds for the left endpoint of intervals of convergence. ∎

3.2.2 (Lack of) results for partial differential equations

The questions of existence and uniqueness of solutions for partial differential equations is far more difficult than for ordinary differential equations. Situations range from relatively simple cases where one can prove existence and uniqueness directly by writing down solutions, to equations where proving an existence and uniqueness result becomes a triumph of analysis, resulting in a paper in the *Annals*

of Mathematics. Thus it is not possible to have a comprehensive discussion of a theory of existence and uniqueness for general partial differential equations. Instead we content ourselves with some mostly vague observations about the nature of the problem.

First we note that all of the examples of Section 3.2.1 can be turned into partial differential equations is an entirely artificial way, merely by artificially adding an extra independent variable. This is not an interesting thing to do, except that it ensures that all of the conclusions 1–5 enumerated after these examples equally apply to partial differential equations.

Let us list some of the difficulties that arise in arriving at existence and uniqueness theorems for partial differential equations.

1. For ordinary differential equations, we saw that appropriate combinations of continuity, boundedness, and Lipschitz hypotheses ensured existence, and often uniqueness, of solutions. For partial differential equations, this is no longer true. A partial differential equation with lots of nice properties can fail to have any solutions. Moreover, this failure of solutions to exist can arise in various ways. So any attempt at a general theorem is dead from the start, and one must make assumptions on the sort of partial differential equation for solutions to even exist, cf. the discussion of elliptic, hyperbolic, and parabolic equations in Section 3.1.4.
2. For ordinary differential equations, we saw that to uniquely prescribe a solution one must specify an initial value of the state at some time to arrive at an initial value problem. For partial differential equations, this process is more difficult. Typically one must specify values of the solution along some surface or some such thing. This is known as prescribing "Cauchy data." However, the type of Cauchy data that is to be specified is not as easy a matter to understand as for ordinary differential equations. For many problems arising from physics, e.g., the heat, wave, and potential equations, the "natural" prescriptions of values for the solution and/or its derivatives at "boundaries" of the domain is often correct. However, these partial differential equations are "nice." In general, finding the analogue of initial conditions for ordinary differential equations is quite hard for partial differential equations.
3. The properties of a solution of an ordinary differential equation as it depends on the independent variable are quite easy: it is of class C^1 (or, more generally, locally absolutely continuous). For partial differential equations, finding the right attributes for a solution beforehand is often crucial to proving existence and uniqueness theorems for an equation.

We shall say nothing more about the subject of existence and uniqueness theorems for partial differential equations, except to say this:

Go to

http://www.claymath.org/millennium-problems/navier-stokes-equation

to win $1,000,000! •

Exercises

3.2.1 For a function $f\colon [0,1] \to \mathbb{R}$, consider an ordinary differential equation with right-hand side

$$\begin{aligned} \widehat{F}\colon [0,1] \times \mathbb{R} &\to \mathbb{R} \\ (t,x) &\mapsto f(t). \end{aligned}$$

Answer the following questions.

(a) Suppose that f is almost everywhere differentiable with zero derivative almost everywhere, but that f is not constant (Example I-3.2.27). Show that there exist at least two solutions to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(0) = 0.$$

(b) Show that, if f is absolutely continuous, then $\xi\colon [0,1] \to \mathbb{R}$ satisfies the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(0) = 0$$

if and only if

$$\xi(t) = \int_0^t f(\tau)\, \mathrm{d}\tau.$$

3.2.2 Which of the following maps are locally Lipschitz?

(a) $f\colon \mathbb{R} \to \mathbb{R}$
$x \mapsto \sqrt{|x|}$;

(b) $f\colon \mathbb{R}_{>0} \to \mathbb{R}$
$x \mapsto \sqrt{|x|}$;

(c) $f\colon \mathbb{R} \to \mathbb{R}$
$x \mapsto |x|$;

(d) $f\colon [0,\pi] \to \mathbb{R}$
$x \mapsto \sin(x)$;

(e) $f\colon \mathbb{R}_{>0} \to \mathbb{R}$
$x \mapsto x^{-1}$.

3.2.3 For the ordinary differential equations F with right-hand sides $\widehat{F}$ as given, determine which, if either, of the parts of Theorem 3.2.8 apply, and indicate what conclusions, if any, you can make about existence and uniqueness of solutions for F. Here are the right-hand sides:

(a) $\widehat{F}\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
$(t,x) \mapsto \sqrt{|t|}x$;

(b) $\widehat{F}\colon \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}$
$(t, x) \mapsto \dfrac{x}{t}$;

(c) $\widehat{F}\colon \mathbb{R} \times [0, 1] \to \mathbb{R}$
$(t, x) \mapsto \begin{cases} 1, & x \in [0, \frac{1}{2}], \\ -1, & x \in (\frac{1}{2}, 1]; \end{cases}$

(d) $\widehat{F}\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
$(t, x) \mapsto |xt|$;

(e) $\widehat{F}\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
$(t, x) \mapsto x^2$.

3.2.4 For the ordinary differential equations of Examples 3.1.3–1 to 9, show that the hypotheses of Theorem 3.2.8 hold, and so these equations possess unique solutions, at least for small times around any initial time.

3.2.5 In each of Examples 3.2.1–3.2.6, state the hypotheses of Theorem 3.2.8 that are violated by the example.

3.2.6 Let $\boldsymbol{F}$ be an ordinary differential equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^m$$

and suppose that, for each $\boldsymbol{x}_0 \in X$, there exist $M, r \in \mathbb{R}_{>0}$ such that

$$\left| \frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{x}) \right| \leq M, \qquad j, k \in \{1, \dots, n\},\ (t, \boldsymbol{x}) \in \mathbb{T} \times \mathsf{B}(r, \boldsymbol{x}_0).$$

Show that

$$\frac{\partial}{\partial t} \Phi^{\boldsymbol{F}}(t_0, t, \boldsymbol{x}) + \sum_{j=1}^{n} \widehat{F}_j(t, \boldsymbol{x}) \frac{\partial}{\partial x_j} \Phi^{\boldsymbol{F}}(t_0, t, \boldsymbol{x}) = 0.$$

3.2.7 We consider a partial differential equation

$$\boldsymbol{F}\colon \mathbb{R}^3 \times \mathbb{R} \times \mathrm{L}(\mathbb{R}^3; \mathbb{R}) \to \mathbb{R}^3$$

defined by

$$\boldsymbol{F}((x_1, x_2, x_3), u, (u_1, u_2, u_3)) = (u_1 - f_1(\boldsymbol{x}), u_2 - f_2(\boldsymbol{x}), u_3 - f_3(\boldsymbol{x})),$$

for a given continuously differentiable mapping $\boldsymbol{f}\colon \mathbb{R}^3 \to \mathbb{R}^3$. Show that, if $\boldsymbol{F}$ possesses a solution u of class C^2, then $\nabla \times \boldsymbol{f} = \boldsymbol{0}$.

Section 3.3

Classification of difference equations

Next we conduct the classification exercise, conducted in the preceding section for differential equations, to difference equations. As with differential equations, the objective here is to clarify what a difference equation *is* and what a solution *is*. We also describe various important classes of difference equations.

Do I need to read this section? The language we present in this section will be often used below.

3.3.1 Variables in difference equations

We first describe the sort of domains that are used for difference equations. Let $n \in \mathbb{Z}_{>0}$ and let $h_1, \dots, h_n \in \mathbb{R}_{>0}$. We then denote $\boldsymbol{h} = (h_1, \dots, h_n)$ and

$$\mathbb{Z}^n(\boldsymbol{h}) = \mathbb{Z}(h_1) \times \cdots \times \mathbb{Z}(h_n).$$

Thus $\mathbb{Z}^n(\boldsymbol{h})$ is a lattice in $\mathbb{R}^n$ with lattice gaps depending on the component, as depicted in Figure 3.5. With this notation, we make the following definition.

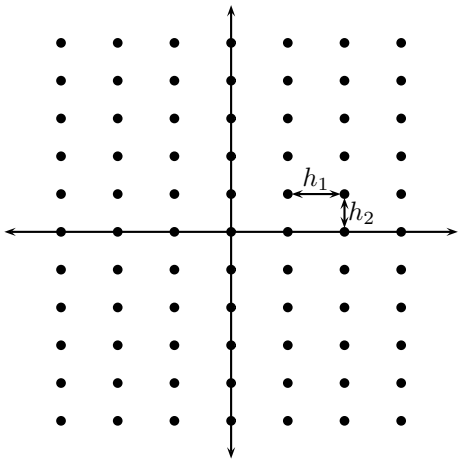


Figure 3.5 A lattice in $\mathbb{R}^2$

3.3.1 Definition (Discrete domain) A *discrete domain* in $\mathbb{R}^n$ is a subset of the form $D = \mathbb{Z}^n(\boldsymbol{h}) \cap U$, where $U \subseteq \mathbb{R}^n$ is an open set and for some $\boldsymbol{h} \in \mathbb{R}^n_{>0}$. •

Note that this amounts to saying that a discrete domain is a subset of $\mathbb{Z}^n(\boldsymbol{h})$. We phrase the definition as we do to emphasise the idea that one might think of a discrete domain as a discretisation of a domain for a differential equation.

The notions of "derivative" and "order" of a differential equation are clear. For difference equations, we use the following notions.

3.3.2 Definition (Forward and backward differences) Let $D \subseteq \mathbb{Z}^n(\boldsymbol{h})$ and let $\boldsymbol{f} \colon D \to \mathbb{R}^m$. For $j \in \{1, \dots, n\}$, denote

$$D_j^+ = \{\boldsymbol{x} \in D \mid \boldsymbol{x} + h_j \boldsymbol{e}_j \in D\}$$

and

$$D_j^- = \{\boldsymbol{x} \in D \mid \boldsymbol{x} - h_j \boldsymbol{e}_j \in D\}.$$

(i) The ***j*th *forward difference*** for $\boldsymbol{f}$ is

$$\begin{aligned} \Delta_j^+ \boldsymbol{f} \colon D_j^+ &\to \mathbb{R}^m \\ \boldsymbol{x} &\mapsto \frac{\boldsymbol{f}(\boldsymbol{x} + h_j \boldsymbol{e}_j) - \boldsymbol{f}(\boldsymbol{x})}{h_j}. \end{aligned}$$

(ii) The ***j*th *backward difference*** for $\boldsymbol{f}$ is

$$\begin{aligned} \Delta_j^- \boldsymbol{f} \colon D_j^- &\to \mathbb{R}^m \\ \boldsymbol{x} &\mapsto \frac{\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x} - h_j \boldsymbol{e}_j)}{h_j}. \end{aligned}$$

•

The idea of a forward and backward difference is that it is the discrete analogue of the derivative. When thinking of time derivatives, it is natural, for reasons of causality, to work with backward differences. However, for spatial derivatives, forward differences are also useful. Just like one constructs higher-order derivatives by iterating differentiation, one can iterate forward and backward differences. Let us set up the notation for doing this.

Let $\boldsymbol{\alpha} \in \{+, -\}^k$ be a sequence of k entries of pluses and minuses. Write

$$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k).$$

We shall use the following awkward, but sensible, notation, for $\alpha \in \{+, -\}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$:

$$\boldsymbol{x} \alpha \boldsymbol{y} = \begin{cases} \boldsymbol{x} + \boldsymbol{y}, & \alpha = +, \\ \boldsymbol{x} - \boldsymbol{y}, & \alpha = -. \end{cases}$$

This notation can be extended to any situation where objects can be added. Now $D \subseteq \mathbb{Z}^n(\boldsymbol{h})$ be a discrete domain, let $\boldsymbol{\alpha} \in \{+, -\}^k$, and let $\boldsymbol{j} \in \{1, \dots, n\}^k$. For $l \in \{1, \dots, k\}$, let $I(k, l)$ denote the set of sublists of length l of the list $(1, \dots, k)$. Thus, if $k = 4$, then

$$\begin{aligned} I(4,1) &= \{(1), (2), (3), (4)\}, \\ I(4,2) &= \{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}, \\ I(4,3) &= \{(1,2,3), (1,2,4), (1,3,4), (2,3,4)\}, \\ I(4,4) &= \{(1,2,3,4)\}. \end{aligned}$$

Then, denote

$$D_j^{\boldsymbol{\alpha}} = \left\{ \boldsymbol{x} \in D \;\middle|\; \sum_{r=1}^{l} \boldsymbol{x} \alpha_{i_r} h_{j_{i_r}} \boldsymbol{e}_{j_{i_r}} \in D,\ (i_1, \dots, i_l) \in I(k,l),\ l \in \{1, \dots, k\} \right\}.$$

The notation required to do this in generality is tedious, but the idea is not complicated, *per se*. Let us illustrate in the case of $m = 4$ and $\boldsymbol{\alpha} = (+,-,+,-)$. Here, a point $\boldsymbol{x} \in D$ is in $D_j^{\boldsymbol{\alpha}}$ if and only if the points

$$\begin{aligned}
&\boldsymbol{x} + h_{j_1} \boldsymbol{e}_{j_1},\\
&\boldsymbol{x} - h_{j_2} \boldsymbol{e}_{j_2},\\
&\boldsymbol{x} + h_{j_3} \boldsymbol{e}_{j_3},\\
&\boldsymbol{x} - h_{j_4} \boldsymbol{e}_{j_4},\\
&\boldsymbol{x} + h_{j_1} \boldsymbol{e}_{j_1} - h_{j_2} \boldsymbol{e}_{j_2},\\
&\boldsymbol{x} + h_{j_1} \boldsymbol{e}_{j_1} + h_{j_3} \boldsymbol{e}_{j_3},\\
&\boldsymbol{x} + h_{j_1} \boldsymbol{e}_{j_1} - h_{j_4} \boldsymbol{e}_{j_4},\\
&\boldsymbol{x} - h_{j_2} \boldsymbol{e}_{j_2} + h_{j_3} \boldsymbol{e}_{j_3},\\
&\boldsymbol{x} - h_{j_2} \boldsymbol{e}_{j_2} - h_{j_4} \boldsymbol{e}_{j_4},\\
&\boldsymbol{x} + h_{j_3} \boldsymbol{e}_{j_3} - h_{j_4} \boldsymbol{e}_{j_4},\\
&\boldsymbol{x} + h_{j_1} \boldsymbol{e}_{j_1} - h_{j_2} \boldsymbol{e}_{j_2} + h_{j_3} \boldsymbol{e}_{j_3},\\
&\boldsymbol{x} + h_{j_1} \boldsymbol{e}_{j_1} - h_{j_2} \boldsymbol{e}_{j_2} - h_{j_4} \boldsymbol{e}_{j_4},\\
&\boldsymbol{x} + h_{j_1} \boldsymbol{e}_{j_1} + h_{j_3} \boldsymbol{e}_{j_3} - h_{j_4} \boldsymbol{e}_{j_4},\\
&\boldsymbol{x} - h_{j_2} \boldsymbol{e}_{j_2} + h_{j_3} \boldsymbol{e}_{j_3} - h_{j_4} \boldsymbol{e}_{j_4},\\
&\boldsymbol{x} + h_{j_1} \boldsymbol{e}_{j_1} - h_{j_2} \boldsymbol{e}_{j_2} + h_{j_3} \boldsymbol{e}_{j_3} - h_{j_4} \boldsymbol{e}_{j_4}
\end{aligned}$$

are all in D. For $\boldsymbol{x} \in D_j^{\boldsymbol{\alpha}}$, we denote

$$D_j^{\boldsymbol{\alpha}}(\boldsymbol{x}) = \left\{ \sum_{r=1}^{l} \boldsymbol{x} \alpha_{i_r} h_{j_{i_r}} \boldsymbol{e}_{j_{i_r}} \;\middle|\; (i_1, \dots, i_l) \in I(k,l),\ l \in \{1, \dots, k\} \right\} \subseteq D.$$

We can then define

$$\begin{aligned}
\Delta_j^{\boldsymbol{\alpha}} \boldsymbol{f} \colon D_j^{\boldsymbol{\alpha}} &\to \mathbb{R}^m\\
\boldsymbol{x} &\mapsto \Delta_{j_1}^{\alpha_1} \cdots \Delta_{j_k}^{\alpha_k} \boldsymbol{f}(\boldsymbol{x}).
\end{aligned}$$

We call $\Delta_j^{\boldsymbol{\alpha}} \boldsymbol{f}(\boldsymbol{x})$ the **k*th-order partial difference*** for $\boldsymbol{f}$ associated with $\boldsymbol{\alpha}$ and $\boldsymbol{j}$. These partial differences, depending on $\boldsymbol{\alpha}$ and $\boldsymbol{j}$, are rather like the partial derivatives of calculus.

As the partial derivatives from calculus conglomerate to give the various higher-order derivatives, the partial differences do the same. To do this, we make a

construction. We denote by A_2 the Abelian group with elements $\{+,-\}$ with the following group operations:

$$++ = +, \quad +- = -, \quad -+ = -, \quad -- = +.$$

Now consider the tensor product $A_2 \otimes \mathbb{R}^n$ of $\mathbb{Z}$-modules.[4] By virtue of being a tensor product, $A_2 \otimes \mathbb{R}^n$ has the structure of an Abelian group and so elements can be added. It becomes a $\mathbb{R}$-vector space with scalar multiplication defined by $a(\alpha \otimes v) = \alpha \otimes (av)$. A basis for this vector space is given by

$$\{+ \otimes e_j \mid \ j \in \{1, \dots, n\}\}, \quad \{- \otimes e_j \mid \ j \in \{1, \dots, n\}\},$$

and so $\dim_{\mathbb{R}}(A_2 \otimes \mathbb{R}^n) = 2n$.

With the preceding construction, we make the following definition.

3.3.3 Definition (Total difference) For $h \in \mathbb{R}^n_{>0}$, for a discrete domain $D \subseteq \mathbb{Z}^n(h)$, for $f\colon D \to \mathbb{R}^m$, and for $k \in \mathbb{Z}_{>0}$, the ***k*th-total difference** of f at

$$x \in \cap\{D^{\alpha}_{j} \mid \ \alpha \in \{+,-\}^k, \ j \in \{1, \dots, n\}^k\}$$

is $\Delta^k f(x) \in \mathrm{L}^m(A_2 \otimes \mathbb{R}^n; \mathbb{R}^m)$ defined by

$$\Delta^k f(x)(\alpha_1 \otimes e_{j_1}, \dots, \alpha_k \otimes e_{j_k}) = \Delta^{\alpha}_{j} f(x), \qquad \alpha \in \{+,-\}^k, \ j \in \{1, \dots, n\}^k. \qquad \bullet$$

We shall denote the first-total difference by Δf.

Note that the definition of $\Delta^k f(x)$, applied to arbitrary arguments, is made by making use of multilinearity.

Let us next see that, like usual partial derivatives, the partial differences are symmetric in a certain sense. To establish this, we introduce the following notation for $\alpha \in \{+,-\}^k$, $j \in \{1, \dots, n\}^k$, and $\sigma \in \mathfrak{S}_k$:

$$\sigma(\alpha) = (\alpha_{\sigma(1)}, \dots, \alpha_{\sigma(k)}), \quad \sigma(j) = (j_{\sigma(1)}, \dots, j_{\sigma(k)}).$$

With this notation, we have the following lemma.

3.3.4 Lemma (Symmetry of the total difference) *For* $\mathbf{h} \in \mathbb{R}^{\mathrm{n}}_{>0}$, *for a discrete domain* $\mathrm{D} \subseteq \mathbb{Z}^{\mathrm{n}}(\mathbf{h})$, *for* $\mathbf{f}\colon \mathrm{D} \to \mathbb{R}^{\mathrm{m}}$, *for* $\mathrm{k} \in \mathbb{Z}_{>0}$, *for* $\boldsymbol{\alpha} \in \{+,-\}^{\mathrm{k}}$, *for* $\mathbf{j} \in \{1, \dots, \mathrm{n}\}^{\mathrm{k}}$, *and for* $\sigma \in \mathfrak{S}_{\mathrm{k}}$, *we have*

$$\Delta^{\sigma(\boldsymbol{\alpha})}_{\sigma(\mathbf{j})} \mathbf{f}(\mathbf{x}) = \Delta^{\boldsymbol{\alpha}}_{\mathbf{j}} \mathbf{f}(\mathbf{x}), \qquad \mathbf{x} \in \bigcap_{\sigma' \in \mathfrak{S}_{\mathrm{m}}} \mathrm{D}^{\sigma'(\boldsymbol{\alpha})}_{\sigma'(\mathbf{j})}.$$

[4] We refer to Section I-5.6.3 for the presentation of tensor products of vector spaces, and this notion is easily extended to tensor products of Abelian groups. In any case, we concretely describe the desired tensor product here.

Proof We prove the result by induction on k. For $k = 2$ the assertion of the lemma is a mere calculation. Suppose the lemma true for $k = r$ and let $\boldsymbol{\alpha} \in \{+,-\}^{r+1}$ and $\boldsymbol{j} \in \{1,\dots,n\}^{r+1}$. By the induction hypothesis,

$$\Delta^{\sigma(\boldsymbol{\alpha})}_{\sigma(j)} \boldsymbol{f}(\boldsymbol{x}) = \Delta^{\boldsymbol{\alpha}}_{\boldsymbol{j}} \boldsymbol{f}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \bigcap_{\sigma' \in \mathfrak{S}_{r+1}} D^{\sigma'(\boldsymbol{\alpha})}_{\sigma'(\boldsymbol{j})},$$

for any permutation σ of the last r terms in the list $\{1,\dots,r+1\}$. We shall show that this equality holds for any *transposition* of $\{1,\dots,r+1\}$. To do this, we need only show that it holds for transpositions involving the first component in the list. By the induction hypothesis, it suffices to show that this holds for the transposition (2 1). However, the conclusion holds in this case by applying the result for $k = 2$ to the function

$$\Delta^{\alpha_3}_{j_3} \cdots \Delta^{\alpha_{r+1}}_{j_{r+1}} \boldsymbol{f}.$$

Since $\mathfrak{S}_{r+1}$ is generated by transpositions by Theorem I-4.1.36, the lemma follows. ■

Based on the lemma, let us denote

$$\begin{aligned} \mathrm{L}^k_{\otimes\mathrm{sym}}(A_2 \otimes \mathbb{R}^n; \mathbb{R}^m) = \Big\{ \boldsymbol{A} \in \mathrm{L}^k(A_2 \otimes \mathbb{R}^n; \mathbb{R}^m) \Big| \; & \boldsymbol{A}(\alpha_{\sigma(1)} \otimes \boldsymbol{e}_{j_{\sigma(1)}}, \dots, \alpha_{\sigma(k)} \otimes \boldsymbol{e}_{j_{\sigma(k)}}) \\ & = \boldsymbol{A}(\alpha_1 \otimes \boldsymbol{e}_{j_1}, \dots, \alpha_k \otimes \boldsymbol{e}_{j_k}),\ \boldsymbol{\alpha} \in \{+,-\}^k,\ \boldsymbol{j} \in \{1,\dots,n\}^k \Big\}. \end{aligned}$$

In the usual way, we denote

$$\mathrm{L}^{\le k}_{\otimes\mathrm{sym}}(A_2 \otimes \mathbb{R}^n; \mathbb{R}^m) = \oplus_{j=1}^k \mathrm{L}^j_{\otimes\mathrm{sym}}(A_2 \otimes \mathbb{R}^n; \mathbb{R}^m).$$

Notationally, it is complicated to reproduce for partial differences the notation for partial derivatives from Section 3.1.1. However, for a function u defined on a subset $D' \subseteq D$ of a discrete domain, we denote the kth partial differences by

$$u^{\boldsymbol{\alpha}}_{\boldsymbol{j}}, \qquad \boldsymbol{\alpha} \in \{+,-\}^k,\ \boldsymbol{j} \in \{1,\dots,n\}^k,$$

keeping in mind the symmetries of Lemma 3.3.4. The subset D' must be selected in such a manner that the partial differences are well-defined. We shall be careful about this in the next section. Here, let us list the coordinates for $\mathrm{L}^{\le 2}_{\otimes\mathrm{sym}}(\mathbb{R}^2; \mathbb{R})$, so that one can see how this works:

$$u^+_1, u^-_1, u^+_2, u^-_2, u^{++}_{11} u^{+-}_{11}, u^{--}_{11}, u^{++}_{12}, u^{+-}_{12}, u^{-+}_{12}, u^{--}_{12}, u^{++}_{22}, u^{+-}_{22}, u^{--}_{22}.$$

We shall say that a function

$$\boldsymbol{u} \colon \mathrm{L}^k_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^n) \to \mathbb{R}^l$$

depends on $\boldsymbol{(\alpha, j) \in \{+,-\}^k \times \{1,\dots,n\}^k}$ if the function obtained by fixing all partial differences other than $\boldsymbol{u^{\alpha}_j}$ at some value is not a constant function.

3.3.2 Difference equations and solutions

In Section 3.1.2, we considered a differential equation as a function of the unknown and its derivatives. We shall do the same for difference equations, instead working with functions of the unknown and its partial differences. A complication that arises is that certain partial differences may not be defined at certain points in a discrete domain since partial differences depend on distant points. With this in mind, we make the following definition.

3.3.5 Definition (Difference equation) A ***difference equation*** consists of a mapping

$$\boldsymbol{F}\colon D_{\boldsymbol{F}} \times U \times \mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l,$$

where $k, l, m, n \in \mathbb{Z}_{>0}$, and for some $D_{\boldsymbol{F}} \subseteq D$ and $U \subseteq \mathbb{R}^m$, with $D \subseteq \mathbb{Z}^n(\boldsymbol{h})$ for some $\boldsymbol{h} \in \mathbb{R}^n_{>0}$, with the requirement that, for each $(\boldsymbol{x}, \boldsymbol{u}) \in D_{\boldsymbol{F}} \times U$ and $r \in \{1, \dots, k\}$,

$$\boldsymbol{F}|\{\boldsymbol{x}\} \times \{\boldsymbol{u}\} \times \mathrm{L}^{\leq r}_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \text{ depends on } (\boldsymbol{\alpha}, \boldsymbol{j}) \in \{+, -\}^r \times \{1, \dots, n\}^r \iff D^{\boldsymbol{\alpha}}_{\boldsymbol{j}}(\boldsymbol{x}) \subseteq D.$$

We also have the following terminology:

(i) n is the number of ***independent variables***;
(ii) m is the number of ***unknowns*** or ***states***;
(iii) k is the ***order***;
(iv) l is the number of ***equations***;
(v) $D \subseteq \mathbb{Z}^n(\boldsymbol{h})$ is the ***domain*** for the difference equation;
(vi) $D_{\boldsymbol{F}} \subseteq D$ is the ***free domain*** for the difference equation;
(vii) $U \subseteq \mathbb{R}^m$ is the ***state space*** for the difference equation. •

The free domain in the definition is made such that all partial differences upon which the equation depends are well-defined at points in $D_{\boldsymbol{F}}$.

In order to understand the relevance of this definition, it is perhaps best to think first about solutions.

3.3.6 Definition (Solution to a difference equation) Let

$$\boldsymbol{F}\colon D_{\boldsymbol{F}} \times U \times \mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l$$

be a difference equation with $D_{\boldsymbol{F}} \subseteq D \subseteq \mathbb{Z}^n(\boldsymbol{h})$. A ***solution*** to the difference equation is a function $\boldsymbol{u}\colon D' \to U$ defined on a subset $D' \subseteq D_{\boldsymbol{F}}$ such that

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \Delta\boldsymbol{u}(\boldsymbol{x}), \dots, \Delta^k\boldsymbol{u}(\boldsymbol{x})) = \boldsymbol{0}, \qquad \boldsymbol{x} \in D'.$$ •

We shall not systematically go through the difference equation examples from Sections 1.1.17–1.1.20. Most of these difference equations are ordinary difference equations that we will work with in detail in Section 3.3.3 below. Instead we shall give a single example that illustrates the essential ideas.

3.3.7 Example (Discrete heat equation in a finite rod) We consider the heat flow in a rod of length ℓ. We discretise the rod into N equal length segments. Thus we take $h_1 = \frac{\ell}{N}$. We also discretise time into intervals of length h_2. This gives the discrete domain

$$D = \{(j_1h_1, j_2h_2) \in \mathbb{Z}^2(h_1, h_2) \mid \ j_1 \in \{0, 1, \dots, N\}, \ j_2 \in \mathbb{Z}_{\geq 0}\}.$$

As we saw in Section 1.1.20, the physics suggests the equation

$$\begin{aligned} &u(j_1h_1, j_2h_2) - u(j_1h_1, (j_2 - 1)h_2) \\ &\quad = k\,(u((j_1 - 1)h_1, (j_2 - 1)h_2) - 2u(j_1h_1, (j_2 - 1)h_2) + u((j_1 + 1)h_1, (j_2 - 1)h_2)) \end{aligned} \tag{3.12}$$

governing the temperature distribution $u\colon D \to \mathbb{R}$, where $k \in \mathbb{R}_{>0}$ is the diffusion coefficient. We recognise this to be equivalent to

$$\Delta_1^- u(j_1h_1, j_2h_2) = k\Delta_{22}^{+-} u(j_1h_1 j_2h_2).$$

We render this a difference equation by taking

$$D_F = \{(j_1h_1, j_2h_2) \in D \mid \ j_1 \in \{1, \dots, N - 1\}, \ j \in \mathbb{Z}_{>0}\}$$

and

$$\begin{aligned} &F\colon D_F \times \mathbb{R} \times \mathrm{L}^{\leq 2}_{\otimes\mathrm{sym}}(\mathbb{R}^2; \mathbb{R}) \to \mathbb{R} \\ &\quad ((x, y), u, (u_1^+, u_1^-, u_2^+, u_2^-, u_{11}^{++} u_{11}^{+-}, u_{11}^{--}, u_{12}^{++}, u_{12}^{+-}, u_{12}^{-+}, u_{12}^{--}, u_{22}^{++}, u_{22}^{+-}, u_{22}^{--})) \mapsto u_1^- - ku_{22}^{+-}. \end{aligned}$$

A solution of this equation defined on D will be a function $u\colon D \to \mathbb{R}$ satisfying (3.12). •

3.3.3 Ordinary difference equations

In this section we specialise to a class of difference equations with one independent variable. Apart from having only one independent variable, the equation we consider are characterised by employing only forward differences. We shall use the following notation for iterated forward differences:

$$\Delta^{k,+} f = \underbrace{\Delta_1^+ \cdots \Delta_1^+}_{k \text{ times}} f.$$

We begin by giving an explicit formula for iterated forward differences.

3.3.8 Lemma (Iterated forward differences in one variable) *Let* $\mathrm{h} \in \mathbb{R}_{>0}$, *let* $\mathrm{D} \subseteq \mathbb{Z}(\mathrm{h})$ *be a discrete domain, let* $\mathbf{f}\colon \mathrm{D} \to \mathbb{R}^{\mathrm{m}}$, *and let* $\mathrm{k} \in \mathbb{Z}_{\geq 0}$. *If* $\mathrm{t} \in \mathbb{T}$ *is such that* $\mathrm{t} + \mathrm{jh} \in \mathrm{D}$, $\mathrm{j} \in \{0, 1, \dots, \mathrm{k}\}$, *then*

$$\Delta^{\mathrm{k},+}\mathbf{f}(\mathrm{t}) = \frac{1}{\mathrm{h}^{\mathrm{k}}} \sum_{\mathrm{j}=0}^{\mathrm{k}} (-1)^{\mathrm{k}-\mathrm{j}} \binom{\mathrm{k}}{\mathrm{j}} \mathbf{f}(\mathrm{t} + \mathrm{jh}).$$

Proof This is proved by induction on k. For $k = 0$ the assertion is that $f(t) = f(t)$, which certainly holds. Assume now the assertion holds for $k = r$. Then, using Exercise I-2.2.2,

$$\begin{aligned}
\Delta^{r+1,+}\boldsymbol{f}(t) &= \frac{1}{h}\left(\Delta^{r,+}\boldsymbol{f}(t+h) - \Delta^{r,+}\boldsymbol{f}(t)\right) \\
&= \frac{1}{h^{r+1}}\left(\sum_{j=0}^{r}(-1)^{r-j}\binom{r}{j}\boldsymbol{f}(t+jh+h) - \sum_{j=0}^{r}(-1)^{r-j}\binom{r}{j}\boldsymbol{f}(t+jh)\right) \\
&= -\frac{1}{h^{r+1}}\left(\sum_{j=1}^{r+1}(-1)^{r-j}\binom{r}{j-1} + \sum_{j=0}^{r}(-1)^{r-j}\binom{r}{j}\right)\boldsymbol{f}(t+jh) \\
&= -\frac{1}{h^{r+1}}\left((-1)^r\boldsymbol{f}(t) + \sum_{j=1}^{r}(-1)^j\left(\binom{r}{j} + \binom{r}{j-1}\right)\boldsymbol{f}(t+jh) - \boldsymbol{f}(t+(r+1)h)\right) \\
&= -\frac{1}{h^{r+1}}\left((-1)^r\boldsymbol{f}(t) + \sum_{j=1}^{r}(-1)^{r-j}\binom{r+1}{j}\boldsymbol{f}(t+jh) - \boldsymbol{f}(t+(r+1)h)\right) \\
&= \frac{1}{h^{r+1}}\sum_{j=0}^{r+1}(-1)^{r-j}\binom{r+1}{j}\boldsymbol{f}(t+jh),
\end{aligned}$$

as desired. ∎

If we restrict to forward differences in one variable, then the total differences take values in $\mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}^m)$ (by taking those elements of $\mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R};\mathbb{R}^m)$ associated with $\boldsymbol{\alpha} = (+,\ldots,+)$). We shall use the following variables to label points in $\mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}^m) \subseteq \mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R};\mathbb{R}^m)$:

$$(\boldsymbol{x}^{(+,1)},\ldots,\boldsymbol{x}^{(+,k)}),$$

echoing the notation used for derivative variables.

3.3.3.1 General ordinary difference equations

We can now give the definition.

3.3.9 Definition (Ordinary difference equation) An ***ordinary difference equation*** is a difference equation $\boldsymbol{F}$ subject to the following conditions:

(i) there is one independent variable, i.e., $n = 1$;

(ii) the independent variable takes values in a subset $\mathbb{T} = I \cap \mathbb{Z}(h)$ called the ***time-domain***, where $I \subseteq \mathbb{R}$ is an interval;

(iii) the free domain is
$$\mathbb{T}_{\boldsymbol{F}} = \{t \in \mathbb{T} \mid t + kh \in \mathbb{T}\};$$

(iv) the ***state space*** is an open subset $X \subseteq \mathbb{R}^m$;

(v) there are the same number of equations as states, i.e., $l = m$;

(vi) $\boldsymbol{F}$ depends only on forward partial differences;

(vii) if the order of the difference equation is k, for each

$$(t, \boldsymbol{x}, \boldsymbol{x}^{(+,1)}, \dots, \boldsymbol{x}^{(+,k-1)}) \in \mathbb{T}_{\boldsymbol{F}} \times X \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m),$$

the equation

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(+,1)}, \dots, \boldsymbol{x}^{(+,k)}) = \boldsymbol{0}$$

can be uniquely solved to give

$$\boldsymbol{x}^{(+,k)} = \overline{\boldsymbol{F}}(t, \boldsymbol{x}, \boldsymbol{x}^{(+,1)}, \dots, \boldsymbol{x}^{(+,k-1)}).$$

We call $\overline{\boldsymbol{F}} \colon \mathbb{T}_{\boldsymbol{F}} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$ the ***difference right-hand side*** for the ordinary difference equation. •

The reader will notice that we have not immediately defined the "right-hand side" as we did for ordinary differential equations. This is because we wish to make an immediate alteration in the way in which we think about ordinary difference equations. To do this, we note from Lemma 3.3.8 that we have

$$\Delta^{r,+} \boldsymbol{f}(t) = h^{-r} \boldsymbol{f}(t + rh) + \frac{1}{h^r} \sum_{j=0}^{r-1} (-1)^{r-j} \binom{r}{j} \boldsymbol{f}(t - jh), \qquad r \in \{0, 1, \dots, k\}. \tag{3.13}$$

and thus we have a bijective mapping between the representations

$$(\boldsymbol{f}(t), \boldsymbol{f}(t + h), \dots, \boldsymbol{f}(t + kh)) \leftrightarrow (\boldsymbol{f}(t), \Delta^{1,+} \boldsymbol{f}(t), \dots, \Delta^{k,+} \boldsymbol{f}(t))$$

of the forward differences up to order k. Thus a solution $\xi \colon \mathbb{T} \to X$ to a difference equation $\boldsymbol{F}$ satisfies

$$\boldsymbol{G}(t, \xi(t), \xi(t + h), \dots, \xi(t + kh)) = \boldsymbol{0}$$

where the mapping $\boldsymbol{G}$ is determined by $\boldsymbol{F}$, using the relations (3.13). (We shall not explicitly determine this formula, since is does not play a rôle in our study of ordinary difference equations.) The property (vii) in the definition of an ordinary difference equation, along with (3.13), then gives

$$\xi(t + kh) = \widehat{\boldsymbol{F}}(t, \xi(t + h), \dots, \xi(t + (k - 1)h)),$$

for a mapping $\widehat{\boldsymbol{F}} \colon \mathbb{T}_{\boldsymbol{F}} \times X \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$. This is the mapping we call the ***right-hand side*** for $\boldsymbol{F}$.

Corresponding to our switching from using forward differences to use shifts in time, we make a change in our variables for difference equations. Specifically, we write the shifted variables as $\boldsymbol{x}^{(j)}$, $j \in \{0, 1, \dots, k\}$, and note that these are defined in terms of $\boldsymbol{x}^{(+,j)}$, $j \in \{0, 1, \dots, k\}$, by

$$\boldsymbol{x}^{(+,j)} = \frac{1}{h^j} \sum_{l=0}^{j} (-1)^{j-l} \binom{j}{l} \boldsymbol{x}^{(l)},$$

as per Lemma 3.3.8. Thus the variables we use mirror in appearance those for ordinary differential equations.

Let us summarise the preceding discussion with the following result.

3.3.10 Proposition (Solutions to ordinary difference equations) *Let* $\mathbf{F}$ *be an ordinary difference equation with time-domain* $\mathbb{T} \subseteq \mathbb{Z}(\mathrm{h})$*, state space* $\mathrm{X} \subseteq \mathbb{R}^{\mathrm{m}}$*, and right-hand side* $\widehat{\mathbf{F}}$*. Then the following statements are equivalent for a map* $\xi\colon \mathbb{T}' \to \mathrm{X}$ *defined on a sub-time-domain* $\mathbb{T}' \subseteq \mathbb{T}$*:*

(i) ξ *is a solution for* $\mathbf{F}$*:*

(ii) ξ *satisfies the equation*

$$\xi(\mathrm{t} + \mathrm{kh}) = \widehat{\mathbf{F}}(\mathrm{t}, \xi(\mathrm{t} + \mathrm{h}), \dots, \xi(\mathrm{t} + (\mathrm{k} - 1)\mathrm{h})), \qquad \mathrm{t} \in \mathbb{T}_{\mathbf{F}}.$$

Proof This follows by arguments mirroring those in Proposition 3.1.7, making use of the discussion above leading to the definition of $\widehat{F}$. ■

The reader can work out how to formulate ordinary difference equations and solutions for the problems of Sections 1.1.17–1.1.19 in Exercises 3.3.1–3.3.3

As with ordinary differential equations, multiple ordinary difference equations can give rise to the same right-hand side. The following definition resolves this ambiguity in favour of one distinguished form of the difference equation.

3.3.11 Definition (Normalised ordinary difference equations) An ordinary difference equation

$$F\colon \mathbb{T} \times X \times \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \times \mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

is ***normalised*** if

$$F(t, x, x^{(1)}, \dots, x^{(k)}) = x^{(k)} - \widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)})$$

for all

$$(t, x, x^{(1)}, \dots, x^{(k-1)}) \in \mathbb{T} \times X \times \mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m).$$ •

Of course, given an ordinary difference equation F, we can effectively replace it with the normalised ordinary difference equation F^* defined by

$$F^*(t, x, x^{(1)}, \dots, x^{(k)}) = x^{(k)} - \widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}),$$

and the solutions of F and F^* agree.

Next we consider a simplification of the structure of ordinary difference equations, a simplification that will commonly hold in practice, and which will be common for us to work with subsequently.

3.3.12 Definition (Autonomous ordinary difference equation) An ordinary difference equation

$$\boldsymbol{F}\colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

is ***autonomous*** if there exists $\boldsymbol{F}_0\colon X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$ such that

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) = \boldsymbol{F}_0(\boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)})$$

for every $(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) \in \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m)$. An ordinary difference equation that is not autonomous is ***nonautonomous***. •

We can characterise autonomous ordinary difference equations by their right-hand sides.

3.3.13 Proposition (Right-hand sides of autonomous ordinary difference equations) *If an ordinary difference equation*

$$\mathbf{F}\colon \mathbb{T} \times \mathrm{X} \times \mathrm{L}^{\leq \mathrm{k}}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

with right-hand side

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{X} \times \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

is autonomous, then there exists

$$\widehat{\mathbf{F}}_0\colon \mathrm{X} \times \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

such that

$$\widehat{\mathbf{F}}(\mathrm{t}, \mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}) = \widehat{\mathbf{F}}_0(\mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}).$$

for every $(\mathrm{t}, \mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}) \in \mathbb{T} \times \mathrm{X} \times \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}})$.

Proof The proof is a simple adaptation of that for Proposition 3.1.11. ■

As with ordinary differential equations, the converse of this result is not generally true, although this is not interesting, cf. Exercise 3.1.20.

3.3.3.2 Linear ordinary difference equations Now we turn to an important class of ordinary difference equations, a class that will occupy much of our subsequent attention to difference equations below.

3.3.14 Definition (Linear ordinary difference equation) Let

$$\boldsymbol{F}\colon \mathbb{T} \times \mathbb{R}^m \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

be an ordinary difference equation with state space $X = \mathbb{R}^m$. The ordinary difference equation $\boldsymbol{F}$ is:

(i) ***linear*** if, for each $t \in \mathbb{T}$, the map

$$\begin{aligned} \boldsymbol{F}_t \colon \mathbb{R}^m \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) &\to \mathbb{R}^m \\ (\boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) &\mapsto \boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) \end{aligned}$$

is affine;

(ii) ***linear homogeneous*** if, for each $t \in \mathbb{T}$, the map $\boldsymbol{F}_t$ is linear;

(iii) ***linear inhomogeneous*** if it is linear but not linear homogeneous. •

Let us characterise linearity in terms of the right-hand side of the ordinary difference equation.

3.3.15 Proposition (Right-hand sides of linear ordinary difference equations) *Let*

$$\mathbf{F} \colon \mathbb{T} \times \mathbb{R}^{\mathrm{m}} \oplus \mathrm{L}^{\leq \mathrm{k}}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

be an ordinary difference equation with right-hand side

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times \mathbb{R}^{\mathrm{m}} \oplus \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{m}}$$

The following statements hold:

(i) if $\mathbf{F}$ *is linear, then, for each* $\mathrm{t} \in \mathbb{T}$*, the map*

$$\begin{aligned} \widehat{\mathbf{F}}_{\mathrm{t}} \colon \mathbb{R}^{\mathrm{m}} \oplus \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^{\mathrm{m}}) &\to \mathbb{R}^{\mathrm{m}} \\ (\mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}) &\mapsto \widehat{\mathbf{F}}(\mathrm{t}, \mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\mathrm{k}-1)}) \end{aligned}$$

is affine;

(ii) if $\mathbf{F}$ *is linear homogeneous, then, for each* $\mathrm{t} \in \mathbb{T}$*, the map* $\widehat{\mathbf{F}}_{\mathrm{t}}$ *is linear;*

(iii) if $\mathbf{F}$ *is linear inhomogeneous, then, for each* $\mathrm{t} \in \mathbb{T}$*, the map* $\widehat{\mathbf{F}}_{\mathrm{t}}$ *is affine but not linear.*

Proof The proof can be carried our like that for Proposition 3.1.13. ■

As with Proposition 3.3.13, the converses to the statements in the preceding result are generally false, cf. Exercise 3.1.21.

As we indicated after the proof of Proposition 3.1.13, one can be more explicit about the form of a linear ordinary difference equation. To wit, a difference equation

$$\boldsymbol{F} \colon \mathbb{T} \times \mathbb{R}^m \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

is linear if and only if there exist maps

$$\boldsymbol{A}_j \colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m), \qquad j \in \{0, 1, \dots, k\},$$

and $\boldsymbol{b} \colon \mathbb{T} \to \mathbb{R}^m$ such that

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}) = \boldsymbol{A}_k(t)(\boldsymbol{x}^{(k)}) + \cdots + \boldsymbol{A}_1(t)(\boldsymbol{x}^{(1)}) + \boldsymbol{A}_0(t)(\boldsymbol{x}) + \boldsymbol{b}(t). \tag{3.14}$$

The right-hand side is then

$$-\boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_{k-1}(t)(\boldsymbol{x}^{(k-1)}) - \cdots - \boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_0(t)(\boldsymbol{x}) - \boldsymbol{A}_k^{-1}(t)(\boldsymbol{b}(t)).$$

Solutions to this ordinary difference equation are then functions $t \mapsto \boldsymbol{x}(t)$ satisfying

$$\boldsymbol{x}(t + kh) = -\boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_{k-1}(t)\left(\boldsymbol{x}(t - (k-1)h)\right) - \cdots - \boldsymbol{A}_k^{-1}(t) \circ \boldsymbol{A}_0(t)\left(\boldsymbol{x}(t)\right) - \boldsymbol{A}_k^{-1}(t)(\boldsymbol{b}(t)).$$

We shall study equations like this in great detail subsequently, particularly in the case when the linear maps $\boldsymbol{A}_0, \boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ are independent of t. Indeed, equations like this have a particular name.

3.3.16 Definition (Constant coefficient linear ordinary difference equation) A linear ordinary difference equation given by (3.14) is a ***constant coefficient linear ordinary difference equation*** if the functions $\boldsymbol{A}_0, \boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ are independent of t. •

3.3.3.3 Linear ordinary difference equations in vector spaces In Section 3.1.3.3 we considered linear ordinary differential equations whose state space is an abstract finite-dimensional $\mathbb{R}$-vector space. We wish to carry out the same sort of abstraction for difference equations, and for the same reasons.

The definitions we need are the following.

3.3.17 Definition (System of linear ordinary difference equations) Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let $\mathbb{T} \subseteq \mathbb{Z}(h)$ be a discrete time-domain, and let V be an n-dimensional $\mathbb{F}$-vector space.

(i) A ***system of linear ordinary difference equations*** in V is a map $F\colon \mathbb{T}\times\mathsf{V}\oplus\mathsf{V} \to \mathsf{V}$ of the form

$$F(t, x, x^{(1)}) = \mathsf{A}_1(t)(x^{(1)}) + \mathsf{A}_0(t)(x) - b_0(t)$$

for maps $\mathsf{A}_0, \mathsf{A}_1\colon \mathbb{T} \to \mathrm{L}(\mathsf{V};\mathsf{V})$ and $b_0\colon \mathbb{T} \to \mathsf{V}$, where $\mathsf{A}_1(t)$ is invertible for every $t \in \mathbb{T}$.

(ii) The ***right-hand side*** of a system of linear ordinary difference equations F is the map $\widehat{F}\colon \mathbb{T} \times \mathsf{V} \to \mathsf{V}$ is the map defined by

$$\widehat{F}(t, x) = -\mathsf{A}_1(t)^{-1} \circ \mathsf{A}_0(t)(x) + \mathsf{A}_1(t)^{-1}(b_0(t)).$$

We shall typically denote $\mathsf{A}(t) = -\mathsf{A}_1(t)^{-1} \circ \mathsf{A}_0(t)$ and $b(t) = \mathsf{A}_1(t)^{-1}(b_0(t))$.

(iii) The system of linear ordinary difference equations F

(a) is ***homogeneous*** if $b(t) = 0$ for every $t \in \mathbb{T}$,

(b) is ***inhomogeneous*** if $b(t) \neq 0$ for some $t \in \mathbb{T}$, and

(c) has ***constant coefficients*** if A is a constant map.

(iv) A ***solution*** for a system of linear ordinary difference equations F is a map $\xi \in \mathsf{V}^{\mathbb{T}'}$ defined on a sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}$ and satisfying

$$\xi(t + h) = \mathsf{A}(t)(\xi(t)) + b(t), \qquad t \in \mathbb{T}' \cap \mathbb{T}_F.$$ •

Note that, because we are considering difference equations of order 1, we always have

$$\mathbb{T}_F = \{t \in \mathbb{T} \mid \ t + h \in \mathbb{T}\}.$$

3.3.4 Partial difference equations

We consider in this section the analogue for difference equations of the sorts of differential equations considered in Section 3.1.4. To do so, we introduce some useful general notation for the various variables and for the total differences. Independent variables will be denoted by $\boldsymbol{x}$ and states or unknowns by $\boldsymbol{u}$. Then the list of the coordinates representing the total differences up to order k of the dependent variables with respect to the independent variables will be denoted by

$$(\boldsymbol{u}, \boldsymbol{u}^{(\otimes,1)}, \dots, \boldsymbol{u}^{(\otimes,k)}) \in U \times \mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m).$$

3.3.4.1 General partial difference equations

We begin with the definition.

3.3.18 Definition (Partial difference equation) A ***partial difference equation*** is a difference equation

$$\boldsymbol{F}\colon D \times U \times \mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l$$

with the following properties:

(i) $n > 1$;

(ii) there exists $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{u}^{(\otimes,1)}, \dots, \boldsymbol{u}^{(\otimes,k-1)}) \in D \times U \times \mathrm{L}^{\leq k-1}_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m)$ such that the function

$$\boldsymbol{u}^{(k)} \mapsto \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{u}^{(\otimes,1)}, \dots, \boldsymbol{u}^{(\otimes,k-1)}, \boldsymbol{u}^{(\otimes,k)})$$

is not constant. •

As with partial differential equations, the second condition is that $\boldsymbol{F}$ should not be everywhere independent of the highest-order derivative. This is a condition that, while technically required for a sensible notion of order for a partial difference equation, is always met in practice.

As with general partial differential equations, there is not a lot one can say about general partial difference equations. Thus we turn to a consideration of such difference equations in the presence of additional structure.

3.3.4.2 Linear and quasilinear partial difference equations

Let us provide the appropriate definitions of linearity for partial difference equations.

3.3.19 Definition (Linear partial difference equation) Let

$$\boldsymbol{F}\colon D \times \mathbb{R}^m \oplus \mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) \to \mathbb{R}^l$$

be a partial difference equation with state space $U = \mathbb{R}^m$. The partial difference equation $\boldsymbol{F}$ is:

(i) ***linear*** if, for each $\boldsymbol{x} \in D$, the map

$$\begin{aligned}\boldsymbol{F}_{\boldsymbol{x}}\colon \mathbb{R}^m \oplus \mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R}^n; \mathbb{R}^m) &\to \mathbb{R}^l\\ (\boldsymbol{u}, \boldsymbol{u}^{(\otimes,1)}, \dots, \boldsymbol{u}^{(\otimes,k)}) &\mapsto \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{u}^{(\otimes,1)}, \dots, \boldsymbol{u}^{(\otimes,k)})\end{aligned}$$

is affine;

(ii) ***linear homogeneous*** if, for each $x \in D$, the map F_x is linear;
(iii) ***linear inhomogeneous*** if it is linear but not linear homogeneous. •

3.3.20 Definition (Quasilinear partial difference equation) A partial difference equation

$$F\colon D \times U \times \mathrm{L}^{\leq k}_{\otimes\mathrm{sym}}(\mathbb{R}^n;\mathbb{R}^m) \to \mathbb{R}^l$$

is ***quasilinear*** if, for each

$$(x, u, u^{(\otimes,1)}, \dots, u^{(\otimes,k-1)}) \in D \times U \times \mathrm{L}^{\leq k-1}_{\otimes\mathrm{sym}}(\mathbb{R}^n;\mathbb{R}^m),$$

the map

$$u^{(\otimes,k)} \mapsto F(x, u, u^{(\otimes,1)}, \dots, u^{(\otimes,k)})$$

is affine. •

We can immediately deduce from the definitions the following forms for the various flavours of linear and quasilinear partial difference equations.

3.3.21 Proposition (Linear partial difference equations) *Let*

$$\mathbf{F}\colon \mathrm{D} \times \mathbb{R}^{\mathrm{m}} \oplus \mathrm{L}^{\leq \mathrm{k}}_{\otimes\mathrm{sym}}(\mathbb{R}^{\mathrm{n}};\mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{l}}$$

be a partial difference equation with state space $\mathrm{U} = \mathbb{R}^{\mathrm{m}}$. *Then the following statements hold:*

(i) $\mathbf{F}$ *is linear if and only if there exist maps*

$$\mathbf{A}_{\mathrm{j}}\colon \mathrm{D} \to \mathrm{L}(\mathrm{L}^{\mathrm{j}}_{\otimes\mathrm{sym}}(\mathbb{R}^{\mathrm{n}};\mathbb{R}^{\mathrm{m}});\mathbb{R}^{\mathrm{l}}), \qquad \mathrm{j} \in \{0, 1, \dots, \mathrm{k}\},$$

and $\mathbf{b}\colon \mathrm{D} \to \mathbb{R}^{\mathrm{l}}$, *with* $\mathbf{A}_{\mathrm{k}}$ *not identically zero, such that*

$$\mathbf{F}(\mathbf{x}, \mathbf{u}, \mathbf{u}^{(\otimes,1)}, \dots, \mathbf{u}^{(\otimes,\mathrm{k})}) = \mathbf{A}_{\mathrm{k}}(\mathbf{x})(\mathbf{u}^{(\otimes,\mathrm{k})}) + \cdots + \mathbf{A}_1(\mathbf{x})(\mathbf{u}^{(\otimes,1)}) + \mathbf{A}_0(\mathbf{x})(\mathbf{u}) + \mathbf{b}(\mathbf{x}); \quad (3.15)$$

(ii) $\mathbf{F}$ *is linear homogeneous if and only if it has the form from part (i) with* $\mathbf{b}(\mathbf{x}) = \mathbf{0}$ *for every* $\mathbf{x} \in \mathrm{D}$;

(iii) $\mathbf{F}$ *is linear inhomogeneous if and only if it has the form from part (i) with* $\mathbf{b}(\mathbf{x}) \neq \mathbf{0}$ *for some* $\mathbf{x} \in \mathrm{D}$.

3.3.22 Proposition (Quasilinear partial difference equations) *A partial difference equation*

$$\mathbf{F}\colon \mathrm{D} \times \mathrm{U} \times \mathrm{L}^{\leq \mathrm{k}}_{\otimes\mathrm{sym}}(\mathbb{R}^{\mathrm{n}};\mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{l}}$$

is quasilinear if and only if there exist maps

$$\mathbf{A}_1\colon \mathrm{D} \times \mathrm{U} \times \mathrm{L}^{\leq \mathrm{k}-1}_{\otimes\mathrm{sym}}(\mathbb{R}^{\mathrm{n}};\mathbb{R}^{\mathrm{m}}) \to \mathrm{L}(\mathrm{L}^{\mathrm{k}}_{\otimes\mathrm{sym}}(\mathbb{R}^{\mathrm{n}};\mathbb{R}^{\mathrm{m}});\mathbb{R}^{\mathrm{l}}),$$
$$\mathbf{A}_0\colon \mathrm{D} \times \mathrm{U} \times \mathrm{L}^{\leq \mathrm{k}-1}_{\otimes\mathrm{sym}}(\mathbb{R}^{\mathrm{n}};\mathbb{R}^{\mathrm{m}}) \to \mathbb{R}^{\mathrm{l}},$$

with $\mathbf{A}_1$ *not identically zero, such that*

$$\mathrm{F}(\mathbf{x}, \mathbf{u}, \mathbf{u}^{(\otimes,1)}, \dots, \mathbf{u}^{(\otimes,\mathrm{k})}) = \mathbf{A}_1(\mathbf{x}, \mathbf{u}, \mathbf{u}^{(\otimes,1)}, \dots, \mathbf{u}^{(\otimes,\mathrm{k}-1)})(\mathbf{u}^{(\otimes,\mathrm{k})}) + \mathbf{A}_0(\mathbf{x}, \mathbf{u}, \mathbf{u}^{(\otimes,1)}, \dots, \mathbf{u}^{(\otimes,\mathrm{k}-1)}).$$

The notion of having constant coefficients that we encountered for ordinary difference equations also makes sense for partial difference equations.

3.3.23 Definition (Constant coefficient linear partial difference equation) A linear partial difference equation given by (3.15) is a ***constant coefficient linear partial difference equation*** if the functions $A_0, A_1, \dots, A_k$ are constant. •

3.3.4.3 Elliptic, hyperbolic, and parabolic second-order linear partial difference equations Many of the partial difference equations that arise from physics are linear second-order equations with a single unknown, and there are various classifications that can be applied to such equations that bear on the attributes of the solutions to these equations.

Let us explicitly write the form of a class of such equations. Unlike with partial differential equations, for difference equations one must keep track of forward and backward partial differences. This makes it rather tedious to consider all possible second-order linear partial difference equations. What we will do, therefore, is simply consider the classical examples of such equations.

3.3.24 Examples (Elliptic, hyperbolic, and parabolic partial difference equations) We shall not be too fussy here with the free domain D_F, but focus instead on the equation $\boldsymbol{F}$.

1. The standard example of an elliptic partial difference equation is the ***potential equation***, or ***Laplace's equation***. The domain $D \subseteq \mathbb{Z}^n(\boldsymbol{h})$ is normally thought of as being "space" in this case, so we denote coordinates for D by $(x_1, \dots, x_n)$. Frequently, $\boldsymbol{h} = (h, \dots, h)$, i.e., the discretisation is the same in all coordinate directions. We consider the second partial differences

$$\Delta^{+,-}_{j,k} u(\boldsymbol{x}) = \frac{1}{h^2}\Big(u(\boldsymbol{x} - h\boldsymbol{e}_j) - 2u(\boldsymbol{x}) + u(\boldsymbol{x} + h\boldsymbol{e}_k)\Big), \qquad j, k \in \{1, \dots, n\}.$$

   Then the difference equation is given by

$$\boldsymbol{F}(\boldsymbol{x}, u, u^{(\otimes,1)}, u^{(\otimes,2)}) = u^{+,-}_{1,1} + \cdots + u^{+,-}_{n,n}.$$

   Thus $u \colon D' \to \mathbb{R}$ is a solution if it satisfies

$$\sum_{j=1}^{n} (u(\boldsymbol{x} - h\boldsymbol{e}_j) - 2u(\boldsymbol{x}) + u(\boldsymbol{x} + h\boldsymbol{e}_j)) = 0.$$

2. The standard example of an hyperbolic partial difference equation is the ***wave equation***. In this case, the domain D is normally thought of as encoding time and space, and so we denote coordinates by $(t, x_1, \dots, x_n)$. Here we use backward differences in time, and the same spatial differences as in the potential equation above. The discretisation is assumed to be h_1 in the time coordinate and h_2 in the spatial coordinate. The difference equation is then given by

$$F((t, \boldsymbol{x}), u, u^{(\otimes,1)}, u^{(\otimes,2)}) = -u^{-,-}_{t,t} + u^{+,-}_{1,1} + \cdots + u^{+,-}_{n,n}.$$

Solutions u thus satisfy the equation

$$u(t-2h_1,\boldsymbol{x})-2u(t-h_1,\boldsymbol{x})+u(t,\boldsymbol{x})=\frac{h_1^2}{h_2^2}\sum_{j=1}^{n}(u(t,\boldsymbol{x}-h_2\boldsymbol{e}_j)-2u(t,\boldsymbol{x})+u(t,\boldsymbol{x}+h_2\boldsymbol{e}_j)).$$

3. The usual example of a parabolic difference equation is the ***heat equation***, which we saw modelled the temperature distribution in a rod in Section 1.1.20. In this case, like the wave equation, the domain D is coordinatised by time and space: $(t,x_1,\dots,x_n)$. The differential equation is

$$F((t,\boldsymbol{x}),u,u^{(\otimes,1)},u^{(\otimes,2)})=-u_t^-+u_{1,1}^{+,-}+\cdots+u_{n,n}^{+,-}.$$

Solutions $u\colon D'\to\mathbb{R}$ satisfy

$$u(t,\boldsymbol{x})-u(t-h_1,\boldsymbol{x})=\frac{h_1}{h_2^2}\sum_{j=1}^{n}(u(t,\boldsymbol{x}-h_2\boldsymbol{e}_j)-2u(t,\boldsymbol{x})+u(t,\boldsymbol{x}+h_2\boldsymbol{e}_j)).$$ •

3.3.5 How to think about difference equations

We shall consider systematic ways of solving some ordinary difference equations, but before we do so, it is interesting to think about some conceptual aspects of difference equations. Many of the problems attached to difference equations resemble those for differential equations discussed in Section 3.1.5.

Let us enumerate some ways of thinking about difference equations.

1. *Character of difference equations:* Note that differential equations involve... well... derivatives. In contrast, difference equations are purely algebraic equations. This sometimes makes it more straightforward to obtain numerical solutions using the computer, since the need to approximate derivatives is obviated. However, matters like existence and uniqueness of solutions are still relevant for difference equations, especially partial difference equations (we shall consider the matter of existence and uniqueness of solutions for ordinary difference equations in Section 3.4). It is also generally no easier to obtain "closed-form" solutions, even when these can be obtained.
2. *Analysis:* The issues of examining steady-state behaviour, approximating solutions, and studying equilibria and their stability—discussed when we discussed how to think about differential equations—arise also for ordinary difference equations.
3. *Numerical solution:* For ordinary difference equations, numerical solution is natural, and indeed the numerical solution of ordinary differential equations produces an ordinary difference equation, i.e., the ordinary differential equation is replaced with an ordinary difference equation with a small discretisation interval. For partial difference equations, the matter of numerical solution is not an entirely trivial one, and is the subject of much study.

For ordinary differential equations, we saw in Example 3.1.25 that one can represent solutions as curves in the state space. For ordinary difference equations, the same idea is valid, except curves are not continuous but discrete. Let us illustrate this with an example.

3.3.25 Example (Ordinary difference equations and discrete curves)

Exercises

3.3.1 For the completely unrealistic rabbit population model of Section 1.1.17, do the following:
(a) identify the time-domain, the free time-domain, and the state space for the ordinary difference equation;
(b) write $\boldsymbol{F}$ using the ordinary difference equation notation for derivatives;
(c) show that $\boldsymbol{F}$ is an ordinary difference equation;
(d) write down the right-hand side;
(e) write the condition for a solution using Proposition 3.3.10.

3.3.2 For the simple bank balance model of Section 1.1.18, do the following:
(a) identify the time-domain, the free time-domain, and the state space for the ordinary difference equation;
(b) write $\boldsymbol{F}$ using the ordinary difference equation notation for derivatives;
(c) show that $\boldsymbol{F}$ is an ordinary difference equation;
(d) write down the right-hand side;
(e) write the condition for a solution using Proposition 3.3.10.

3.3.3 For the simple national income model of Section 1.1.18, do the following:
(a) identify the time-domain, the free time-domain, and the state space for the ordinary difference equation;
(b) write $\boldsymbol{F}$ using the ordinary difference equation notation for derivatives;
(c) show that $\boldsymbol{F}$ is an ordinary difference equation;
(d) write down the right-hand side;
(e) write the condition for a solution using Proposition 3.3.10.

3.3.4 Consider signals defined on a discrete time-domain contained in $\mathbb{Z}(h)$. Consider the forward difference operator in a single independent variable

$$\Delta^{1,+} f(t) = \frac{1}{h}(f(t+h) - f(t)).$$

Prove the following statements:
(a) $\Delta^{1,+}(\alpha) = 0$, if α is a constant function;
(b) $\Delta^{1,+}(\mathsf{alt})(jh) = \frac{2}{h}\mathsf{alt}(t+h)$, where $\mathsf{alt}(t) = (-1)^{t/h}$;
(c) $\Delta^{1,+}(\mathsf{P}_a)(t) = \frac{a^h-1}{h}\mathsf{P}_a(t)$, where $\mathsf{P}_a(t) = a^{t/h}$;
(d) $\Delta^{1,+}(\mathrm{pow}_j)(t) = \sum_{l=0}^{j-1} h^{l-j} t^j$, where $\mathrm{pow}_j(t) = t^j$;

(e) $\Delta^{1,+}(fg)(t) = (\Delta^{1,+}f)(t)g(t) + f(t+h)(\Delta^{1,+}g)(t)$.

3.3.5 Prove the higher-order Leibniz Rule for iterated forward differences:

$$\Delta^{k,+}(fg)(t) = \sum_{j=0}^{k} \binom{k}{j} \Delta^{k-j,+}f(t+jh)\Delta^{j,+}g(t).$$

3.3.6 State and prove the "Fundamental Theorem of Calculus" for forward finite differences in a single variable.

3.3.7 Let

$$\boldsymbol{F}\colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}^m) \to \mathbb{R}^m$$

be an ordinary difference equation with right-hand side $\widehat{\boldsymbol{F}}$. Suppose that $\mathbb{T} \subseteq \mathbb{Z}(h)$. As usual, let t be the independent variable and $\boldsymbol{x}$ the state, with $\boldsymbol{x}^{(+,j)} \in \mathrm{L}^{j}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}^m)$ being the coordinate for the jth derivative. As per Remark 3.1.5, we can think of $\boldsymbol{x}^{(+,j)}$ as being an element of $\mathbb{R}^m$.

We will associate to $\boldsymbol{F}$ a first-order ordinary difference equation $\boldsymbol{F}_1$ with time domain $\mathbb{T}'$ and state space

$$X_1 = X \times \underbrace{\mathbb{R}^m \times \cdots \times \mathbb{R}^m}_{k-1 \text{ times}}.$$

To do so, answer the following questions.

(a) Denote coordinates for the state space X_1 by $\boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_{k-1}$, and relate these to $(\boldsymbol{x}, \boldsymbol{x}^{(+,1)}, \dots, \boldsymbol{x}^{(+,k-1)})$ by

$$\boldsymbol{y}_0 = \boldsymbol{x},\ \boldsymbol{y}_j = \boldsymbol{x}^{(j)}, \qquad j \in \{1, \dots, k-1\}.$$

If $t \mapsto \boldsymbol{x}(t)$ is a solution for $\boldsymbol{F}$, write down the corresponding difference equations that must be satisfied by $(\boldsymbol{y}_0, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{k-1})$.
Hint: *For each* $\mathrm{j} \in \{0, 1, \dots, \mathrm{k}-1\}$, *write down* $\mathbf{y}_{\mathrm{j}}(\mathrm{t})$ *in terms of* $\mathbf{x}(\mathrm{t} - \mathrm{lh})$, $\mathrm{l} \in \{0, 1, \dots, \mathrm{k}-1\}$, *and express the result in terms of the coordinates for* X_1.

(b) What is the right-hand side $\widehat{\boldsymbol{F}}_1$ corresponding to the equations you derived in part (a)?

(c) Write down a first-order ordinary difference equation $\boldsymbol{F}_1$ with time domain $\mathbb{T}'$ and state space X_1 whose right-hand side is the function $\widehat{\boldsymbol{F}}_1$ you determined in part (b). Part of the problem is to define $\mathbb{T}'$ appropriately.

(d) State *precisely* the relationship between solutions for $\boldsymbol{F}$ and solutions for $\boldsymbol{F}_1$.

(e) Show that $\boldsymbol{F}_1$ can be taken to be linear if $\boldsymbol{F}$ is linear, and show that $\boldsymbol{F}_1$ is homogeneous if and only if $\boldsymbol{F}$ is, in this case.

In the next exercise we shall show how autonomous ordinary differential equations are special in terms of their solutions. In order for the exercise to make sense, we require the existence and uniqueness theorem we state below, Theorem 3.4.2.

3.3.5 Let

$$\boldsymbol{F}\colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

be an autonomous ordinary difference equation satisfying the conditions of Theorem 3.4.2, let

$$(\boldsymbol{x}_0, \boldsymbol{x}_0^{(1)}, \dots, \boldsymbol{x}_0^{(k-1)}) \in X \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m),$$

and let $t_1, t_2 \in \mathbb{T}$. Let $\xi_1\colon \mathbb{T} \to X$ and $\xi_2\colon \mathbb{T} \to X$ be solutions for $\boldsymbol{F}$ satisfying

$$\xi_1(t_1) = \xi_2(t_2) = \boldsymbol{x}_0, \quad \xi_1(t_1 + jh) = \xi_2(t_2 + jh) = \boldsymbol{x}_0^{(j)}, \qquad j \in \{1, \dots, k-1\}.$$

Answer the following questions.

(a) Show that $\xi_2(t) = \xi_1(t + t_1 - t_2)$ for all $t \in \mathbb{T}$ for which $\xi(t)$ is defined and for which $t + t_1 - t_2 \in \mathbb{T}$.

(b) Assuming that $\mathbb{T} = \mathbb{R}$ and that all solutions are defined for all time for simplicity, express your conclusion from part (a) as a condition on the flow $\Phi^{\boldsymbol{F}}$.

3.3.6 Consider the ordinary difference equations of Sections 1.1.17–1.1.19.

(a) Which of the equations is autonomous?

(b) Which of the equations is linear?

(c) Which of the equations is linear and homogeneous?

(d) Which of the equations is linear and inhomogeneous?

(e) Which of the equations is a linear constant coefficient equation?

3.3.7 *(Mini-project)* We consider a nonlinear generalisation of a model of the economy known as the Leontief input-output model. In this model, examined by Chander [1983], the economy is divided into n individual sectors and in each sector of the economy a single good is produced that is either traded, consumed, or reinvested. For $j \in \{1, \dots, n\}$, by $x_j \in \mathbb{R}_{\geq 0}$ we denote the quantity of the jth good. For $j, k \in \{1, \dots, k\}$, we denote by $a_{jk}(x_k)$ the amount of the jth good used to quantity x_k of the kth good. For $j \in \{1, \dots, n\}$, we denote by d_j the final demand for the jth good. If the economy is in equilibrium, i.e., exactly as much of each good is produced as needed, then it holds that

$$x_j - \sum_{k=1}^{n} a_{jk}(x_k) = d_j, \qquad j \in \{1, \dots, n\}.$$

We make two assumptions.

1. We assume that, for $j, k \in \{1, \dots, n\}$, $a_{jk}\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ is of class C^1, $a_{jk}(0) = 0$, $a'_{jk}(s) \geq 0$ for $s \in \mathbb{R}_{\geq 0}$. This roughly means that one requires more of good j if one produces more of good k.

2. We assume that there exists $p_1, \dots, p_n \in \mathbb{R}_{\geq 0}$ and $v_1, \dots, v_n \in \mathbb{R}_{>0}$ such that

$$p_k \geq \sum_{k=1}^{n} p_j a'_{jk}(s) + v_j, \qquad s \in \mathbb{R}_{\geq 0},\ k \in \{1, \dots, n\}.$$

This assumption has the following interpretation. The quantity p_j is the price of good j (which can be set) so that the sum in the preceding expression can be thought of as the rate of increase in production cost for good k when the prices are set to $p_1, \dots, p_n$. The assumption, then, is that prices can be set in such a way that positive value (the quantities $v_1, \dots, v_n$) is added when production is increased.

If we suppose that at the end of the kth cycle, we are at a nonequilibrium production vector $\boldsymbol{x}(k)$, then we adjust the production at the beginning of the $(k+1)$st cycle by

$$\boldsymbol{x}(k+1) = \boldsymbol{a}(\boldsymbol{x}(k)) + \boldsymbol{d}.$$

We wish to assemble all of this into an ordinary difference equation.

(a) What is the state space X for the system?

(b) What is the time-domain $\mathbb{T}$ for the system?

(c) What are the dynamics $\boldsymbol{f}$?

Do some explorations as follows.

(d) Do some research to describe what the model is used for and how the ordinary differential equation model should behave to be useful.

(e) With the stated assumptions, prove the following theorem.

Theorem *For each* $\mathbf{d} \in \mathbb{R}^n_{\geq 0}$, *there exists a unique* $\mathbf{x}^* \in \mathbb{R}^n_{\geq 0}$ *such that*

(i) $\mathbf{x}^* = \mathbf{a}(\mathbf{x}^*) + \mathbf{d}$ *and*

(ii) for any $\mathbf{x}(0) \in \mathbb{R}^n_{\geq 0}$, *it holds that* $\lim_{k\to\infty} \mathbf{x}(k) = \mathbf{x}^*$.

Hint: *Use the Contraction Mapping Theorem, Theorem* III-1.1.23.

Section 3.4

Existence and uniqueness of solutions for difference equations

We consider in this section the matter of existence and uniqueness of solutions for difference equations, mirroring what we did in Section 3.2 for differential equations. As we shall see, the theory for difference equations does not track exactly that for differential equations. None of the complications of Section 3.2.1.1 arise for difference equations. On the other hand, sets of solutions for difference equations possess a different sort of structure than their counterparts for differential equations.

Do I need to read this section? As with Section 3.2 for differential equations, the theoretical results of this section are not required knowledge for much of what we shall do subsequently. However, the results for ordinary difference equations are far simpler than those for ordinary differential equations. Moreover, the flow notation we introduce in Definition 3.4.3 will be useful at many points in the text. •

3.4.1 Results for ordinary difference equations

It is possible to give a pretty complete characterisation of existence and uniqueness of solutions for ordinary difference equations. This is because the results are, in summary, solutions exist and are unique. To set this up properly, we first need a precise formulation of the problem whose solutions exist and are unique. We restrict ourselves to ordinary difference equations of order one, noting that this can be done without loss of generality by Exercises 3.3.7 and 3.4.2.

3.4.1 Definition (Initial value problem for ordinary difference equations) Let F be an ordinary difference equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^m$$

with time-domain $\mathbb{T} \subseteq \mathbb{Z}(h)$. Let $t_0 \in \mathbb{T}_F$ and $x_0 \in X$. A map $\xi\colon \mathbb{T}' \to X$ is a ***solution*** for F with ***initial value*** x_0 at t_0 if it satisfies the following conditions:

(i) $\mathbb{T}' \subseteq \mathbb{T}$ is a sub-time-domain;

(ii) $\xi(t + h) = \widehat{F}(t, \xi(t))$ for each $t \in \mathbb{T}_F \cap \mathbb{T}'$;

(iii) $\xi(t_0) = x_0$.

In this case, we say that ξ is a solution to the ***initial value problem***

$$\xi(t + h) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x_0. \qquad \bullet$$

3.4.1.1 Principal existence and uniqueness theorems for ordinary difference equations

It is now possible to state the analogue of Theorem 3.2.8 for ordinary difference equations.

3.4.2 Theorem (Existence and uniqueness of solutions for ordinary difference equations) *Let $X \subseteq \mathbb{R}^m$ be open, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $\mathbf{F}$ be a first-order ordinary difference equation with right-hand side*

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times X \to \mathbb{R}^m.$$

Then, for each $(t_0, x_0) \in \mathbb{T}_{\mathbf{F}} \times X$, there exists a sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}$ with $t_0 \in \mathbb{T}'$, and a solution $\xi \colon \mathbb{T}' \to \mathbb{R}^m$ for $\mathbf{F}$ such that $\xi(t_0) = x_0$. Moreover, if $\mathbb{T}''$ is another such sub-time-domain and $\eta \colon \mathbb{T}'' \to \mathbb{R}^m$ is another such solution, then $\eta(t) = \xi(t)$ for all $t \in \mathbb{T}'' \cap \mathbb{T}'$. Finally, if $\widehat{\mathbf{F}}$ takes values in X, then the preceding conclusions hold for any sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}_{\geq t_0}$.

Proof Define $\xi \colon \mathbb{T}_{\geq t_0} \to \mathbb{R}^m$ recursively by

$$\xi(t_0) = x_0, \quad \xi(t_0 + kh) = \widehat{F}(t_0 + (k-1)h, \xi(t_0 + (k-1)h)).$$

Clearly this construction is well-defined for $t \in \mathbb{T}_{\geq t_0}$, as long as $\xi(t + h) \in X$, and this gives the existence assertion. Moreover, by its very construction, ξ is the *only* solution to the initial value problem. Finally, if $\widehat{F}$ takes values in X, this construction can be made for every $t \in \mathbb{T}_{\geq t_0}$. ■

Let us contrast this to the result of Theorem 3.2.8 for ordinary differential equations.

1. Unlike Theorem 3.2.8, there are no hypotheses required on the right-hand side $\widehat{F}$. That is, we do not require any regularity of $\widehat{F}$, either in t or in x. We shall see in Section 3.4.1.2 that there are additional conditions one can place on $\widehat{F}$ that will give additional properties of solutions or, more properly, the set of solutions.
2. The sub-time-domain $\mathbb{T}'$ on which a solution exists is only asserted to consist of times larger than t_0. This is because, in general, solutions may not exist for times smaller than t_0. We shall discuss matters like this in Section 3.4.1.2.
3. Unlike with ordinary differential equations where solutions generally exist only for small times, solutions for ordinary difference equations exist for *all* times larger than t_0, as long as solutions remain in X. Ordinary difference equations for which $\widehat{F}$ takes values in X will be called ***complete***.

3.4.1.2 Flows for ordinary difference equations In Section 3.2.1.3 we went to some significant lengths to define the flow for an ordinary differential equation. For ordinary difference equations, we have similar characterisations, but there are some differences that we will mention as we go along.

First of all, we can directly define the flow by virtue of the fact that there are no hypotheses required for F.

3.4.3 Definition (Flow of an ordinary difference equation) Let $\boldsymbol{F}$ be an ordinary difference equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^m.$$

(i) The ***interval of existence*** for the initial value problem

$$\xi(t+h) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x}_0,$$

is

$$J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) = \cup\{\mathbb{T}' \subseteq \mathbb{T} | \text{ there is a solution for the initial value problem defined on } \mathbb{T}'\}.$$

(ii) The ***domain of solutions*** for $\boldsymbol{F}$ is

$$D_{\boldsymbol{F}} = \{(t, t_0, \boldsymbol{x}_0) \in \mathbb{T} \times \mathbb{T} \times X \mid \ t \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0)\}.$$

(iii) We use the notation

$$D_{\boldsymbol{F}}(t, t_0) = \{\boldsymbol{x} \in X \mid \ (t, t_0, \boldsymbol{x}) \in D_{\boldsymbol{F}}\}.$$

(iv) The ***flow*** of $\boldsymbol{F}$ is the map $\Phi^{\boldsymbol{F}}\colon D_{\boldsymbol{F}} \to \mathbb{R}^m$ defined by asking that $\Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}_0)$ is the solution, evaluated at t, of the initial value problem

$$\xi(\tau + h) = \widehat{\boldsymbol{F}}(\tau, \xi(\tau)), \quad \xi(t_0) = \boldsymbol{x}_0. \qquad \bullet$$

As we have already discussed, for general ordinary difference equations, solutions are not defined for times *smaller* than the initial time. The following simple example illustrates this.

3.4.4 Example (Ordinary difference equation with restricted domain of definition) Let $X \subseteq \mathbb{R}^m$ be an open set, let $\mathbb{T} \subseteq \mathbb{Z}(h)$ be a time-domain, and let $\boldsymbol{F}$ be the first-order ordinary difference equation with right-hand side

$$\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) = \bar{\boldsymbol{x}}, \qquad (t, \boldsymbol{x}) \in \mathbb{T} \times X,$$

for some $\bar{\boldsymbol{x}} \in X$. Note that, for any initial value problem

$$\xi(t+h) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x}_0,$$

the solution satisfies $\xi(t) = \bar{\boldsymbol{x}}$ for $t \in \mathbb{T}_{>t_0}$. Therefore, if $\boldsymbol{x}_0 \neq \bar{\boldsymbol{x}}$, it is not possible for the solution to the initial value problem to be defined for $t < t_0$. Indeed, suppose such a solution is defined at $t_0 - h$. Then we must have

$$\xi(t_0) = \widehat{\boldsymbol{F}}(t_0 - h, \xi(t_0 - h)) = \boldsymbol{x}_0 \neq \bar{\boldsymbol{x}}. \qquad \bullet$$

The example motivates the introduction of the following special and important class of ordinary difference equations.

3.4.5 Definition (Invertible ordinary difference equation) An ordinary difference equation F with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^m$$

is ***invertible*** if it is complete and if, for every $t \in \mathbb{T}$, the mapping $x \mapsto \widehat{F}(t, x)$ is a bijection. •

The importance of invertible ordinary difference equations is given to us by the following result.

3.4.6 Theorem (Existence and uniqueness of solutions for invertible ordinary difference equations) *Let $X \subseteq \mathbb{R}^m$ be open, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $\mathbf{F}$ be an invertible first-order ordinary difference equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times X \to X.$$

Then, for each $(t_0, x_0) \in \mathbb{T}_{\mathbf{F}} \times X$, there exists a unique solution $\xi\colon \mathbb{T} \to X$ for $\mathbf{F}$ such that $\xi(t_0) = x_0$. In particular, $D_{\mathbf{F}} = \mathbb{T} \times \mathbb{T} \times X$.

Proof Let

$$\widehat{F}^{-1}\colon \mathbb{T} \times X \to X$$

be defined by

$$\widehat{F}^{-1}(t, \widehat{F}(t, x)) = \widehat{F}(t, \widehat{F}^{-1}(t, x)) = x, \qquad (t, x) \in \mathbb{T} \times X.$$

Define $\xi\colon \mathbb{T}_{\geq t_0} \to X$ recursively by

$$\xi(t_0) = x_0, \quad \xi(t_0 + kh) = \widehat{F}(t_0 + kh, \xi(t_0 + (k-1)h)),$$

for $t \geq t_0$ and by

$$\xi(t_0) = x_0, \quad \xi(t_0 - kh) = \widehat{F}^{-1}(t_0 - kh, \xi(t_0 - (k-1)h)),$$

for $t \leq t_0$. Clearly this construction is well-defined on $\mathbb{T}$, and gives the existence assertion. Moreover, by its very construction, ξ is the *only* solution to the initial value problem. ■

Let us give some properties of flows that follow directly from the definition.

3.4.7 Proposition (Elementary properties of flow for ordinary difference equations) *Let $\mathbf{F}$ be an ordinary difference equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times X \to \mathbb{R}^m.$$

Then the following statements hold:

(i) for each $(t_0, x_0) \in \mathbb{T} \times X$, $(t_0, t_0, x_0) \in D_{\mathbf{F}}$ and $\Phi^{\mathbf{F}}(t_0, t_0, x_0) = x_0$;

(ii) *if, for* $t_1, t_2 \in \mathbb{T}$ *with* $t_1 \leq t_2$, $(t_2, t_1, \mathbf{x}) \in \mathrm{D}_{\mathbf{F}}$, *then, for* $t_3 \in \mathbb{T}$ *with* $t_2 \leq t_3$, $(t_3, t_2, \Phi^{\mathbf{F}}(t_2, t_1, \mathbf{x})) \in \mathrm{D}_{\mathbf{F}}$ *if and only if* $(t_3, t_1, \mathbf{x}) \in \mathrm{D}_{\mathbf{F}}$ *and, if this holds, then*

$$\Phi^{\mathbf{F}}(t_3, t_1, \mathbf{x}) = \Phi^{\mathbf{F}}(t_3, t_2, \Phi^{\mathbf{F}}(t_2, t_1, \mathbf{x})).$$

(iii) *if* $\mathbf{F}$ *is invertible and if* $(t_2, t_1, \mathbf{x}) \in \mathrm{D}_{\mathbf{F}}$, *then* $(t_1, t_2, \Phi^{\mathbf{F}}(t_2, t_1, \mathbf{x})) \in \mathrm{D}_{\mathbf{F}}$ *and* $\Phi^{\mathbf{F}}(t_1, t_2, \Phi^{\mathbf{F}}(t_2, t_1, \mathbf{x})) = \mathbf{x}$.

Proof The proof mirrors that of Proposition 3.2.12, with suitable notational modifications. ■

An important facet of flows for ordinary differential equations is their regularity, and great effort was devoted to proving this regularity in the proof of Theorem 3.2.13. For ordinary difference equations, we have the following simple result concerning the regularity of flows.

3.4.8 Theorem (Properties of flows of ordinary difference equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times \mathrm{X} \to \mathbb{R}^{\mathrm{m}}.$$

If $\widehat{\mathbf{F}}$ *is continuous, then the following statements hold:*

(i) *for* $(t_0, \mathbf{x}_0) \in \mathbb{T} \times \mathrm{X}$, $\mathrm{J}_{\mathbf{F}}(t_0, \mathbf{x}_0)$ *is a sub-time-domain of* $\mathbb{T}$;

(ii) *for* $(t_0, \mathbf{x}_0) \in \mathbb{T} \times \mathrm{X}$, *the curve*

$$\begin{aligned} \gamma_{(t_0, \mathbf{x}_0)} \colon \mathrm{J}_{\mathbf{F}}(t_0, \mathbf{x}_0) &\to \mathbb{R}^{\mathrm{m}} \\ t &\mapsto \Phi^{\mathbf{F}}(t, t_0, \mathbf{x}_0) \end{aligned}$$

is well-defined and continuous;

(iii) *for* $t, t_0 \in \mathbb{T}$ *with* $t \geq t_0$, $\mathrm{D}_\Sigma(t, t_0)$ *is open;*

(iv) *for* $t, t_0 \in \mathbb{T}$ *with* $t \geq t_0$ *for which* $\mathrm{D}_\Sigma(t, t_0) \neq \emptyset$, Φ^Σ_{t,t_0} *is continuous;*

(v) *for* $t_0 \in \mathbb{T}$, $\mathrm{D}_\Sigma(t_0)$ *is relatively open in* $\mathbb{T} \times \mathrm{X} \times \mathscr{U}$;

(vi) *for* $t_0 \in \mathbb{T}$, *the map*

$$\begin{aligned} \Phi^\Sigma(t_0) \colon \mathrm{D}_\Sigma(t_0) &\to \mathrm{X} \\ (t, \mathbf{x}) &\mapsto \Phi^\Sigma(t, t_0, \mathbf{x}) \end{aligned}$$

is well-defined and continuous;

(vii) D_Σ *is relatively open in* $\mathbb{T} \times \mathbb{T} \times \mathrm{X} \times \mathscr{U}$;

(viii) *the map*

$$\Phi^\Sigma \colon \mathrm{D}_\Sigma \to \mathrm{X}$$

is continuous;

(ix) *for* $(t_0, \mathbf{x}_0) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$ *and for* $\epsilon \in \mathbb{R}_{>0}$, *there exists* $r, \rho \in \mathbb{R}_{>0}$ *such that*

$$\sup \mathrm{J}_\Sigma(t_0, \mathbf{x}) > \sup \mathrm{J}_\Sigma(t_0, \mathbf{x}_0) - \epsilon, \quad \inf \mathrm{J}_\Sigma(t_0, \mathbf{x}) < \inf \mathrm{J}_\Sigma(t_0, \mathbf{x}_0) + \epsilon,$$

for all $\mathbf{x} \in \mathrm{B}(r, \mathbf{x}_0)$.

continuous maps for discrete topology

Proof Parts (i) and (ii) follow since $\mathbb{T}$ is discrete, and so all functions from any subset of $\mathbb{T}$ are continuous ().

(iii) Let $x \in D_F(t, t_0)$. Thus

$$\Phi^F(t, t_0, x) \in X.$$

Since

$$\Phi^F(t, t_0, x) = \Phi^F_{t,t-h} \circ \cdots \circ \Phi^F_{t_0+h,t_0}(x)$$

and since

$$\Phi^F_{\tau+h,\tau}(y) = \widehat{F}(\tau, y), \qquad \tau \in \mathbb{T}_F,\ y \in X,$$

and we thus conclude that

$$x \mapsto \Phi^F(t, t_0, x)$$

is continuous. Combining this with openness of X, we conclude that there is a neighbourhood of x that maps to X, giving the desired conclusion.

(iv) This we proved in the preceding part of the proof.

(v) Let $(t, x) \in D_F(t_0)$. Thus

$$\Phi^F(t, t_0, x) \in X.$$

For $\tau \in \mathbb{T}$, define

$$\begin{aligned} \Phi_\tau \colon X &\to X \\ y &\mapsto \widehat{F}(\tau, x) \end{aligned}$$

so that

$$\Phi^F(\tau + h, \tau, x) = \Phi_\tau(x).$$

Thus

$$\begin{aligned} \Phi^F(t_0 + h, t_0, x) &= \Phi_{t_0}(x), \\ \Phi^F(t_0 + 2h, t_0, x) &= \Phi_{t_0+h} \circ \Phi_{t_0}(x), \\ &\vdots \\ \Phi^F(t, t_0, x) &= \Phi_{t-h} \circ \Phi_{t-2h} \circ \cdots \circ \Phi_{t_0+h} \circ \Phi_{t_0}(x). \end{aligned}$$

This shows that

$$x \mapsto \Phi^F(t, t_0, x)$$

is continuous by continuity of $\widehat{F}$. Now, openness of X gives a neighbourhood N of x to X. This gives the neighbourhood $\{t\} \times N$ in $D_\Sigma(t_0)$ that maps to X, keeping in mind that the topology on $\mathbb{T}$ is the discrete topology.

(vi) This was proved in the preceding part of the proof.

(vii) The proof here can be carried out as was the proof of part (v).

(viii) This follows from part (vii) in the same manner as part (vi) follows from part (v).

(ix) In this discrete-time case, the assertion will follow if we can show that, for $t, t_0 \in \mathbb{T}$ with $t \geq t_0$ and for $x \in D_\Sigma(t, t_0)$, there is a neighbourhood N of x in X such that, if $x' \in N$, then $x' \in D_\Sigma(t, t_0)$. This, however, follows from part (v). ■

For ordinary differential equations in Theorem 3.2.13 (and in Section 5.1.1.4 below), we devote much effort to determining the manner of dependence on initial conditions. For ordinary difference equations, these questions are trivial since, by the very definition of the flow of an ordinary difference equation, the regularity of

$$D_F(t, t_0) \ni \boldsymbol{x} \mapsto \Phi^F(t, t_0, \boldsymbol{x}) \in \mathbb{R}^m$$

is exactly determined by regularity of the right-hand side

$$X \ni \boldsymbol{x} \mapsto \widehat{\boldsymbol{F}}(t, \boldsymbol{x}) \in \mathbb{R}^m,$$

provided the sort of regularity is preserved by composition, e.g., continuity, differentiability, being an homeomorphism, etc.

3.4.2 (Lack of) results for partial difference equations

Much of the discussion for partial differential equations from Section 3.2.2 carries over to partial difference equations. What one does not have for partial difference equations is the subtleties concerning regularity that often arise in the study of partial differential equations. Nonetheless, matters of existence and uniqueness in any general setting for partial difference equations are not possible. We shall not, therefore, engage in any deep discussion of these matters here.

Exercises

3.4.1

3.4.2 Let

$$\boldsymbol{F} \colon \mathbb{T} \times X \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

be an ordinary difference equation with right-hand side $\widehat{\boldsymbol{F}}$. Suppose that $\mathbb{T} \subseteq \mathbb{Z}(h)$. As per Exercise 3.3.7, let $\boldsymbol{F}_1$ be the associated first-order ordinary difference equation with state space

$$X_1 = X \times \underbrace{\mathbb{R}^m \times \cdots \times \mathbb{R}^m}_{k-1 \text{ times}}.$$

Answer the following questions.

(a) Formulate the notion of an initial value problem for the kth-order ordinary difference equation $\boldsymbol{F}$.

(b) State *precisely* the relationship between the initial conditions for the initial value problem from part (a) and the initial value problem for $\boldsymbol{F}_1$ from Definition 3.4.1.

3.4.3 Consider the partial difference equation

$$\boldsymbol{F} \colon \mathbb{Z}^3 \times \mathbb{R} \times \mathbb{R}^{\otimes,1} \to \mathbb{R}^3$$

given by

$$\boldsymbol{F}((x_1, x_2, x_3), u, (u_1^+, u_2^+, u_3^+, u_1^-, u_2^-, u_3^-)) = (u_1^- - f_1(\boldsymbol{x}), u_2^- - f_2(\boldsymbol{x}), u_3^- - f_3(\boldsymbol{x})).$$

Show that if $\boldsymbol{F}$ has a solution u, then

$$\Delta_j^- f_k(\boldsymbol{x}) = \Delta_k^- f_j(\boldsymbol{x}), \qquad j, k \in \{1, 2, 3\},\ \boldsymbol{x} \in \mathbb{Z}^3.$$

Chapter 4

Scalar ordinary differential and ordinary difference equations

In this chapter, we begin our studies in earnest, doing what one does with differential and difference equations: where possible, solve them and/or understand the nature of their solutions or sets of solutions. We shall study ordinary differential and difference equations with a single state and arbitrary order.

For differential equations, in the notation of Section 3.1.3, we consider an ordinary differential equation with time domain $\mathbb{T} \subseteq \mathbb{R}$, state space $U \subseteq \mathbb{R}$, and with right-hand side

$$\widehat{F} \colon \mathbb{T} \times U \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$

that gives an equation

$$\frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t) = \widehat{F}\left(t, \xi(t), \frac{\mathrm{d}\xi}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t)\right)$$

that must be satisfied by solutions $t \mapsto \xi(t)$.

For difference equations, in the notation of Section 3.3.3, we consider an ordinary difference equation with time domain $\mathbb{T} \subseteq \mathbb{Z}(h)$, state space $U \subseteq \mathbb{R}$, and with right-hand side

$$\widehat{F} \colon \mathbb{T} \times U \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$

that gives an equation

$$\xi(t) = \widehat{F}\,(t, \xi(t-h), \dots, \xi(t-kh))$$

that must be satisfied by solutions $t \mapsto \xi(t)$.

There is not much one can say in any generality about such equations, except to say that we can use Theorems 3.2.8 and 3.4.2 to assert the existence and uniqueness of solutions, at least for small times (making use of Exercises 3.1.23 and 3.3.7). Thus we focus in this chapter on special equations for which one *can* say something useful. In Section 4.1 we consider very special classes of first-order differential and difference equations that can, in some sense, be solved. In Sections 4.2 and 4.3 we consider linear differential equations, first homogeneous equations then inhomogeneous equations. The results in these sections are echoed for difference

equations in Sections 4.6 and 4.7. In Section 4.4 we consider an important aspect of the subject of differential equations, where distributions play a fundamental rôle.

Do I need to read this chapter? Many of the results and techniques in this chapter are prerequisite for our treatments of system theory, particularly in Sections 6.6, 6.8, 6.7, and 6.9. •

Contents

Section 4.1

General first-order scalar ordinary differential and difference equations

In this section we study the simplest sort of differential and difference equations, those of first-order with a scalar state. Even for these very simple systems, there are limits on what one can say. However, there are some interesting and useful special classes of these equations for which some techniques exist to obtain formulae for solutions.

Do I need to read this section? The techniques in this section are useful, when they can be applied. They should, therefore, be regarded as an essential part of the toolkit for any mathematician or scientist. •

4.1.1 First-order scalar ordinary differential equations

In this short section we consider a very special class of first-order scalar differential equation, one that can sometimes be solved explicitly. The following definition encodes what we are after.

4.1.1 Definition (Separable scalar differential equation) A differential equation $F\colon \mathbb{T}\times U\times \mathrm{L}^1_{\mathrm{sym}}(\mathbb{R};\mathbb{R})\to\mathbb{R}$ is ***separable*** if it has the form

$$\boldsymbol{F}(t,x,x^{(1)}) = f_1(x)x^{(1)} - f_0(t).$$ •

We note that a separable differential equation is an ordinary differential equation if and only if $f_1(x)$ is nonzero for every $x\in U$, because in this case we can solve for $x^{(1)}$ for a given $(t,x)\in\mathbb{T}\times U$ by

$$x^{(1)} = \frac{f_0(t)}{f_1(x)} = \widehat{F}(t,x).$$

Note that $t\mapsto x(t)$ is a solution to a separable differential equation if

$$f_1(x(t))\frac{\mathrm{d}x}{\mathrm{d}t}(t) = f_0(t), \qquad x(t_0)=x_0$$

for some $(t_0,x_0)\in\mathbb{T}\times U$. There is a naïve way to "solve" such an equation. First do some (*a priori* meaningless) manipulations:

$$f_1(x)\frac{\mathrm{d}x}{\mathrm{d}t} = f_0(t) \implies \int_{x_0}^{x(t)} f_1(\xi)\,\mathrm{d}\xi = \int_{t_0}^{t} f_0(\tau)\,\mathrm{d}\tau.$$

If F_1 and F_0 are antiderivatives of f_1 and f_0, respectively, we have

$$F_1(x(t)) - F_1(x_0) = F_0(t) - F_0(t_0).$$

This is an equation that you pray you can solve for $x(t)$.

This naïve procedure does, in fact, work, as the following result indicates.

4.1.2 Proposition (Solutions for separable differential equations) *Let $\mathbb{T} \subseteq \mathbb{R}$ be a time-domain, let $U \subseteq \mathbb{R}$ be an open set, let $f_0 \colon \mathbb{T} \to \mathbb{R}$ and $f_1 \colon U \to \mathbb{R}$ be continuous functions for which $f_1(x) \neq 0$ for every $x \in U$. Let F_0 and F_1 be antiderivatives of f_0 and f_1, respectively. Let $(t_0, x_0) \in \mathbb{T} \times U$. Then the following statements hold:*

(i) if $\mathbb{T}' \subseteq \mathbb{T}$ is a subinterval containing t_0 and if a class C^1-function $\xi \colon \mathbb{T}' \to U$ satisfies

$$F_1(\xi(t)) - F_1(x_0) = F_0(t) - F_0(t_0), \qquad t \in \mathbb{T}',$$

then ξ is a solution to the separable ordinary differential equation

$$F(t, x, x^{(1)}) = f_1(x)x^{(1)} - f_0(t)$$

satisfying the initial condition $\xi(t_0) = x_0$;

(ii) if there exists a subinterval $\mathbb{T}' \subseteq \mathbb{T}$ and a solution $\xi \colon \mathbb{T}' \to U$ to F satisfying $\xi(t_0) = x_0$, then

$$F_1(\xi(t)) - F_1(x_0) = F_0(t) - F_0(t_0), \qquad t \in \mathbb{T}'.$$

Proof (i) Let us define

$$G \colon \mathbb{T} \times U \to \mathbb{R}$$

by

$$G(t, x) = F_1(x) - F_1(x_0) - F_0(t) + F_0(t_0),$$

noting that $G(t, \xi(t)) = 0$. Note that G is of class C^1 and that

$$\frac{\partial G}{\partial x}(t, \xi(t)) \neq 0, \qquad t \in \mathbb{T}'$$

Thus, by the Implicit Function Theorem, there exists a relatively open interval $\mathbb{T}'_t \subseteq \mathbb{T}'$ ref containing t and a unique map $\xi_t \colon \mathbb{T} \to U$ of class C^1 such that $\xi_t(t) = \xi(t)$ and that $G(\tau, \xi_t(\tau)) = 0$ for all $\tau \in \mathbb{T}'_t$. Therefore, by the Chain Rule,

$$0 = \frac{\mathrm{d}}{\mathrm{d}\tau} G(\tau, \xi_t(\tau)) = \frac{\mathrm{d}}{\mathrm{d}\tau}(F_1(\xi_t(\tau)) - F_1(x_0) - F_0(\tau) + F_0(t_0)) = f_1(\xi_t(\tau))\dot{\xi}_t(\tau) - f_0(\tau),$$

giving ξ_t as a solution to $\boldsymbol{F}$.

It remains to show that $\xi(\tau) = \xi_t(\tau)$ for every $t \in \mathbb{T}'$ and every $\tau \in \mathbb{T}'_t$. Let $\mathbb{T}'' \subset \mathbb{T}'$ be the largest subinterval such that $\xi(\tau) = \xi_t(\tau)$ for every $t \in \mathbb{T}''$ and every $\tau \in \mathbb{T}'_t$. We claim that $\mathbb{T}'' = \mathbb{T}'$. We need only show that $\mathbb{T}' \subseteq \mathbb{T}''$. Let $t \in \mathbb{T}'$. By construction, we have $\xi_t(t) = \xi(t)$. Note that, for every $\tau \in \mathbb{T}'_t$ we have $G(\tau, \xi(\tau)) = 0$. Moreover, $\xi|\mathbb{T}'_t$ is of class C^1. Thus the uniqueness part of the Implicit Function Theorem gives $\xi_t(\tau) = \xi(\tau)$ for all $\tau \in \mathbb{T}'_t$. Therefore, $t \in \mathbb{T}''$. From this we conclude that, indeed $\xi(\tau) = \xi_t(\tau)$ for every $\tau \in \mathbb{T}'_t$, and this shows that ξ is a solution for F, since ξ_t is a solution for F.

(ii) We have, for all $t \in \mathbb{T}'$,

$$\begin{aligned} & f_1(\xi(t))\dot{\xi}(t) - f_0(t) = 0 \\ \implies \quad & \frac{\mathrm{d}}{\mathrm{d}t}(F_1(\xi(t)) - F_0(t)) = 0 \\ \implies \quad & F_1(\xi(t)) - F_1(x_0) - F_0(t) + F_0(t_0) \end{aligned}$$

since ξ is continuous, and using the Fundamental Theorem of Calculus. ■

Now let us look at some examples.

4.1.3 Examples (Separable ordinary differential equations)

1. Consider the ordinary differential equation

$$F(t, x, x^{(1)}) = x^{(1)} - ax$$

for $a \in \mathbb{R}$, which is defined for $(t, x) \in \mathbb{R} \times \mathbb{R}$, i.e., $\mathbb{T} = \mathbb{R}$ and $U = \mathbb{R}$. Solutions of this differential equation satisfy

$$\dot{x}(t) = ax(t).$$

This is not immediately in the form of a separable equation, but it can be converted into the separable equation

$$\tilde{F}(t, x, x^{(1)}) = \frac{x^{(1)}}{x} - a,$$

but only at the cost of limiting the state space to be $\tilde{U} = \mathbb{R} \setminus \{0\}$. But let us do this and see what happens. We have $f_1(x) = x^{-1}$ and $f_0(t) = a$ and so $F_1(x) = \ln(|x|)$ and $F_0(t) = at$. Thus, by Proposition 4.1.2, a solution $t \mapsto \xi(t)$ with values in $\tilde{U}$ will satisfy

$$\begin{aligned} & \ln(|\xi(t)|) - \ln(|\xi(t_0)|) = a(t - t_0) \\ \iff \quad & \ln\left(\left|\frac{\xi(t)}{\xi(t_0)}\right|\right) = a(t - t_0) \\ \iff \quad & \left|\frac{\xi(t)}{\xi(t_0)}\right| = \mathrm{e}^{a(t-t_0)} \\ \iff \quad & |\xi(t)| = |\xi(t_0)|\mathrm{e}^{a(t-t_0)}. \end{aligned}$$

Now, since ξ must be of class C^1, in particular continuous, it follows that the sign of $\xi(t)$ must be the same as that of $\xi(t_0)$, and so we have

$$\xi(t) = \xi(t_0)\mathrm{e}^{a(t-t_0)}.$$

Note that this only applies when $\xi(t_0) \neq 0$. However, if $\xi(t_0) = 0$ then we immediately have the solution as $\xi(t) = 0$ for all t.
We will encounter this differential equation as a special case of various other sorts of differential equations in the sequel.

2. Next we consider the differential equation

$$F(t, x, x^{(1)}) = x^{(1)} - x^2$$

with $(t, x) \in \mathbb{R} \times \mathbb{R}$ that we initially investigated in Example 3.2.5. Again, this equation is not in the form of a separable ordinary differential equation, but can be converted into the separable equation

$$\tilde{F}(t, x, x^{(1)}) = \frac{x^{(1)}}{x^2} - 1$$

with $f_0(x) = x^{-2}$ and $f_1(t) = 1$. Again, in making this conversion, we must restrict our state to be in $\tilde{U} = \mathbb{R} \setminus \{0\}$. We then have

$$F_1(x) = -x^{-1}, \quad F_0(t) = t.$$

Therefore, skipping the details, a solution $t \mapsto \xi(t)$ satisfies

$$-\frac{1}{\xi(t)} + \frac{1}{\xi(t_0)} = t - t_0 \quad \Longrightarrow \quad \xi(t) = \frac{\xi(0)}{\xi(t_0)(t_0 - t) + 1},$$

just as in Example 3.2.5. As we saw in this previous example, the solution cannot be defined on the entire time interval $\mathbb{R}$. Also, we can recover the solution with the initial condition $\xi(t_0) = 0$ by noting that, in this case, the solution is $\xi(t) = 0$.

3. Here we consider the differential equation

$$F(t, x, x^{(1)}) = x^{(1)} - x^{1/3}$$

first considered in Example 3.2.6. As with our other examples, this one is not separable by can be converted to a separable equation on the reduced state space $U' = \mathbb{R} \setminus \{0\}$:

$$\tilde{F}(t, x, x^{(1)}) = \frac{x^{(1)}}{x^{1/3}} - 1.$$

We then have

$$F_1(x) = \frac{3x^{2/3}}{2}, \quad F_0(t) = t$$

and so solutions $t \mapsto \xi(t)$ are determined by

$$\frac{3\xi(t)^{2/3}}{2} - \frac{3\xi(t_0)^{2/3}}{2} = t - t_0 \quad \Longrightarrow \quad \xi(t) = \frac{(2t - 2t_0 + 3\xi(t_0)^{2/3})^{3/2}}{3\sqrt{3}}.$$

Again, if we include the possibility that $\xi(t_0) = 0$, we arrive at the situation described in Example 3.2.6.

4. Finally, we consider the separable ordinary differential equation

$$F(t, x, x^{(1)}) = (x^4 + x^2 + 1)x^{(1)} - \mathrm{e}^{-t^2}$$

with $f_1(x) = x^4 + x^2 + 1$ and $f_0(t) = \mathrm{e}^{-t^2}$ with $(t, x) \in \mathbb{R} \times \mathbb{R}$. Here we have

$$F_1(x) = \frac{x^5}{5} + \frac{x^3}{3} + x, \quad F_0(t) = \frac{\sqrt{\pi}}{2} \operatorname{erf}(t),$$

where erf is the ***error function*** defined by

$$\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t \mathrm{e}^{-\tau^2} \, \mathrm{d}\tau.$$

Thus a solution $t \mapsto \xi(t)$ satisfies

$$\frac{\xi(t)^5}{5} + \frac{\xi(t)^3}{3} + \xi(t) - \frac{\xi(t_0)^5}{5} - \frac{\xi(t_0)^3}{3} - \xi(t_0) = \frac{\sqrt{\pi}}{2}(\operatorname{erf}(t) - \operatorname{erf}(t_0)).$$

This is an implicit equation that will be unpleasant to solve. Note that one might have five possible solutions for $\xi(t)$ at a given time, since we have the solution as the root of a fifth-order polynomial. •

4.1.2 First-order scalar ordinary difference equations

Techniques analogous to those in the preceding section for differential equations are less interesting for difference equations. Indeed, the governing equation for a difference equation analogous to a separable differential equation is

$$f_1(\xi((j-1)h))\Delta^- \xi(j) = f_0(j), \quad \xi(k_0 h) = x_0,$$

with f_1 being nowhere zero. Using the definition of Δ^- this gives

$$\xi(jh) = \xi((j-1)h) + h \frac{f_0(j)}{f_1((j-1)h)}.$$

This can be immediately solved to give

$$\xi(kh) = x_0 + h \sum_{j=k_0}^{k-1} \frac{f_0(jh)}{f_1((j-1)h)},$$

which is not all that interesting, and makes no particular use of the "separation" property of the equation.

In Example 4.7.5 we shall consider general linear scalar first-order difference equations.

Exercises

4.1.1 Solve the following initial value problems, taking care to provide the domain of definition for the solution:

(a) $t\dot{\xi}(t) = 2(\xi(t) - 4)$, $\xi(1) = 5$;

(b) $(t^2 + 1)\dot{\xi}(t) = t\xi(t)$, $\xi(0) = 1$;

(c) $\dot{\xi}(t) = \xi(t)\tan(t)$, $\xi(0) = 1$;

(d) $\dot{\xi}(t) = t\xi(t) + 2t + \xi(t) + 2$, $\xi(0) = -1$.

4.1.2 Solve the following initial value problems, taking care to provide the domain of definition for the solution:

(a) $\dot{\xi}(t) + t\xi(t) = t$, $\xi(1) = 5$;

(b) $t\dot{\xi}(t) + \xi(t) = t + 1$, $\xi(1) = 0$;

(c) $\dot{\xi}(t) + \mathrm{e}^t\xi(t) = \mathrm{e}^t$, $\xi(0) = x_0$;

(d) $(1 + t)\dot{\xi}(t) + \tan(t)\xi(t) = \sec(t)$, $\xi(\frac{\pi}{4}) = 0$.

Section 4.2

Scalar linear homogeneous ordinary differential equations

Now we turn to scalar linear ordinary differential equations, looking first in this section at the homogeneous case. That is to say, we consider differential equations with $\mathbb{T} \subseteq \mathbb{R}$ an interval, the state space $U = \mathbb{R}$, and right-hand sides of the form

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x \tag{4.1}$$

for functions $a_0, a_1, \dots, a_{k-1} \colon \mathbb{T} \to \mathbb{R}$. Thus solutions $t \mapsto \xi(t)$ satisfy

$$\frac{\mathrm{d}^k \xi(t)}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t)\xi(t) = 0.$$

In this section we shall (1) investigate the character of the solutions, (2) investigate the set of all solutions in the general case, and (3) provide a procedure for, in principle, solving the equations in the constant coefficient case.

Do I need to read this section? This section contains tools that are standard for anyone claiming to know something about ordinary differential equations. •

4.2.1 Equations with time-varying coefficients

We start by working with the general situation where the coefficients $a_0, a_1, \dots, a_{k-1}$ depend on time. In this case, we will study the properties of solutions and sets of solutions, and as well introduce an important tool, the "Wronskian," for dealing with linear ordinary differential equations.

4.2.1.1 Solutions and their properties We begin by listing the general properties of solutions. First let us be sure that the equations with which we are dealing possess solutions.

4.2.1 Proposition (Local existence and uniqueness of solutions for scalar linear homogeneous ordinary differential equations) *Consider the linear homogeneous ordinary differential equation* F *with right-hand side (4.1) and suppose that* $\mathrm{a}_0, \mathrm{a}_1, \dots, \mathrm{a}_{k-1} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. *Let*

$$(\mathrm{t}_0, \mathrm{x}_0, \mathrm{x}_0^{(1)}, \dots, \mathrm{x}_0^{(k-1)}) \in \mathbb{T} \times \mathbb{R} \oplus \mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}).$$

Then there exists an interval $\mathbb{T}' \subseteq \mathbb{T}$ *and a map* $\xi \colon \mathbb{T}' \to \mathbb{R}$ *of class* C^{k-1}, *with locally absolutely continuous* $(k-1)$*st derivative, that is a solution for* F *and which satisfies*

$$\xi(\mathrm{t}_0) = \mathrm{x}_0, \frac{\mathrm{d}\xi}{\mathrm{d}t}(\mathrm{t}_0) = \mathrm{x}_0^{(1)}, \dots, \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(\mathrm{t}_0) = \mathrm{x}_0^{(k-1)}.$$

Moreover, if $\tilde{\mathbb{T}}' \subseteq \mathbb{T}$ *is another subinterval and if* $\tilde{\xi}\colon \tilde{\mathbb{T}}' \to \mathbb{R}$ *is another* C^{k-1}*-solution, with locally absolutely continuous* $(k-1)$*st derivative, for* F *satisfying*

$$\tilde{\xi}(t_0) = x_0, \frac{d\tilde{\xi}}{dt}(t_0) = x_0^{(1)}, \dots, \frac{d^{k-1}\tilde{\xi}}{dt^{k-1}}(t)(t_0) = x_0^{(k-1)},$$

then $\tilde{\xi}(t) = \xi(t)$ *for every* $t \in \tilde{\mathbb{T}}' \cap \mathbb{T}'$.

Proof This is Exercise 4.2.1. ■

The proposition indicates the importance of the following class of functions for a continuous time-domain $\mathbb{T}$ and for $r \in \mathbb{Z}_{\geq 0}$:

$$\mathsf{AC}^r_{\mathrm{loc}}(\mathbb{T};\mathbb{R}) = \{f \in \mathsf{C}^r(\mathbb{T};\mathbb{R}) \mid f^{(r)} \text{ is locally absolutely continuous}\}.$$

As we have seen in Example 3.2.5, a solution to a general ordinary differential equation will not be defined for all times in $\mathbb{T}$, even for seemingly "nice" differential equations. One might then wonder whether linear ordinary differential equations are sufficiently nice to permit solutions defined for all time. This is, indeed, the case.

4.2.2 Proposition (Global existence of solutions for scalar linear homogeneous ordinary differential equations) *Consider the linear homogeneous ordinary differential equation* F *with right-hand side (4.1) and suppose that* $a_0, a_1, \dots, a_{k-1} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathbb{R})$. *If* $\xi \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T}';\mathbb{R})$ *is a solution for* F*, then there exists a solution* $\overline{\xi} \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T};\mathbb{R})$ *for which* $\overline{\xi}|\mathbb{T}' = \xi$.

Proof Note that, as per Exercise 3.1.23, we can convert the differential equation F into a first-order differential equation linear homogeneous differential equation with states $(x, x^{(1)}, \dots, x^{(k-1)})$. Thus the result will follow from the analogous result for first-order systems of equations, and this is stated and proved as Proposition 5.2.2. ■

Now that we know the domain of definition of a scalar linear homogeneous ordinary differential equation, we can talk in a reasonable manner about the set of *all* solutions of such equations, as the structure of these is what is most interesting about the equations. Thus we consider a scalar linear homogeneous ordinary differential equation

$$F\colon \mathbb{T} \times \mathbb{R} \oplus \mathsf{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) \to \mathbb{R}$$

with right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x,$$

where $a_0, a_1, \dots, a_{k-1} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathbb{R})$. Let us denote by

$$\mathrm{Sol}(F) = \left\{ \xi \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T};\mathbb{R}) \,\middle|\, \frac{d^k\xi}{dt^k}(t) + a_{k-1}(t)\frac{d^{k-1}\xi}{dt^{k-1}}(t) + \cdots + a_1(t)\frac{d\xi}{dt}(t) + a_0(t)\xi(t) = 0, \text{ a.e. } t \in \mathbb{T} \right\}$$

the set of solutions for F. The following result is then the main structural result for the class of differential equations we are considering in this section. We note that $\mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T};\mathbb{R})$ is a $\mathbb{R}$-vector space by virtue of Propositions I-3.2.10 and III-2.9.29.

4.2.3 Theorem (Vector space structure of sets of solutions) *Consider the linear homogeneous ordinary differential equation* F *with right-hand side (4.1) and suppose that the functions* $\mathrm{a}_0, \mathrm{a}_1, \dots, \mathrm{a}_{k-1} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathbb{R})$. *Then* Sol(F) *is a* k*-dimensional subspace of* $\mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T};\mathbb{R})$.

Proof We first show that Sol(F) is a subspace. Let $\xi, \xi_1.\xi_2 \in \mathrm{Sol}(F)$ and $\alpha \in \mathbb{R}$. Then we immediately have

$$\begin{aligned}\frac{\mathrm{d}^k(\xi_1+\xi_2)}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1})\xi_1+\xi_2}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}(\xi_1+\xi_2)}{\mathrm{d}t}(t) + a_0(t)(\xi_1+\xi_2)(t)\\ = \frac{\mathrm{d}^k\xi_1}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi_1}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi_1}{\mathrm{d}t}(t) + a_0(t)\xi_1(t)\\ + \frac{\mathrm{d}^k\xi_2}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi_2}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi_2}{\mathrm{d}t}(t) + a_0(t)\xi_2(t) = 0+0=0\end{aligned}$$

and

$$\begin{aligned}\frac{\mathrm{d}^k(\alpha\xi)}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}(\alpha\xi)}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}(\alpha\xi)}{\mathrm{d}t}(t) + a_0(t)(\alpha\xi)(t)\\ = \alpha\left(\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t)\xi(t)\right) = 0,\end{aligned}$$

using linearity of differentiation.

Next we prove that the dimension of Sol(F) is k. We shall do this by showing that, for a given $t_0 \in \mathbb{T}$, the map

$$\begin{aligned}\sigma_{t_0} : \mathrm{Sol}(F) &\to \mathbb{R}^k\\ \xi &\mapsto \left(\xi(t_0), \frac{\mathrm{d}\xi}{\mathrm{d}t}(t_0), \dots, \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t_0)\right)\end{aligned}$$

is an isomorphism of $\mathbb{R}$-vector spaces. Since the map is surjective by the existence part of Proposition 4.2.1, it suffices to show that it is an injective linear map. Linearity of σ_{t_0} is immediate since the identities

$$\begin{aligned}\left((\xi_1+\xi_2)(t_0), \frac{\mathrm{d}(\xi_1+\xi_2)}{\mathrm{d}t}(t_0), \dots, \frac{\mathrm{d}^{k-1}(\xi_1+\xi_2)}{\mathrm{d}t^{k-1}}(t_0)\right)\\ = \left(\xi_1(t_0), \frac{\mathrm{d}\xi_1}{\mathrm{d}t}(t_0), \dots, \frac{\mathrm{d}^{k-1}\xi_1}{\mathrm{d}t^{k-1}}(t_0)\right) + \left(\xi_2(t_0), \frac{\mathrm{d}\xi_2}{\mathrm{d}t}(t_0), \dots, \frac{\mathrm{d}^{k-1}\xi_2}{\mathrm{d}t^{k-1}}(t_0)\right),\end{aligned}$$

by definition of the vector space structure for Sol(F). To show that σ_{t_0} is injective, it suffices so show that, if $\sigma_{t_0}(\xi) = \mathbf{0}$, then ξ is the zero vector in Sol(F) (by Exercise I-4.5.23) i.e., that $\xi(t) = 0$ for all $t \in \mathbb{T}$. So, suppose that $\sigma_{t_0}(\xi) = \mathbf{0}$. Then

$$\xi(t_0) = 0,\ \frac{\mathrm{d}\xi}{\mathrm{d}t}(t_0) = 0, \dots, \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t_0) = 0.$$

Consider the function $\zeta\colon \mathbb{T} \to \mathbb{R}$ given by $\zeta(t) = 0$ for all $t \in \mathbb{T}$. Then $\zeta \in \mathrm{Sol}(F)$ and

$$\zeta(t_0) = 0, \ \frac{\mathrm{d}\zeta}{\mathrm{d}t}(t_0) = 0, \dots, \frac{\mathrm{d}^{k-1}\zeta}{\mathrm{d}t^{k-1}}(t_0) = 0.$$

Therefore, by Proposition 4.2.1, $\xi = \zeta$, giving the theorem. ■

Being a finite-dimensional $\mathbb{R}$-vector space, the set $\mathrm{Sol}(F)$ of solutions to the scalar linear homogeneous differential equation F is capable of possessing a basis. One has a special name for a basis of $\mathrm{Sol}(F)$, i.e., a set of k linearly independent solutions for F.

4.2.4 Definition (Fundamental set of solutions) Consider the linear homogeneous ordinary differential equation F with right-hand side (4.1) and suppose that $a_0, a_1, \dots, a_{k-1} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. A set $\{\xi_1, \dots, \xi_k\}$ of linearly independent elements of $\mathrm{Sol}(F)$ is a ***fundamental set of solutions*** for F. •

There is not much more one can say easily, in general, about scalar linear homogeneous ordinary differential equations with coefficients that depend on time. There is, however, one case where they can be solved "explicitly," and this is when $k = 1$.

4.2.5 Example (First-order scalar linear homogeneous equations) The differential equation we consider here is given by

$$\begin{aligned} F\colon \mathbb{T} \times \mathbb{R} \oplus \mathrm{L}^1_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) &\to \mathbb{R} \\ (t, x, x^{(1)}) &\mapsto x^{(1)} + a(t)x, \end{aligned}$$

for $a \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. Thus a solution $t \mapsto \xi(t)$ satisfies

$$\dot{\xi}(t) + a(t)\xi(t) = 0.$$

Note that F is equivalent to the separable equation

$$\tilde{F}(t, x, x^{(1)}) = \frac{x^{(1)}}{x} + a(t)$$

with $f_1(x) = x^{-1}$ and $f_0(t) = -a(t)$. Thus we can apply the methods of Section 4.1.1 to solve this equation; indeed, note that Example 4.1.3–1 is a special case that we have already treated in this manner.[1] Let $t_0 \in \mathbb{T}$ and $x_0 \in \mathbb{R}$. We have the antiderivatives

$$F_1(x) = \ln(|x|) - \ln(|x_0|), \quad F_0(t) = -\int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau.$$

[1] An astute reader will note that this is not quite true since a is not continuous, but only locally integrable. Nonetheless, we shall produce the unique solution, which one can verify by substituting it into the differential equation.

In the same manner as Example 4.1.3–1, we conclude that

$$\xi(t) = \xi(t_0)\mathrm{e}^{-\int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau}.$$

Note that this solution is also valid when $\xi(t_0) = 0$, although this is not covered by this solution method, since we had to eliminate 0 from the state space to make the equation a separable equation. •

4.2.1.2 The Wronskian, and its properties and uses

In this section we present a fairly simple construction that turns out to have great importance in the treatment of linear differential equations. We first make a simple general definition that seems to not be *a priori* relating to differential equations.

4.2.6 Definition (Wronskian[2]) Let $\mathbb{T} \subseteq \mathbb{R}$ be an interval and let $f_1, \dots, f_k \in \mathrm{C}^{k-1}(\mathbb{T}; \mathbb{R})$ for $k \in \mathbb{Z}_{>0}$. The ***Wronskian*** for the functions $f_1, \dots, f_k$ is the function $\mathrm{W}(f_1, \dots, f_k)\colon \mathbb{T} \to \mathbb{R}$ defined by

$$\mathrm{W}(f_1, \dots, f_k)(t) = \det \begin{bmatrix} f_1(t) & f_2(t) & \cdots & f_k(t) \\ \dfrac{\mathrm{d}f_1}{\mathrm{d}t}(t) & \dfrac{\mathrm{d}f_2}{\mathrm{d}t}(t) & \cdots & \dfrac{\mathrm{d}f_k}{\mathrm{d}t}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\mathrm{d}^{k-1}f_1}{\mathrm{d}t^{k-1}}(t) & \dfrac{\mathrm{d}^{k-1}f_2}{\mathrm{d}t^{k-1}}(t) & \cdots & \dfrac{\mathrm{d}^{k-1}f_k}{\mathrm{d}t^{k-1}}(t) \end{bmatrix}. \quad \bullet$$

An essential feature of the Wronskian is that it gives a sufficient condition for measuring the linear independence of finite sets of functions in the space of functions. More precisely, we have the following result, which again is not *a priori* related to differential equations.

4.2.7 Proposition (The Wronskian and linear independence) *Let $\mathbb{T} \subseteq \mathbb{R}$ be an interval and let $\mathrm{f}_1, \dots, \mathrm{f}_k \in \mathrm{C}^{k-1}(\mathbb{T}; \mathbb{R})$ for $\mathrm{k} \in \mathbb{Z}_{>0}$. If $\mathrm{W}(\mathrm{f}_1, \dots, \mathrm{f}_k)(\mathrm{t}) \neq 0$ for some $\mathrm{t} \in \mathbb{T}$, then the set $\{\mathrm{f}_1, \dots, \mathrm{f}_k\}$ is linearly independent in $\mathrm{C}^{k-1}(\mathbb{T}; \mathbb{R})$.*

Proof We prove the contrapositive, i.e., that, if the functions $\{f_1, \dots, f_k\}$ are linearly dependent, then $\mathrm{W}(f_1, \dots, f_k)(t) = 0$ for all $t \in \mathbb{T}$.

So suppose that $\{f_1, \dots, f_k\}$ is linearly dependent, and let $c_1, \dots, c_k \in \mathbb{R}$, not all zero, be such that

$$c_1 f_1 + \cdots + c_k f_k = 0.$$

Then, for any $j \in \{1, \dots, k-1\}$,

$$c_1 \frac{\mathrm{d}^j f_1}{\mathrm{d}t^j} + \cdots + c_n \frac{\mathrm{d}^j f_n}{\mathrm{d}t^j} = 0.$$

[2] After Josef Hoëné de Wronski (1778–1853). Wronski was a "philosopher mathematician," and as a consequence he (1) published a lot of rubbish and (2) had a high opinion of himself. Nevertheless, he apparently had a few good days, and the Wronskian, one supposes, must be a result of one of these.

Assembling these relationships for $j \in \{0, 1, \dots, k-1\}$ gives the single equation

$$\begin{bmatrix} f_1(t) & f_2(t) & \cdots & f_k(t) \\ \frac{\mathrm{d}f_1}{\mathrm{d}t}(t) & \frac{\mathrm{d}f_2}{\mathrm{d}t}(t) & \cdots & \frac{\mathrm{d}f_k}{\mathrm{d}t}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{d}^{k-1}f_1}{\mathrm{d}t^{k-1}}(t) & \frac{\mathrm{d}^{k-1}f_2}{\mathrm{d}t^{k-1}}(t) & \cdots & \frac{\mathrm{d}^{k-1}f_k}{\mathrm{d}t^{k-1}}(t) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

This means that the matrix on the left has a nontrivial kernel (since this kernel contains $(c_1, \dots, c_k)$) and so must have zero determinant. ■

Note that the converse of the preceding result is not generally true, as demonstrated by the following example.

4.2.8 Example (The Wronskian is not adequate to characterise linear independence) Let $\mathbb{T} = [-1, 1]$ and consider the two functions $f_1, f_2 \colon [-1, 1] \to \mathbb{R}$ of class C^1 defined by

$$f_1(t) = t^2, \quad f_2(t) = t|t|.$$

We have

$$\frac{\mathrm{d}f_1}{\mathrm{d}t}(t) = 2t, \quad \frac{\mathrm{d}f_2}{\mathrm{d}t} = 2|t|$$

We thus have

$$\mathrm{W}(f_1, f_2)(t) = \det \begin{bmatrix} t^2 & t|t| \\ 2t & 2|t| \end{bmatrix} = 2t^2|t| - 2t^2|t| = 0.$$

However, the set $\{f_1, f_2\}$ *is* linearly independent. Indeed, suppose that $c_1, c_2 \in \mathbb{R}$ satisfy

$$c_1 f_1(t) + c_2 f_2(t) = 0, \qquad t \in [-1, 1].$$

Then, taking $t = -1$, we get $c_1 - c_2 = 0$ and taking $t = 1$ we get $c_1 + c_2 = 0$. The only way both of these equations can be satisfied is when $c_1 = c_2 = 0$. •

Thus the Wronskian is not quite the thing for precisely characterising the linear independence of general sets of functions. However, it is just the thing when the set of functions under consideration are solutions to a scalar linear homogeneous ordinary differential equation.

4.2.9 Proposition (Wronskians and linear independence in Sol(F)) *Consider the linear homogeneous ordinary differential equation* F *with right-hand side (4.1) and suppose that* $\mathrm{a}_0, \mathrm{a}_1, \dots, \mathrm{a}_{k-1} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. *Then the following statements are equivalent for* $\xi_1, \dots, \xi_k \in$ Sol(F)*:*

(i) $\{\xi_1, \dots, \xi_k\}$ *is linearly independent;*

(ii) $\mathrm{W}(\xi_1, \dots, \xi_k)(\mathrm{t}) \neq 0$ *for some* $\mathrm{t} \in \mathbb{T}$*;*

(iii) $\mathrm{W}(\xi_1, \dots, \xi_k)(\mathrm{t}) \neq 0$ *for all* $\mathrm{t} \in \mathbb{T}$.

Proof (i) $\implies$ (ii) We prove the contrapositive, i.e., we prove that, if $W(\xi_1, \dots, \xi_k)(t) = 0$ for all $t \in \mathbb{T}$, then $\{\xi_1, \dots, \xi_k\}$ is linearly dependent.

So suppose that $W(\xi_1, \dots, \xi_k)(t) = 0$ for all $t \in \mathbb{T}$, which means that there exists $c_1, \dots, c_k \in \mathbb{R}$, not all zero, such that

$$\begin{bmatrix} \xi_1(t) & \xi_2(t) & \cdots & \xi_k(t) \\ \frac{\mathrm{d}\xi_1}{\mathrm{d}t}(t) & \frac{\mathrm{d}\xi_2}{\mathrm{d}t}(t) & \cdots & \frac{\mathrm{d}\xi_k}{\mathrm{d}t}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{d}^{k-1}\xi_1}{\mathrm{d}t^{k-1}}(t) & \frac{\mathrm{d}^{k-1}\xi_2}{\mathrm{d}t^{k-1}}(t) & \cdots & \frac{\mathrm{d}^{k-1}\xi_k}{\mathrm{d}t^{k-1}}(t) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

for all $t \in \mathbb{T}$. If we simply expand this out, we see that it is equivalent to

$$c_1\sigma_t(\xi_1) + \cdots + c_k\sigma_t(\xi_k) = 0$$

for all $t \in \mathbb{T}$, recalling the isomorphism $\sigma_t \colon \mathrm{Sol}(F) \to \mathbb{R}^k$, defined for some $t \in \mathbb{T}$, from the proof of Theorem 4.2.3. Since σ_t is linear, this gives

$$\sigma_t(c_1\xi_1 + \cdots + c_k\xi_k) = 0, \qquad t \in \mathbb{T}.$$

Injectivity of σ_t then gives

$$c_1\xi_1 + \cdots + c_k\xi_k = 0,$$

showing linear dependence of $\{\xi_1, \dots, \xi_k\}$.

(ii) $\implies$ (iii) From Proposition 4.2.7, noting that $\xi_1, \dots, \xi_k$ are of class $\mathsf{AC}^{k-1}_{\mathrm{loc}}$, and so of class C^{k-1}, the assumption of (ii) implies that $\{\xi_1, \dots, \xi_k\}$ is linearly independent. Suppose now that there exists $t' \in \mathbb{T}$ such that $W(\xi_1, \dots, \xi_k)(t') = 0$. Then there exists $c_1, \dots, c_k \in \mathbb{R}$, not all zero, such that

$$\begin{bmatrix} \xi_1(t') & \xi_2(t') & \cdots & \xi_k(t') \\ \frac{\mathrm{d}\xi_1}{\mathrm{d}t}(t') & \frac{\mathrm{d}\xi_2}{\mathrm{d}t}(t') & \cdots & \frac{\mathrm{d}\xi_k}{\mathrm{d}t}(t') \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{d}^{k-1}\xi_1}{\mathrm{d}t^{k-1}}(t') & \frac{\mathrm{d}^{k-1}\xi_2}{\mathrm{d}t^{k-1}}(t') & \cdots & \frac{\mathrm{d}^{k-1}\xi_k}{\mathrm{d}t^{k-1}}(t') \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{4.2}$$

Now, define $\xi \colon \mathbb{T} \to \mathbb{R}$ by

$$\xi = c_1\xi_1 + \cdots + c_k\xi_k.$$

By Theorem 4.2.3, $\xi \in \mathrm{Sol}(F)$. Moreover, the equation (4.2) gives

$$\xi(t') = 0, \ \frac{\mathrm{d}\xi}{\mathrm{d}t}(t') = 0, \dots, \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^k}(t') = 0.$$

By Proposition 4.2.1, we conclude that $\xi(t) = 0$ for all $t \in \mathbb{T}$. This contradicts the linear independence of $\{\xi_1, \dots, \xi_k\}$.

(iii) $\implies$ (i) This follows from Proposition 4.2.7, noting that $\xi_1, \dots, \xi_k$ are of class $\mathsf{AC}^{k-1}_{\mathrm{loc}}$, and so of class C^{k-1}. ■

The following result gives an interesting characterisation of the Wronskian, further illustrating the fact that, when applied to solutions of scalar linear homogeneous ordinary differential equations, it serves to characterise linear independence of sets of solutions.

4.2.10 Proposition (Abel's formula) *Consider the scalar linear homogeneous ordinary differential equation* F *with right-hand side (4.1) and suppose that* $a_0, a_1, \dots, a_{k-1} \in L^1_{loc}(\mathbb{T}; \mathbb{R})$. *If* $\{\xi_1, \dots, \xi_k\}$ *are linearly independent, then, for any* $t_0, t \in \mathbb{T}$,

$$W(\xi_1, \dots, \xi_k)(t) = W(\xi_1, \dots, \xi_k)(t_0) e^{-\int_{t_0}^{t} a_{k-1}(\tau)\, d\tau}.$$

Proof This is Exercise 5.2.4, which can be proved using some attributes of systems of linear ordinary differential equations in Section 5.2. ∎

One of the sort of peculiar features of the Wronskian is that it can be used to actually write down a differential equation, at least when the coefficient functions are continuous, which guarantees that the solutions are of class C^k.

4.2.11 Proposition (A Wronskian representation of a differential equation) *Consider the scalar linear homogeneous ordinary differential equation* F *with right-hand side (4.1) and suppose that the functions* $a_0, a_1, \dots, a_{k-1} : \mathbb{T} \to \mathbb{R}$ *are continuous. Let* $\{\xi_1, \dots, \xi_k\}$ *be a fundamental set of solutions for* F. *Then, for* $\xi \in C^k(\mathbb{T}; \mathbb{R})$ *and* $t \in \mathbb{T}$,

$$\frac{d^k \xi}{dt^k}(t) + a_{k-1}(t) \frac{d^{k-1}\xi}{dt^{k-1}}(t) + \cdots + a_1(t) \frac{d\xi}{dt}(t) + a_0 \xi(t) = \frac{W(\xi_1, \dots, \xi_k, \xi)(t)}{W(\xi_1, \dots, \xi_k)(t)}.$$

In particular,

$$\mathrm{Sol}(F) = \left\{ \xi \in C^k(\mathbb{T}; \mathbb{R}) \;\middle|\; \frac{W(\xi_1, \dots, \xi_k, \xi)(t)}{W(\xi_1, \dots, \xi_k)(t)} = 0,\ t \in \mathbb{T} \right\}.$$

Proof First of all, note by Proposition 4.2.9 that $W(\xi_1, \dots, \xi_k)(t)$ is never zero, so this is valid to appear in denominators, as in the statement of the proposition.

We shall prove the last assertion first. First suppose that $\xi \in \mathrm{Sol}(F)$, then

$$\xi = c_1 \xi_1 + \cdots + c_k \xi_k$$

for some (unique) constants $c_1, \dots, c_k \in \mathbb{R}$. Therefore, the functions $\{\xi, \xi_1, \dots, \xi_k\}$ are linearly dependent, cf.

$$-c_1 \xi_1 - \cdots - c_k \xi_k + 1\xi = 0.$$

Therefore, differentiating this equation k-times gives

$$\begin{bmatrix} \xi_1(t) & \xi_2(t) & \cdots & \xi_k(t) & \xi(t) \\ \frac{d\xi_1}{dt}(t) & \frac{d\xi_2}{dt}(t) & \cdots & \frac{d\xi_k}{dt}(t) & \frac{d\xi}{dt}(t) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{d^{k-1}\xi_1}{dt^{k-1}}(t) & \frac{d^{k-1}\xi_2}{dt^{k-1}}(t) & \cdots & \frac{d^{k-1}\xi_k}{dt^{k-1}}(t) & \frac{d^{k-1}\xi}{dt^{k-1}}(t) \\ \frac{d^k\xi_1}{dt^k}(t) & \frac{d^k\xi_2}{dt^k}(t) & \cdots & \frac{d^k\xi_k}{dt^k}(t) & \frac{d^k\xi}{dt^k}(t) \end{bmatrix} \begin{bmatrix} -c_1 \\ -c_2 \\ \vdots \\ -c_k \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

for all $t \in \mathbb{T}$. From this we immediately conclude that $W(\xi_1, \dots, \xi_k, \xi)(t) = 0$ for all $t \in \mathbb{T}$, and so

$$\xi \in \left\{ \tilde{\xi} \in C^k(\mathbb{T}; \mathbb{R}) \;\middle|\; \frac{W(\xi_1, \dots, \xi_k, \tilde{\xi})(t)}{W(\xi_1, \dots, \xi_k)(t)} = 0 \right\}.$$

Now note that, if we expand the determinant $\mathrm{W}(\xi_1, \dots, \xi_k, \xi)$ about the last column, we get an expression of the form

$$\begin{aligned}&\mathrm{W}(\xi_1, \dots, \xi_k, \xi)(t)\\ &\qquad = \mathrm{W}(\xi_1, \dots, \xi_k)(t)\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t) + b_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + b_1(t)\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t)\xi(t)\end{aligned}$$

for some continuous functions $a_0, a_1, \dots, a_{k-1} \colon \mathbb{T} \to \mathbb{R}$. By Proposition 4.2.9 it follows that

$$\left\{\xi \in \mathrm{C}^k(\mathbb{T}; \mathbb{R}) \;\middle|\; \frac{\mathrm{W}(\xi_1, \dots, \xi_k, \xi)(t)}{\mathrm{W}(\xi_1, \dots, \xi_k)(t)} = 0,\ t \in \mathbb{T}\right\}$$

is the set of solutions to a kth-order scalar linear homogeneous ordinary differential equation. Moreover, since we clearly have $\mathrm{W}(\xi_1, \dots, \xi_k, \xi_j) = 0$ for every $j \in \{1, \dots, k\}$, (it is the determinant of a $(k+1) \times (k+1)$ matrix with two equal columns), it follows that $\{\xi_1, \dots, \xi_k\}$ is a fundamental set of solutions for this differential equation. Thus we have shown that

$$\mathrm{Sol}(F) = \left\{\xi \in \mathrm{C}^k(\mathbb{T}; \mathbb{R}) \;\middle|\; \frac{\mathrm{W}(\xi_1, \dots, \xi_k, \xi)(t)}{\mathrm{W}(\xi_1, \dots, \xi_k)(t)} = 0\right\}.$$

To prove the first assertion, we shall show that the set of solutions for a kth-order scalar linear homogeneous ordinary differential equation uniquely determines its coefficients. That is, we show that if two such equations F and G with right-hand sides

$$\begin{aligned}\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) &= -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x,\\ \widehat{G}(t, x, x^{(1)}, \dots, x^{(k-1)}) &= -b_{k-1}(t)x^{(k-1)} - \cdots - b_1(t)x^{(1)} - b_0(t)x\end{aligned}$$

satisfy $\mathrm{Sol}(F) = \mathrm{Sol}(G)$, then $a_j = b_j$, $j \in \{0, 1, \dots, k-1\}$. Let us consider the differential equation

$$H(t, x, x^{(1)}, \dots, x^{(k-1)}) = F(t, x, x^{(1)}, \dots, x^{(k-1)}) - G(t, x, x^{(1)}, \dots, x^{(k-1)}).$$

Note that this is not necessarily a $(k-1)$st-order ordinary differential equation, since we may have $a_{k-1} = b_{k-1}$. However, suppose that $\widehat{F} \neq \widehat{G}$ and let j be the largest element of $\{0, 1, \dots, k-1\}$ such that $a_j \neq b_j$. Thus there exists $t_0 \in \mathbb{T}$ so that $a_j(t_0) \neq b_j(t_0)$. Since a_j and b_j are continuous, there is an interval $\mathbb{T}' \subseteq \mathbb{T}$ around t_0 such that $a_j(t) \neq b_j(t)$ for all $t \in \mathbb{T}'$. We then define an ordinary differential equation H' with right-hand side

$$\begin{aligned}&\widehat{H}' \colon \mathbb{T}' \times \mathbb{R} \oplus \mathrm{L}^{lej-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}\\ &\qquad (t, x, x^{(1)}, \dots, x^{(j-1)}) \mapsto -\frac{a_{j-1}(t) - b_{j-1}(t)}{a_j(t) - b_j(t)}x^{(j-1)} - \cdots - \frac{a_1(t) - b_1(t)}{a_j(t) - b_j(t)}x^{(1)}\\ &\qquad\qquad - \frac{a_0(t) - b_0(t)}{a_j(t) - b_j(t)}x.\end{aligned}$$

This jth-order ordinary differential equation has $\xi_1, \dots, \xi_k$ as linearly independent solutions, and this is in contradiction with Theorem 4.2.3. Thus we must have $\widehat{F} = \widehat{G}$, as claimed. ■

4.2.2 Equations with constant coefficients

Having said about as much as one can say, in general, about the situation with time-varying coefficients, we now turn to the case of constant coefficient scalar linear homogeneous ordinary differential equations. If

$$F\colon \mathbb{T}\times\mathbb{R}\oplus \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R};\mathbb{R})\to\mathbb{R}$$

is such an equation, then its right-hand side must be given by

$$\widehat{F}(t,x,x^{(1)},\dots,x^{(k)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1x^{(1)} - a_0x \tag{4.3}$$

for $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$. Thus a solution $t \mapsto \xi(t)$ satisfies the equation

$$\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0\xi(t) = 0. \tag{4.4}$$

These equations are, of course, a special case of the equations considered in Section 4.2.1, and so all statements made about the general case of time-varying coefficients hold in the special case of constant coefficients. In particular, Propositions 4.2.1 and 4.2.2, and Theorem 4.2.3 hold for equations of the form (4.4). However, for these constant coefficient equations, it is possible to explicitly describe the character of the solutions, and this is what we undertake to do.

The trick, motivated to some extent by Example 4.1.3–1, is to *assume* a solution of the form $\xi(t) = ae^{rt}$ for $a, r \in \mathbb{R}$, and see what happens. A direct substitution into the equation (4.4) shows that, with ξ in this assumed form,

$$\frac{\mathrm{d}^k(ae^{rt})}{\mathrm{d}t^k} + a_{k-1}\frac{\mathrm{d}^{k-1}(ae^{rt})}{\mathrm{d}t^{k-1}} + \cdots + a_1\frac{\mathrm{d}(ae^{rt})}{\mathrm{d}t} + a_0(ae^{rt}) = ae^{rt}(r^k + a_{k-1}r^{k-1} + \cdots + a_1r + a_0) = 0.$$

Since we are looking for nontrivial solutions, we suppose that $a \neq 0$, in which case $\xi(t) = ae^{rt}$ is a solution for F if and only if

$$r^k + a_{k-1}r^{k-1} + \cdots + a_1r + a_0 = 0.$$

With this as backdrop, we make the following definition.

4.2.12 Definition (Characteristic polynomial of a scalar linear homogeneous differential equation with constant coefficients) Consider the linear homogeneous ordinary differential equation F with constant coefficients and with right-hand side (4.3). The ***characteristic polynomial*** of F is

$$P_F = X^k + a_{k-1}X^{k-1} + \cdots + a_1X + a_0 \in \mathbb{R}[X]. \qquad \bullet$$

Now we systematically develop the methodology for solving scalar linear homogeneous ordinary differential equations with constant coefficients.

4.2.2.1 Complexification of scalar linear ordinary differential equations It turns out that to solve constant coefficient linear ordinary differential equations, one needs to work with complex numbers. To do this systematically, we introduce the notion of "complexification," by which a real equation is converted into a complex one. This is rather elementary in this setting, but will be less elementary in Section 5.2.2. Thus it will do not harm, and maybe do some good, to treat this systematically here.

First let us understand the notation for derivatives of $\mathbb{C}$-valued functions of a single real variable, i.e., functions of time. Let $\mathbb{T} \subseteq \mathbb{R}$ be an interval and suppose that we have a mapping $\zeta\colon \mathbb{T} \to \mathbb{C}$. Since we have $\mathbb{C} \simeq \mathbb{R}^2$, it makes sense to say that ζ is of class C^k for any $k \in \mathbb{Z}_{\geq 0}$: it is of class C^k if and only if both its real and imaginary parts are of class C^k. Moreover, if we write ζ as a sum of its real and imaginary parts, $\zeta(t) = \xi(t) + \mathrm{i}\eta(t)$, then we have

$$\frac{\mathrm{d}^j\zeta}{\mathrm{d}t^j} = \frac{\mathrm{d}^j\xi}{\mathrm{d}t^j} + \mathrm{i}\frac{\mathrm{d}^j\eta}{\mathrm{d}t^j}.$$

Thus derivatives of order j are just $\mathbb{C}$-valued functions of t. Thus we can follow the same line of reasoning as Remark 3.1.5 and make the identification $\mathrm{L}^j_{\mathrm{sym}}(\mathbb{R};\mathbb{C}) \simeq \mathbb{C}$.

Here is the basic and quite elementary construction.

4.2.13 Definition (Complexification of scalar linear ordinary differential equation) Consider the linear homogeneous ordinary differential equation F with constant coefficients and with right-hand side (4.3). The ***complexification*** of F is the mapping

$$\begin{aligned} F^{\mathbb{C}}\colon \mathbb{T} \times \mathbb{C} \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R};\mathbb{C}) &\to \mathbb{C}\\ (t, z, z^{(1)}, \dots, z^{(k)}) &\mapsto z^{(k)} + a_{k-1}z^{(k-1)} + \cdots + a_1 z^{(1)} + a_0 z. \end{aligned}$$

A ***solution*** for $F^{\mathbb{C}}$ is $\zeta \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T};\mathbb{C})$ that satisfies

$$\frac{\mathrm{d}^k\zeta}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\zeta}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\zeta}{\mathrm{d}t}(t) + a_0\zeta(t) = 0.$$

By $\mathrm{Sol}(F^{\mathbb{C}})$ we denote the set of solutions for $F^{\mathbb{C}}$. •

Everything we said in Section 4.2.1 about scalar linear homogeneous ordinary differential equations holds in the case of the complex differential equation $F^{\mathbb{C}}$, even when the coefficients are not constant. In particular, Propositions 4.2.1 and 4.2.2, and Theorem 4.2.3 hold in this case to give us the basic attributes of the complex differential equation, merely by replacing the appropriate occurrences of the symbol "$\mathbb{R}$" with the symbol "$\mathbb{C}$." In particular, $\mathrm{Sol}(F^{\mathbb{C}})$ is a k-dimensional $\mathbb{C}$-vector space if F has order k.

An essential result for returning to "reality" after complexification is the following simple result.

4.2.14 Lemma (Real and imaginary parts of complex solutions are solutions) *Consider the linear homogeneous ordinary differential equation* F *with constant coefficients, with right-hand side (4.3) and with complexification* $\mathrm{F}^{\mathbb{C}}$. *If* $\zeta\colon \mathbb{T} \to \mathbb{C}$ *is a solution for* $\mathrm{F}^{\mathbb{C}}$, *then* $\mathrm{Re}(\zeta)$ *and* $\mathrm{Im}(\zeta)$ *are solutions for* F.

Proof Since ζ is a solution for $F^{\mathbb{C}}$, we have

$$\frac{\mathrm{d}^k\zeta}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\zeta}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\zeta}{\mathrm{d}t}(t) + a_0\zeta(t) = 0.$$

Now we note that $\mathrm{Re}\colon \mathbb{C} \to \mathbb{R}$ and $\mathrm{Im}\colon \mathbb{C} \to \mathbb{R}$ are $\mathbb{R}$-linear maps. Since the coefficients $a_0, a_1, \ldots, a_{k-1}$ are real, this gives

$$\begin{aligned} 0 &= \mathrm{Re}\left(\frac{\mathrm{d}^k\zeta}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\zeta}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\zeta}{\mathrm{d}t}(t) + a_0\zeta(t)\right) \\ &= \frac{\mathrm{d}^k \mathrm{Re}(\zeta)}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\mathrm{Re}(\zeta)}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\,\mathrm{Re}(\zeta)}{\mathrm{d}t}(t) + a_0\,\mathrm{Re}(\zeta)(t), \end{aligned}$$

showing that $\mathrm{Re}(\zeta)$ is a solution for F. In like manner, of course, $\mathrm{Im}(\zeta)$ is also a solution for F. ■

4.2.2.2 Differential operator calculus

We introduce a simple object that will be used to say a few simple things about our constant coefficient ordinary differential equations.

4.2.15 Definition (Scalar differential operator with constant coefficients) Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $k \in \mathbb{Z}_{\geq 0}$. A ***k*th-order scalar differential operator with constant coefficients in** $\mathbb{F}$ is a mapping

$$D\colon \mathsf{C}^\infty(\mathbb{T}; \mathbb{F}) \to \mathsf{C}^\infty(\mathbb{T}; \mathbb{F})$$

of the form

$$D(f)(t) = d_k\frac{\mathrm{d}^k f}{\mathrm{d}t}(t) + d_{k-1}\frac{\mathrm{d}^{k-1}f}{\mathrm{d}t^{k-1}}(t) + \cdots + d_1\frac{\mathrm{d}f}{\mathrm{d}t}(t) + d_0 f(t),$$

for $d_0, d_1, \ldots, d_k \in \mathbb{F}$ with $d_k \neq 0$. The ***symbol*** for such an object is

$$\sigma(D) = d_k X^k + d_{k-1}X^{k-1} + \cdots + d_1 X + d_0 \in \mathbb{F}[X]. \qquad \bullet$$

Note that, while the domain and range of D in the preceding definition is the set of infinitely differentiable functions, clearly the definition makes sense when applied to functions that are at least k-times continuously differentiable. Indeed, we can think of D as a mapping from $\mathsf{C}^{k+m}(\mathbb{T}; \mathbb{F})$ to $\mathsf{C}^m(\mathbb{T}; \mathbb{F})$ for any $m \in \mathbb{Z}_{\geq 0}$. The definition as stated just allows us to not fuss about this sort of thing for the purposes of our discussion.

Note that differential operators of the sort we are talking about have a product given by composition. Thus, if D_1 and D_1 are k_1th- and k_2th-order scalar differential operators with constant coefficients, then we define a $(k_1 + k_2)$th-order scalar differential operator D_1D_2 with constant coefficients by $D_1D_2(f) = D_1(D_2(f))$.

A simplifying observation about scalar differential operators with constant coefficients is the following.

4.2.16 Proposition (The symbol of a product is the product of the symbols) *Let* $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, *let* $\mathbb{T} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$. *If* D_1 *and* D_1 *are* k_1*th- and* k_2*th-order scalar differential operators with constant coefficients, then* $\sigma(\mathrm{D}_1\mathrm{D}_2) = \sigma(\mathrm{D}_1)\sigma(\mathrm{D}_2)$.

Proof Let us write

$$\sigma(D_1) = \sum_{j=0}^{k_1} d_{1,j} X^j, \quad \sigma(D_2) = \sum_{j=0}^{k_2} d_{2,j} X^j.$$

Then, for $f \in \mathsf{C}^\infty(\mathbb{T}; \mathbb{F})$,

$$D_1D_2(f) = \sum_{j=0}^{k_1} d_{1,j} \frac{\mathrm{d}^j}{\mathrm{d}t^j} \left(\sum_{l=0}^{k_2} d_{2,l} \frac{\mathrm{d}^l f}{\mathrm{d}t^l} \right) = \sum_{k=0}^{k_1+k_2} \sum_{j=0}^{k} d_{1,j} d_{2,k-j} \frac{\mathrm{d}^k f}{\mathrm{d}t^k}.$$

Since

$$\sigma(D_1)\sigma(D_2) = \sum_{k=0}^{k_1+k_2} \sum_{j=0}^{k} d_{1,j} d_{2,k-j} X^k,$$

the result follows. ■

4.2.17 Corollary (The product for differential operators is commutative) *Let* $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, *let* $\mathbb{T} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$. *If* D_1 *and* D_1 *are* k_1*th- and* k_2*th-order scalar differential operators with constant coefficients, then* $\mathrm{D}_1\mathrm{D}_2 = \mathrm{D}_2\mathrm{D}_1$.

Proof This follows from the following facts: (1) polynomial multiplication is commutative; (2) the mapping that assigns $\sigma(D)$ to D is injective. ■

4.2.2.3 Bases of solutions

Now we construct a family of solutions for a scalar linear homogeneous ordinary differential equation. We do this via a procedure.

4.2.18 Procedure (Basis of solutions for scalar linear homogeneous ordinary differential equations with constant coefficients) Given a scalar linear homogeneous ordinary differential equation

$$F\colon \mathbb{T} \times \mathbb{R} \oplus \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$

with right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x,$$

do the following.

1. Let $F^{\mathbb{C}}$ be the complexification of F,
2. Consider the kth-order scalar differential operator D_F with constant coefficients in $\mathbb{C}$ defined by
$$\sigma(D_{F^{\mathbb{C}}}) = X^k + a_{k-1}X^k + \cdots + a_1X + a_0.$$
3. Let $r_1, \ldots, r_s$ be the distinct roots of $\sigma(D_F)$ and let $m(r_j)$, $j \in \{1, \ldots, s\}$, be the multiplicity of the root r_j. Thus
$$\sigma(D_{F^{\mathbb{C}}}) = (X - r_1)^{m(r_1)} \cdots (X - r_s)^{m(r_s)}.$$
4. Fix $j \in \{1, \ldots, s\}$ and consider the following cases.
   (a) $r_j \in \mathbb{R}$: Define functions $\xi_{r_j,l} \colon \mathbb{T} \to \mathbb{R}$, $l \in \{1, \ldots, m(r_j)\}$, by
   $$\xi_{r_j,l}(t) = t^{l-1}\mathrm{e}^{r_j t}, \qquad l \in \{1, \ldots, m(r_j)\}.$$
   (b) $r_j \in \mathbb{C} \setminus \mathbb{R}$: Note that, since r_j is complex and not real, $\bar{r}_j$ is also a root of $\sigma(D_{F^{\mathbb{C}}})$. We will work only with one of these roots, so we write $r_j = \sigma_j + \mathrm{i}\omega_j$ with $\omega_j > 0$. Define functions $\mu_{r_j,l}, \nu_{r_j,l} \colon \mathbb{T} \to \mathbb{R}$ by
   $$\mu_{r_j,l}(t) = t^{l-1}\mathrm{e}^{\sigma_j t}\cos(\omega_j t), \quad \nu_{r_j,l}(t) = t^{l-1}\mathrm{e}^{\sigma_j t}\sin(\omega_j t), \qquad l \in \{1, \ldots, m(r_j)\}.$$
5. Note that the result of the above steps is k functions. We will show that these functions form a basis for $\mathrm{Sol}(F)$. •

4.2.19 Theorem (Basis of solutions for scalar linear homogeneous ordinary differential equations with constant coefficients) *Given a scalar linear homogeneous ordinary differential equation with constant coefficients*
$$\mathrm{F} \colon \mathbb{T} \times \mathbb{R} \oplus \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$
with right-hand side
$$\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}, \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(k-1)}) = -\mathrm{a}_{k-1}\mathrm{x}^{(k-1)} - \cdots - \mathrm{a}_1\mathrm{x}^{(1)} - \mathrm{a}_0\mathrm{x},$$
define k *functions as in Procedure 4.2.18. Then these functions form a basis for* Sol(F).

Proof First we show that each of the functions defined in Procedure 4.2.18 is a solution for F.

First we consider the functions $\xi_{r_j,l}(t) = t^l\mathrm{e}^{r_j t}$, $l \in \{0, 1, \ldots, m(r_j) - 1\}$, associated with a real root r_j of the characteristic polynomial for F. Since
$$\sigma(D_{F^{\mathbb{C}}}) = (X - r_1)^{m(r_1)} \cdots (X - r_s)^{m(r_s)},$$
by Corollary 4.2.17 we can write
$$\sigma(D_{F^{\mathbb{C}}}) = P(X - r_j)^{m(r_j)}$$
for some $P \in \mathbb{C}[X]$. Therefore, it suffices to show that, for $r \in \mathbb{R}$ and for $m, l \in \mathbb{Z}_{\geq 0}$ with $m \in \mathbb{Z}_{>0}$ and $l < m$, we have
$$\left(\frac{\mathrm{d}}{\mathrm{d}t} - r\right)^m P(t)\mathrm{e}^{rt} = 0, \tag{4.5}$$
where P is any polynomial function of degree $l \in \{0, 1, \ldots, m - 1\}$. To prove (4.5), we first prove a simple lemma.

1 Lemma *Let* $\mathrm{m} \in \mathbb{Z}_{>0}$ *and* $\mathrm{r} \in \mathbb{C}$. *If* $\xi\colon \mathbb{T} \to \mathbb{C}$ *is of class* C^{m} *then*

$$\left(\frac{\mathrm{d}}{\mathrm{dt}} - \mathrm{r}\right)^{\mathrm{m}} (\xi(\mathrm{t})\mathrm{e}^{\mathrm{rt}}) = \mathrm{e}^{\mathrm{rt}} \frac{\mathrm{d}^{\mathrm{m}}\xi}{\mathrm{dt}^{\mathrm{m}}}(\mathrm{t}).$$

Proof We prove this by induction on m. For $m = 1$ we have

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} - r\right)(\xi(t)\mathrm{e}^{rt}) = \frac{\mathrm{d}\xi}{\mathrm{d}t}(t)\mathrm{e}^{rt} + r\xi(t)\mathrm{e}^{rt} - r\xi(t)\mathrm{e}^{rt} = \mathrm{e}^{rt}\frac{\mathrm{d}\xi}{\mathrm{d}t}(t),$$

giving the lemma when $m = 1$. Now suppose that the lemma holds when $m = k$. Then

$$\begin{aligned}\left(\frac{\mathrm{d}}{\mathrm{d}t} - r\right)^{k+1} (\xi(t)\mathrm{e}^{rt}) &= \left(\frac{\mathrm{d}}{\mathrm{d}t} - r\right)\left(\frac{\mathrm{d}}{\mathrm{d}t} - r\right)^{k} (\xi(t)\mathrm{e}^{rt}) \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}t} - r\right)\mathrm{e}^{rt}\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t) \\ &= r\mathrm{e}^{rt}\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t) + \mathrm{e}^{rt}\frac{\mathrm{d}^{k+1}\xi}{\mathrm{d}t^{k+1}}(t) - r\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t) \\ &= \mathrm{e}^{rt}\frac{\mathrm{d}^{k+1}\xi}{\mathrm{d}t^{k+1}}(t),\end{aligned}$$

as desired. ▼

Now, if P is a polynomial function of degree $l \in \{0, 1, \dots, m\}$, by the Lemma 1 we have

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} - r\right)^{m} P(t)\mathrm{e}^{rt} = \mathrm{e}^{rt}\frac{\mathrm{d}^m P}{\mathrm{d}t^m}(t) = 0.$$

Thus shows that the functions $\xi_{r_j,l}(t) = t^l\mathrm{e}^{r_j t}$, $l \in \{0, 1, \dots, m(r_j) - 1\}$, are solutions for F.

Next we consider the functions

$$\mu_{r_j,l} = t^l\mathrm{e}^{\sigma_j t}\cos(\omega_j t), \quad \nu_{r_j,l} = t^l\mathrm{e}^{\sigma_j t}\sin(\omega_j t), \qquad l \in \{0, 1, \dots, m(r_j) - 1\},$$

corresponding to a complex root $r_j = \sigma_j + \mathrm{i}\omega_j$, $\omega_j > 0$, of the characteristic polynomial of F. In this case, we argue, exactly as in the case of a real root above, that the $\mathbb{C}$-valued functions $\zeta_{r_j,l}(t) = t^l\mathrm{e}^{r_j t}$, $l \in \{0, 1, \dots, m(r_j) - 1\}$, are solutions for $F^{\mathbb{C}}$. Then, by Lemma 4.2.14, we have that

$$\begin{aligned}\mu_{r_j,l}(t) &= t^l\mathrm{e}^{\sigma_j t}\cos(\omega_j t) \\ &= \mathrm{Re}(t^l\mathrm{e}^{\sigma_j t}(\cos(\omega_j t) + \mathrm{i}\sin(\omega_j t)) \\ &= \mathrm{Re}(t^l\mathrm{e}^{\sigma_j t}\mathrm{e}^{\mathrm{i}\omega_j t}) = \mathrm{Re}(\zeta_{r_j,l}(t))\end{aligned}$$

and, similarly,

$$\nu_{r_j,l} = t^l\mathrm{e}^{\sigma_j t}\sin(\omega_j t) = \mathrm{Im}(\zeta_{r_j,l}(t))$$

are solutions for F for $l \in \{0, 1, \dots, m(r_j) - 1\}$.

Our above arguments show that the functions produced in Procedure 4.2.18 are solutions. Moreover, since Procedure 4.2.18 produces k solutions for F, by Theorem 4.2.3 it suffices to show that these solutions are linearly independent to show that they form a basis for $\mathrm{Sol}(F)$. We achieve this with the aid of the following lemma.

2 Lemma *Let $\mathbb{T} \subseteq \mathbb{R}$ be an interval containing more than one point. Let $r_1, \dots, r_s \in \mathbb{R}$ be distinct and let $P_1, \dots, P_s$ be $\mathbb{C}$-valued polynomial functions on $\mathbb{T}$. If*

$$P_1(t)e^{r_1 t} + \cdots + P_s(t)e^{r_s t} = 0, \qquad t \in \mathbb{T},$$

then $P_j(t) = 0$ for all $j \in \{1, \dots, s\}$ and $t \in \mathbb{T}$.

Proof We prove the lemma by induction on s. For $s = 1$ we have, for $r_1 \in \mathbb{R}$ and a polynomial function P_1,

$$\begin{aligned} & P_1(s)e^{r_1 t} = 0, \qquad t \in \mathbb{T}, \\ \implies \quad & P_1(t) = 0, \qquad t \in \mathbb{T}, \end{aligned}$$

giving the result in this case. Now suppose that the lemma is true for $s = k$ and suppose that

$$P_1(t)e^{r_1 t} + \cdots + P_k(t)e^{r_k t} + P_{k+1}(t)e^{r_{k+1} t} = 0, \qquad t \in \mathbb{T},$$

for distinct $r_1, \dots, r_k, r_{k+1} \in \mathbb{R}$ and for polynomial functions $P_1, \dots, P_k, P_{k+1}$. Then

$$P_1(t)e^{(r_1 - r_{k+1})t} + \cdots + P_k(t)e^{(r_k - r_{k+1})t} + P_{k+1}(t) = 0, \qquad t \in \mathbb{T}. \tag{4.6}$$

Now let us differentiate this expression m times with respect to t, using the Leibniz Rule for higher-order derivatives stated in Proposition I-3.2.11. After m differentiations we get

$$P_1^m(t)e^{(r_1 - r_{k+1})t} + \cdots + P_k^m(t)e^{(r_k - r_{k+1})t} + \frac{d^m P_{k+1}}{dt^m}(t) = 0, \qquad t \in \mathbb{T},$$

where

$$P_j^m(t) = \sum_{l=0}^{m} (r_j - r_{k+1})^l \binom{m}{l} \frac{d^{m-l} P_j}{dt^{m-l}}(t). \tag{4.7}$$

Since $r_j - r_{k+1} \neq 0$, P_j^m is a polynomial function whose degree is the same as the degree of P_j. Now, for m sufficiently large (larger than the degree of P_{k+1}, to be precise), $\frac{d^m P_{k+1}}{dt^m} = 0$. With m so chosen, we have

$$P_1^m(t)e^{(r_1 - r_{k+1})t} + \cdots + P_k^m(t)e^{(r_k - r_{k+1})t} = 0, \qquad t \in \mathbb{T}.$$

By the induction hypothesis, $P_j^m(t) = 0$ for $j \in \{1, \dots, k\}$ and $t \in \mathbb{T}$. Now, in the expression (4.7) for P_j^m, note that the highest polynomial degree term in t in the sum occurs when $l = m$, and this term is $(r_j - r_{k+1})^m P_j(t)$. For the polynomial P_j^m to vanish, this term in the sum must vanish, i.e., $P_j(t) = 0$ for every $j \in \{1, \dots, k\}$ and $t \in \mathbb{T}$. Finally, (4.6) then gives $P_{k+1}(t) = 0$ for all $t \in \mathbb{T}$, giving the result. ▼

Now we can show that the solutions produced by Procedure 4.2.18 are linearly independent. Suppose that there are s_1 distinct real roots, $r_1, \dots, r_{s_1}$, and s_2 distinct complex roots,

$$\rho_j = \sigma_1 + i\omega_j, \dots, \rho_{s_2} = \sigma_{s_2} + i\omega_{s_2},$$

with $\omega_1, \dots, \omega_{s_2} > 0$, for the characteristic polynomial of F. Thus $s_1 + 2s_2 = k$. Suppose that we have k scalars

$$c_{j,l}, \qquad j \in \{1, \dots, s_1\},\ l \in \{0, 1, \dots, m(r_j) - 1\}, \tag{4.8}$$

and

$$a_{j,l},\ b_{j,l}, \qquad j \in \{1, \dots, s_2\},\ l \in \{0, 1, \dots, m(\rho_j) - 1\}, \tag{4.9}$$

satisfying

$$\begin{aligned}
&(c_{1,0} + c_{1,1}t + \cdots + c_{1,m(r_1)-1}t^{m(r_1)-1})\mathrm{e}^{r_1 t} + \dots \\
&\qquad + (c_{s_1,0} + c_{s_1,1}t + \cdots + c_{s_1,m(r_{s_1})-1}t^{m(r_{s_1})-1})\mathrm{e}^{r_1 t} \\
&\qquad + (a_{1,0} + a_{1,1}t + \cdots + a_{1,m(\rho_1)-1}t^{m(\rho_1)-1})\mathrm{e}^{\sigma_1 t}\cos(\omega_1 t) \\
&\qquad + (b_{1,0} + b_{1,1}t + \cdots + b_{1,m(\rho_1)-1}t^{m(\rho_1)-1})\mathrm{e}^{\sigma_1 t}\sin(\omega_1 t) + \dots \\
&\qquad + (a_{s_2,0} + a_{s_2,1}t + \cdots + a_{s_2,m(\rho_{s_2})-1}t^{m(\rho_{s_2})-1})\mathrm{e}^{\sigma_{s_2} t}\cos(\omega_{s_2} t) \\
&\qquad + (b_{s_2,0} + b_{s_2,1}t + \cdots + b_{s_2,m(\rho_{s_2})-1}t^{m(\rho_{s_2})-1})\mathrm{e}^{\sigma_{s_2} t}\sin(\omega_{s_2} t) = 0, \qquad t \in \mathbb{T}.
\end{aligned}$$

By Lemma 2, the polynomial functions

$$\begin{aligned}
&c_{1,0} + c_{1,1}t + \cdots + c_{1,m(r_1)-1}t^{m(r_1)-1}, \dots, \\
&\qquad c_{s_1,0} + c_{s_1,1}t + \cdots + c_{s_1,m(r_{s_1})-1}t^{m(r_{s_1})-1}, \\
&\qquad a_{1,0} + a_{1,1}t + \cdots + a_{1,m(\rho_1)-1}t^{m(\rho_1)-1}, \\
&\qquad b_{1,0} + b_{1,1}t + \cdots + b_{1,m(\rho_1)-1}t^{m(\rho_1)-1}, \dots, \\
&\qquad a_{s_2,0} + a_{s_2,1}t + \cdots + a_{s_2,m(\rho_{s_2})-1}t^{m(\rho_{s_2})-1}, \\
&\qquad\qquad b_{s_2,0} + b_{s_2,1}t + \cdots + b_{s_2,m(\rho_{s_2})-1}t^{m(\rho_{s_2})-1}
\end{aligned}$$

must all vanish. But this implies that the scalars (4.8) and (4.9) must all vanish. This gives the desired linear independence. ∎

4.2.2.4 Some examples As concerns the general theory of scalar linear homogeneous ordinary differential equations, the matter is settled pretty much by Theorem 4.2.19. It remains to consider a few examples.

We first consider an "academic" example, one that illustrates Procedure 4.2.18, but which has no particular deep meaning.

4.2.20 Example ("Academic" example) We consider the 4th-order scalar linear homogeneous ordinary differential equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}, x^{(2)}, x^{(3)}) = -5x + 8x^{(1)} - 2x^{(2)}.$$

Thus solutions $t \mapsto \xi(t)$ to this equation satisfy

$$\frac{\mathrm{d}^4\xi}{\mathrm{d}t^4}(t) + 2\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}(t) - 8\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + 5\xi(t) = 0.$$

The characteristic polynomial is

$$P_F = X^4 + 2X^2 - 8X + 5$$

which can be verified to have roots and multiplicities

$$r_1 = 1,\ m(r_1) = 2,\ \ \rho_1 = -1 + 2\mathrm{i},\ m(\rho_1) = 1.$$

Of course, we also have the root $\overline{\rho}_1 = -1 - 2\mathrm{i}$, but the bookkeeping for this is dealt with when we produce two solutions corresponding to ρ_1. According to Procedure 4.2.18 the 4 solutions that form a basis for $\mathrm{Sol}(F)$ are then

$$\xi_{r_1,0}(t) = \mathrm{e}^t,\ \xi_{r_1,1}(t) = t\mathrm{e}^t,\ \mu_{\rho_1,0}(t) = \mathrm{e}^{-t}\cos(2t),\ \nu_{\rho_1,0}(t) = \mathrm{e}^{-t}\sin(2t).$$

Thus *any* solution for F can be written as

$$\xi(t) = c_1\mathrm{e}^t + c_2t\mathrm{e}^t + c_3\mathrm{e}^{-t}\cos(2t) + c_4\mathrm{e}^{-t}\sin(2t).$$

To prescribe a *specific* solution, according to Proposition 4.2.1, we specify initial conditions. For simplicity, let us do this at $t = 0$:

$$\xi(0) = x_0,\ \frac{\mathrm{d}\xi}{\mathrm{d}t}(0) = x + 0^{(1)},\ \frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}(0) = x_0^{(2)},\ \frac{\mathrm{d}^3\xi}{\mathrm{d}t}(0) = x_0^{(3)}. \tag{4.10}$$

To use these conditions to determine c_1, c_2, c_3, c_4 is a tedious problem in linear algebra. We compute

$$\begin{aligned}
\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) &= c_1\mathrm{e}^t + c_2(\mathrm{e}^t + t\mathrm{e}^t) + c_3(-\mathrm{e}^{-t}\cos(2t) - 2\mathrm{e}^{-t}\sin(2t))\\
&\quad + c_4(2\mathrm{e}^{-t}\cos(2t) - \mathrm{e}^{-t}\sin(2t)),\\
\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}(t) &= c_1\mathrm{e}^t + c_2(2\mathrm{e}^t + t\mathrm{e}^t) + c_3(-3\mathrm{e}^{-t}\cos(2t) + 4\mathrm{e}^{-t}\sin(2t))\\
&\quad + c_4(-4\mathrm{e}^{-t}\cos(2t) - 3\mathrm{e}^{-t}\sin(2t)),\\
\frac{\mathrm{d}^3\xi}{\mathrm{d}t^3}(t) &= c_1\mathrm{e}^t + c_2(3\mathrm{e}^t + t\mathrm{e}^t) + c_3(11\mathrm{e}^{-t}\cos(2t) + 2\mathrm{e}^{-t}\sin(2t))\\
&\quad + c_4(-2\mathrm{e}^{-t}\cos(2t) + 11\mathrm{e}^{-t}\sin(2t)).
\end{aligned}$$

Evaluating these at $t = 0$ gives the equations

$$\begin{aligned}
c_1 + c_3 &= x_0,\\
c_1 + c_2 - c_3 + 2c_4 &= x_0^{(1)},\\
c_1 + 2c_2 - 3c_3 - 4c_4 &= x_0^{(2)},\\
c_1 + 3c_2 + 11c_3 - 2c_4 &= x_0^{(3)}.
\end{aligned}$$

These are four linear equations in four unknowns that, because of Proposition 4.2.1, we know possesses unique solutions. These can be solved to give

$$\begin{aligned}
c_1 &= \tfrac{1}{16}(15x_0 + x_0^{(1)} + x_0^{(2)} - x_0^{(3)}),\\
c_2 &= \tfrac{1}{8}(-5x_0 + 3x_0^{(1)} + x_0^{(2)} + x_0^{(3)}),\\
c_3 &= \tfrac{1}{16}(x_0 - x_0^{(1)} - x_0^{(2)} + x_0^{(3)}),\\
c_4 &= \tfrac{1}{8}(-x_0 + 2x_0^{(1)} - x_0^{(2)}).
\end{aligned}$$

Go ahead and plug numbers into this bad boy, if this is your thing. •

The next two examples are intended to illustrate the how the behaviour of the roots of the characteristic polynomial affect the behaviour of solutions.

4.2.21 Example (First-order system behaviour) Here we consider a general 1st-order scalar linear homogeneous ordinary differential equation F with right-hand side

$$\widehat{F}(t,x) = -\frac{x}{\tau}$$

for $\tau \in \mathbb{R}$. Solutions $t \mapsto \xi(t)$ satisfy

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} + \tau^{-1}\xi(t) = 0.$$

This is an easy equation to solve. It characteristic polynomial is $P_F = X + \tau^{-1}$ which has the single real root $r_1 = -\tau^{-1}$. Thus, by Procedure 4.2.18, *any* solution has the form $\xi(t) = c\mathrm{e}^{-t/\tau}$. To determine c, we use initial conditions as in Proposition 4.2.1. We take a general initial time t_0 and prescribe $\xi(t_0) = x_0$. Thus

$$\xi(t_0) = c\mathrm{e}^{-t_0/\tau} \quad \Longrightarrow \quad c = x_0\mathrm{e}^{t_0/\tau},$$

and so $\xi(t) = x_0\mathrm{e}^{-(t-t_0)/\tau}$.

Let us think about this solution for a moment. When $\tau > 0$, this is ***exponential decay*** and when $\tau < 0$ it is ***exponential growth***. In Figure 4.1 we graph $\xi(t)$ as a

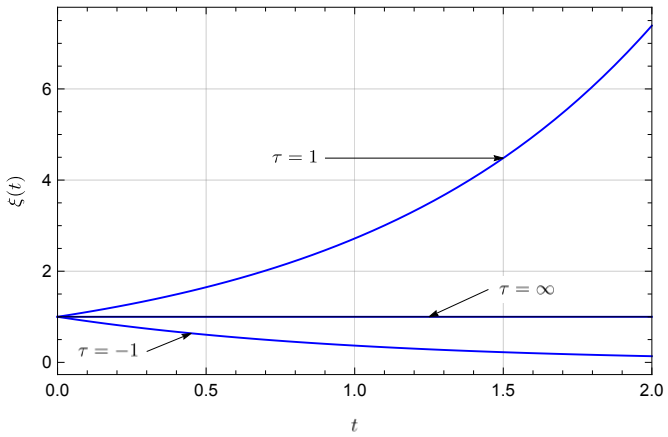


Figure 4.1 Solutions of a first-order scalar linear homogeneous ordinary differential equation with $\xi(0) = 1$

function of t for a few different τ's. Note that τ is not the rate of growth or decay, but the inverse of this. This is sometimes known as the ***time constant*** for the equation, since the units for τ are time. We can see that small (in magnitude) τ's give rise to relatively faster growth or decay. When $\tau = \infty$ (whatever that means), the decay or growth is infinitely slow, i.e., solutions neither grow nor decay. •

4.2.22 Example (Second-order system behaviour) We next consider a certain form of 2nd-order scalar linear homogeneous ordinary differential equation, namely such an equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}) = -\omega_0^2 x - 2\zeta\omega_0 x^{(1)}$$

for $\omega_0 \in \mathbb{R}_{>0}$ and $\zeta \in \mathbb{R}$. The equations (1.1) for a mass-spring-damper and (1.6) for the current in a series RLC circuit are of this general form. A solution $t \mapsto \xi(t)$ satisfies

$$\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}(t) + 2\zeta\omega_0\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + \omega_0^2\xi(t) = 0.$$

The characteristic polynomial is

$$P_F = X^2 + 2\zeta\omega_0 X + \omega_0^2.$$

The roots of this equation are found using the quadratic formula, and their character depends on discriminant which is $\Delta = 2\omega_0^2(\zeta^2 - 1)$. When $\Delta > 0$ the roots are real and when $\Delta < 0$ the roots are complex. To be precise, the roots are the following:

1. $\zeta^2 > 1$: two distinct real roots

$$r_1 = \omega_0(-\zeta + \sqrt{\zeta^2 - 1}),\ m(r_1) = 1,\ \ r_2 = \omega_0(-\zeta - \sqrt{\zeta^2 - 1}),\ m(r_2) = 1;$$

2. $\zeta = 1$: one repeated real root

$$r_1 = -\omega_0\zeta,\ m(r_2) = 2;$$

3. $\zeta^2 < 1$: a complex conjugate pair of roots with

$$\rho_1 = \omega_0(-\zeta + \mathrm{i}\sqrt{1 - \zeta^2}),\ m(\rho_1) = 1.$$

This then gives rise, according to Procedure 4.2.18, to the following solutions of the differential equation:

1. $\zeta^2 > 1$: $\xi(t) = c_1\mathrm{e}^{\omega_0(-\zeta + \sqrt{\zeta^2-1})t} + c_2\mathrm{e}^{\omega_0(-\zeta - \sqrt{\zeta^2-1})t}$;
2. $\zeta^2 = 1$: $\xi(t) = c_1\mathrm{e}^{-\omega_0\zeta t} + c_2 t\mathrm{e}^{-\omega_0\zeta t}$;
3. $\zeta^2 < 1$: $\xi(t) = c_1\mathrm{e}^{-\omega_0\zeta t}\cos(\omega_0\sqrt{1 - \zeta^2}t) + c_2\mathrm{e}^{-\omega_0\zeta t}\sin(\omega_0\sqrt{1 - \zeta^2}t)$.

To determine the constants c_1 and c_2, one applies initial conditions. Let us keep things simple and prescribe initial conditions

$$\xi(0) = x_0,\ \frac{\mathrm{d}\xi}{\mathrm{d}t}(0) = x_0^{(1)}.$$

Skipping the tedious manipulations...

1. $\zeta^2 > 1$:

$$c_1 = \frac{\omega_0(\zeta + \sqrt{\zeta^2 - 1})x_0 + x_0^{(1)}}{2\omega_0\sqrt{\zeta^2 - 1}},$$
$$c_2 = \frac{-\omega_0(\zeta - \sqrt{\zeta^2 - 1})x_0 - x_0^{(1)}}{2\omega_0\sqrt{\zeta^2 - 1}};$$

2. $\zeta^2 = 1$:

$$c_1 = x_0,$$
$$c_2 = \omega_0\zeta x_0 + x_0^{(1)};$$

3. $\zeta^2 < 1$:

$$c_1 = x_0,$$
$$c_2 = \frac{\omega_0\zeta x_0 + x_0^{(1)}}{\omega_0\sqrt{1 - \zeta^2}}.$$

In Figure 4.2 we graph solutions for fixed ω_0 and varying ζ. We $\zeta > 0$ we say the

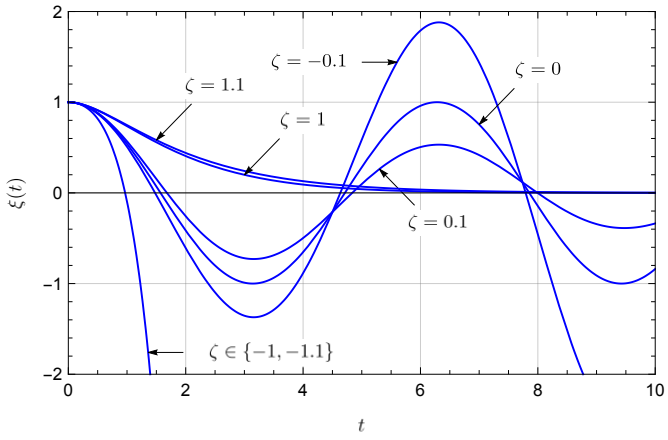


Figure 4.2 Solutions of a second-order scalar linear homogeneous ordinary differential equation with $\omega_0 = 1$, $\xi(0) = 1$, and $\frac{d\xi}{dt}(0) = 0$

equation is ***positively damped***, when $\zeta = 0$ we say the equation is ***undamped***, and when $\zeta < 0$ we say the equation is ***negatively damped***. In practice, systems are positively damped, or possibly undamped. So let us focus on this situation for a moment. Here we break things down into $\zeta < 1$, which is called ***underdamped***, $\zeta = 1$ which is called ***critically damped***, and $\zeta > 1$ which is called ***overdamped***. The underdamped case is distinguished by there being oscillations in the motion, corresponding to the imaginary part of the roots. In the critical and overdamped cases, we do not get this oscillation. •

Exercises

4.2.1 Consider the ordinary differential equation F with right-hand side given by (4.1).

(a) Convert this to a first-order equation with k states, following Exercise 3.1.23.

(b) Show that, if $a_0, a_1, \dots, a_k \in \mathsf{L}^1_{\text{loc}}(\mathbb{T}; \mathbb{R})$, then the resulting first-order equation satisfies the conditions of Theorem 3.2.8 for existence of a unique solution $t \mapsto \xi(t)$ satisfying the initial conditions

$$\xi(t_0) = x_0, \frac{\mathrm{d}\xi}{\mathrm{d}t}(t_0) = x_0^{(1)}, \dots, \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t)(t_0) = x_0^{(k-1)}$$

at time $t_0 \in \mathbb{T}$.

4.2.2 Let $a, b, c, \omega, \phi \in \mathbb{R}$ and define

$$\xi_1(t) = a\cos(\omega t + \phi), \quad \xi_2(t) = b\cos(\omega t) + c\sin(\omega t).$$

Show that $\xi_1, \xi_2 \in \mathrm{Sol}(F)$ where F is the second-order scalar linear homogeneous ordinary differential equation with constant coefficients whose right-hand side is

$$\widehat{F}(t, x, x^{(1)}) = -\omega^2 x.$$

Explain in at least two ways why this is not a violation of Proposition 4.2.1 concerning uniqueness of solutions.

4.2.3 In each of the following cases, show that the functions given are a basis for $\mathrm{Sol}(F)$ with F as given:

(a) take

$$F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} - x$$

and

$$\xi_1(t) = \mathrm{e}^t, \quad \xi_2(t) = \mathrm{e}^{-t};$$

(b) take

$$F(t, x, x^{(1)}, x^{(2)}, x^{(3)}) = x^{(3)} + 4x^{(2)} + 4x^{(1)}$$

and

$$\xi_1(t) = 1, \quad \xi_2(t) = \mathrm{e}^{-2t}, \quad \xi_3(t) = t\mathrm{e}^{-2t}.$$

(c) take

$$F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + \omega^2 x$$

and

$$\xi_1(t) = \cos(\omega t), \quad \xi_2(t) = \sin(\omega t).$$

(d) take

$$F(t, x, x^{(1)}, x^{(2)}) = t^2 x^{(2)} + t x^{(1)} - x$$

and

$$\xi_1(t) = t, \quad \xi_2(t) = t^{-1}$$

(here the time-domain must be an interval not containing 0).

4.2.4 For each of the ordinary differential equations F of Exercise 4.2.3, give the general form of a solution of the differential equation, i.e., the general form of $t \mapsto \xi(t)$ satisfying

$$F\left(t, \xi(t), \frac{\mathrm{d}\xi}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t)\right) = 0.$$

4.2.5 For each of the ordinary differential equations F of Exercise 4.2.3 for which you found a general form of their solution in Exercise 4.2.4, give the solution satisfying the given initial conditions:

(a) $\xi(0) = 1$ and $\dot{\xi}(0) = 1$;
(b) $\xi(0) = 1$, $\dot{\xi}(0) = 1$, and $\ddot{\xi}(0) = 1$;
(c) $\xi(0) = 1$ and $\dot{\xi}(0) = 0$;
(d) $\xi(1) = 1$ and $\dot{\xi}(1) = 1$.

4.2.6 If possible, find the characteristic polynomial for the following scalar ordinary differential equations:

(a) $F(t, x, x^{(1)}) = x^{(1)} + tx$;
(b) $F(t, x, x^{(1)}) = x^{(1)} + 3x$;
(c) $F(t, x, x^{(1)}, x^{(2)}) = 2x^{(2)} - x^{(1)} + 8x$;
(d) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + \frac{a_g}{\ell} \sin(x)$;
(e) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + \omega^2 x$;
(f) $F(t, x, x^{(1)}, \dots, x^{(k)}) = a_k x^{(k)} + \cdots + a_1 x^{(1)} + a_0 x$.

4.2.7 Find the unique lowest order normalised scalar linear homogeneous ordinary differential equation with constant coefficients whose characteristic polynomial has the following roots:

(a) $\{-1, 2\}$;
(b) $\{2 + 2\mathrm{i}, 2 - 2\mathrm{i}, -2\}$;
(c) $\{-\frac{1}{\tau}\}$, $\tau \in \mathbb{R} \setminus \{0\}$;
(d) $\{-a, -a, 2\}$, $a \in \mathbb{R}$;
(e) $\{\omega_0(-\zeta + \mathrm{i}\sqrt{1 - \zeta^2}), \omega_0(-\zeta - \mathrm{i}\sqrt{1 - \zeta^2})\}$, $\omega_0, \zeta \in \mathbb{R}$, $\omega_0 \neq 0$, $|\zeta| \leq 1$;
(f) $\{\sigma + \mathrm{i}\omega, \sigma - \mathrm{i}\omega\}$, $\sigma, \omega \in \mathbb{R}$, $\omega \neq 0$.

4.2.8 Find the unique lowest order normalised scalar linear homogeneous ordinary differential equation with the following functions as a fundamental set of solutions:

(a) $\xi_1(t) = \mathrm{e}^{-t}$ and $\xi_2(t) = \mathrm{e}^{2t}$;

(b) $\xi_1(t) = e^{2t}\cos(2t)$, $\xi_2(t) = e^{2t}\sin(2t)$, $\xi_3(t) = e^{-2t}$;
(c) $\xi_1(t) = e^{-t/\tau}$, $\tau \in \mathbb{R} \setminus \{0\}$;
(d) $\xi_1(t) = e^{-at}$, $\xi_2(t) = te^{-at}$, $a \in \mathbb{R}$, and $\xi_3(t) = e^2 t$;
(e) $\xi_1(t) = e^{-\omega_0\zeta t}\cos(\omega_0\sqrt{1-\zeta^2}t)$ and $\xi_2(t) = e^{-\omega_0\zeta t}\sin(\omega_0\sqrt{1-\zeta^2}t)$, $\omega_0, \zeta \in \mathbb{R}$, $\omega_0 \neq 0$, $|\zeta| \leq 1$;
(f) $\xi_1(t) = e^{\sigma t}\cos(\omega t)$ and $\xi_2(t) = e^{\sigma t}\sin(\omega t)$, $\sigma, \omega \in \mathbb{R}$, $\omega \neq 0$.

4.2.9 In Proposition 4.2.11 it is proved that the set of solutions for a scalar linear inhomogeneous ordinary differential with continuous coefficients uniquely determines the differential equation. Show how you would, given a fundamental set of solutions to a homogeneous such equation, with constant coefficients, recover the coefficients in the differential equation.

4.2.10 Solve the following initial value problems:

(a) $\dot{\xi}(t) + 3\xi(t) = 0$, $\xi(0) = 4$;
(b) $\ddot{\xi}(t) - 4\dot{\xi}(t) + 4\xi(t) = 0$, $\xi(0) = 0$, $\dot{\xi}(0) = 1$;
(c) $\ddot{\xi}(t) - 4\dot{\xi}(t) - 4\xi(t) = 0$, $\xi(0) = 1$, $\dot{\xi}(0) = 1$;
(d) $\dddot{\xi}(t) - 7\ddot{\xi}(t) + 15\dot{\xi}(t) - 9\xi(t) = 0$, $\xi(0) = 1$, $\dot{\xi}(0) = 1$, $\ddot{\xi}(0) = 1$;
(e) $\dddot{\xi}(t) + 3\ddot{\xi}(t) + 4\dot{\xi}(t) + 2\xi(t) = 0$, $\xi(0) = 0$, $\dot{\xi}(0) = 1$, $\ddot{\xi}(0) = 2$;
(f) $\ddddot{\xi}(t) + \dddot{\xi}(t) + \ddot{\xi}(t) + \dot{\xi}(t) + \xi(t) = 0$, $\xi(0) = 0$, $\dot{\xi}(0) = 0$, $\ddot{\xi}(0) = 0$, $\dddot{\xi}(0) = 0$.

Section 4.3

Scalar linear inhomogeneous ordinary differential equations

In this section we still consider scalar linear ordinary differential equations, but now we consider the inhomogeneous case. We still have the time-domain $\mathbb{T}$ and the state space $U = \mathbb{R}$, but now we have a right-hand side of the form

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0 x + b(t) \tag{4.11}$$

for functions $a_0, a_1, a_{k-1}, b \colon \mathbb{T} \to \mathbb{R}$. Thus solutions $t \mapsto \xi(t)$ satisfy

$$\frac{\mathrm{d}^k \xi(t)}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t)\xi(t) = b(t).$$

We shall proceed in this section much as in the preceding section, first saying some things about the general case, and then focussing on the case where F has constant coefficients, as in this case there is more that can be said.

Do I need to read this section? This section contains tools that are standard for anyone claiming to know something about ordinary differential equations. •

4.3.1 Equations with time-varying coefficients

We begin by stating some general properties of general scalar linear inhomogeneous ordinary differential equations.

4.3.1.1 Solutions and their properties First we state the local existence and uniqueness result that one needs to get off the ground for any class of differential equations.

4.3.1 Proposition (Local existence and uniqueness of solutions for scalar linear homogeneous ordinary differential equations) *Consider the linear inhomogeneous ordinary differential equation* F *with right-hand side equation (4.11) and suppose that* $a_0, a_1, \dots, a_{k-1} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. *Let*

$$(t_0, x_0, x_0^{(1)}, \dots, x_{k-1}^{(0)}) \in \mathbb{T} \times \mathbb{R} \oplus \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}).$$

Then there exists an interval $\mathbb{T}' \subseteq \mathbb{T}$ *and a map* $\xi \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T}'; \mathbb{R})$ *that is a solution for* F *and which satisfies*

$$\xi(t_0) = x_0, \frac{\mathrm{d}\xi}{\mathrm{d}t}(t_0) = x_0^{(1)}, \dots, \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t)(t_0) = x_0^{(k-1)}.$$

Moreover, if $\tilde{\mathbb{T}}' \subseteq \mathbb{T}$ *is another subinterval and if* $\tilde{\xi} \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\tilde{\mathbb{T}}'; \mathbb{R})$ *is another solution for* F *satisfying*

$$\tilde{\xi}(t_0) = x_0, \frac{d\tilde{\xi}}{dt}(t_0) = x_0^{(1)}, \dots, \frac{d^{k-1}\tilde{\xi}}{dt^{k-1}}(t)(t_0) = x_0^{(k-1)},$$

then $\tilde{\xi}(t) = \xi(t)$ *for every* $t \in \tilde{\mathbb{T}}' \cap \mathbb{T}'$.

Proof This is Exercise 4.3.1. ∎

As with homogeneous equations, for the scalar linear inhomogeneous ordinary differential equations we can show that solutions exist for all times.

4.3.2 Proposition (Global existence of solutions for scalar linear inhomogeneous ordinary differential equations) *Consider the linear in homogeneous ordinary differential equation* F *with right-hand side equation (4.11) and suppose that* $a_0, a_1, \dots, a_{k-1}, b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. *If* $\xi \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T}'; \mathbb{R})$ *is a solution for* F, *then there exists a solution* $\bar{\xi} \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$ *for which* $\bar{\xi}|\mathbb{T}' = \xi$.

Proof Note that, as per Exercise 3.1.23, we can convert the differential equation F into a first-order differential equation linear homogeneous differential equation with states $(x, x^{(1)}, \dots, x^{(k-1)})$. Thus the result will follow from the analogous result for first-order systems of equations, and this is stated and proved as Proposition 5.3.2. ∎

As in the homogeneous case, we can now talk sensibly about the set of *all* solutions for F. Thus we can define

$$\mathrm{Sol}(F) = \left\{ \xi \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T}; \mathbb{R}) \,\middle|\, \frac{d^k\xi}{dt^k}(t) + a_{k-1}(t)\frac{d^{k-1}\xi}{dt^{k-1}}(t) + \cdots + a_1(t)\frac{d\xi}{dt}(t) + a_0(t)\xi(t) = b(t), \text{ a.e. } t \in \mathbb{T} \right\}$$

which is exactly this set of all solutions for F. While $\mathrm{Sol}(F)$ was a vector space in the homogeneous case, in the inhomogeneous case this is no longer the case. However, the set of all solutions for the homogeneous case plays an important rôle, even in the homogeneous case. To organise this discussion, we let F_{h} be the "homogeneous part" of F. Thus the right-hand side of F_{h} is

$$\widehat{F_{\mathrm{h}}}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x.$$

As in Theorem 4.2.3, $\mathrm{Sol}(F_{\mathrm{h}})$ is a $\mathbb{R}$-vector space of dimension k. We can now state the character of $\mathrm{Sol}(F)$.

4.3.3 Theorem (Affine space structure of sets of solutions) *Consider the linear inhomogeneous ordinary differential equation* F *with right-hand side equation (4.11) and suppose that* $a_0, a_1, \dots, a_{k-1}, b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. *Let* $\xi_{\mathrm{p}} \in \mathrm{Sol}(F)$. *Then*

$$\mathrm{Sol}(F) = \{\xi + \xi_{\mathrm{p}} \mid \xi \in \mathrm{Sol}(F_{\mathrm{h}})\}.$$

Proof First note that, by Theorem 4.2.3, $\mathrm{Sol}(F) \neq \emptyset$ and so there exists some $\xi_{\mathrm{p}} \in \mathrm{Sol}(F)$. We have, of course,

$$\frac{\mathrm{d}^k \xi_{\mathrm{p}}}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi_{\mathrm{p}}}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi_{\mathrm{p}}}{\mathrm{d}t}(t) + a_0(t)\xi_{\mathrm{p}}(t) = b(t). \tag{4.12}$$

Next let $\xi \in \mathrm{Sol}(F)$ so that

$$\frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t)\xi(t) = b(t). \tag{4.13}$$

Subtracting (4.12) from (4.13) we get

$$\frac{\mathrm{d}^k (\xi - \xi_{\mathrm{p}})}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}(\xi - \xi_{\mathrm{p}})}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}(\xi - \xi_{\mathrm{p}})}{\mathrm{d}t}(t) + a_0(t)(\xi - \xi_{\mathrm{p}})(t) = 0,$$

i.e., $\xi - \xi_{\mathrm{p}} \in \mathrm{Sol}(F_{\mathrm{h}})$. In other words, $\xi = \tilde{\xi} + \xi_{\mathrm{p}}$ for $\tilde{\xi} \in \mathrm{Sol}(F_{\mathrm{h}})$.

Conversely, suppose that $\xi = \tilde{\xi} + \xi_{\mathrm{p}}$ for $\tilde{\xi} \in \mathrm{Sol}(F_{\mathrm{h}})$. Then

$$\frac{\mathrm{d}^k \tilde{\xi}}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\tilde{\xi}}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\tilde{\xi}}{\mathrm{d}t}(t) + a_0(t)\tilde{\xi}(t) = 0. \tag{4.14}$$

Adding (4.12) and (4.14) we get

$$\frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t)\xi(t) = b(t),$$

i.e., $\xi \in \mathrm{Sol}(F)$. ∎

It is interesting to make some comments on the preceding theorem in the language of basic problems in linear algebra.

4.3.4 Remark (Comparison of Theorem 4.3.3 with systems of linear algebraic equations) The reader should compare here the result of Theorem 4.3.3 with the situation concerning linear algebraic equations of the form $L(u) = v_0$, for vector spaces U and V, a linear map $L \in \mathrm{L}(\mathsf{U};\mathsf{V})$, and a fixed $v_0 \in \mathsf{V}$. In particular, we can make reference to Section I-5.4.8. In the setting of scalar linear inhomogeneous ordinary differential equations, we have

$$\begin{aligned}
\mathsf{U} &= \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T};\mathbb{R}),\\
\mathsf{V} &= \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathbb{R}),\\
L(f)(t) &= \frac{\mathrm{d}^k f}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1} f}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}f}{\mathrm{d}t}(t) + a_0(t)f(t),\\
v_0 &= b.
\end{aligned}$$

Then Propositions 4.3.1 and 4.3.2 tell us that L is surjective, and so $v_0 \in \mathrm{image}(L)$. Thus we are in case (ii) of Proposition I-5.4.48, which exactly the statement of Theorem 4.3.3. Note that L is not injective, since Theorem 4.2.3 tells us that $\dim_{\mathbb{R}}(\ker(L)) = k$. •

Note that Theorem 4.3.3 tells us that, to solve a scalar linear inhomogeneous ordinary differential equation, we must do two things: (1) find *some* solution for the equation; (2) find *all* solutions for the homogeneous part. Then we know our solution will be found amongst the set of sums of these. Generally, both of these things is impossible, in any general way. We do know, however, that Procedure 4.2.18 can be used, in principle, to find all solutions for the homogeneous part in the constant coefficient case. Thus one need only find some solution of the equation in this case. Upon finding such a solution, one calls it a ***particular solution***. *Note that there are many particular solutions.* Indeed, Proposition 4.2.1 tells us that there is one solution for every set of initial conditions. So one should always speak of *a* particular solution, not *the* particular solution.

4.3.1.2 Finding a particular solution using the Wronskian So... how do we find a particular solution? In this section we outline a general (and not very efficient) way of arriving at some such solution, using the Wronskian of Definition 4.2.6. To state the result, suppose that we have a fundamental set of solutions $\{\xi_1, \dots, \xi_k\}$ for F_{h}, where F has right-hand side (4.11), and denote

$$\mathrm{W}_{b,j}(\xi_1, \dots, \xi_k)(t) = \det \begin{bmatrix} \xi_1(t) & \cdots & \xi_{j-1}(t) & 0 & \xi_{j+1}(t) & \cdots & \xi_k(t) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{d}^{k-2}\xi_1}{\mathrm{d}t^{k-2}}(t) & \cdots & \frac{\mathrm{d}^{k-2}\xi_{j-1}}{\mathrm{d}t^{k-2}}(t) & 0 & \frac{\mathrm{d}^{k-2}\xi_{j+1}}{\mathrm{d}t^{k-2}}(t) & \cdots & \frac{\mathrm{d}^{k-2}\xi_k}{\mathrm{d}t^{k-2}}(t) \\ \frac{\mathrm{d}^{k-1}\xi_1}{\mathrm{d}t^{k-1}}(t) & \cdots & \frac{\mathrm{d}^{k-1}\xi_{j-1}}{\mathrm{d}t^{k-1}}(t) & b(t) & \frac{\mathrm{d}^{k-1}\xi_{j+1}}{\mathrm{d}t^{k-1}}(t) & \cdots & \frac{\mathrm{d}^{k-1}\xi_k}{\mathrm{d}t^{k-1}}(t) \end{bmatrix},$$

for $j \in \{1, \dots, k\}$, i.e., $\mathrm{W}_{b,j}(\xi_1, \dots, \xi_k)(t)$ is the determinant of the matrix used to compute the Wronskian, but with the jth column replaced by $(0, \dots, 0, b(t))$.

We then have the following result.

4.3.5 Proposition (A particular solution using Wronskians) *Consider the linear inhomogeneous ordinary differential equation* F *with right-hand side equation (4.11) and suppose* $a_0, a_1, \dots, a_{k-1}, b \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. *Let* $\{\xi_1, \dots, \xi_k\}$ *be a fundamental set of solutions for* $\mathrm{F_h}$ *and let* $t_0 \in \mathbb{T}$. *Then the function* $\xi_{\mathrm{p}} \colon \mathbb{T} \to \mathbb{R}$ *defined by*

$$\xi_{\mathrm{p}}(t) = \sum_{j=1}^{k} \xi_j(t) \int_{t_0}^{t} \frac{\mathrm{W}_{b,j}(\xi_1, \dots, \xi_k)(\tau)}{\mathrm{W}(\xi_1, \dots, \xi_k)(\tau)} \, \mathrm{d}\tau, \qquad \textit{a.e. } t \in \mathbb{T},$$

is a particular solution for F.

Proof Let us define

$$c_j(t) = \int_{t_0}^{t} \frac{\mathrm{W}_{b,j}(\xi_1, \dots, \xi_k)(\tau)}{\mathrm{W}(\xi_1, \dots, \xi_k)(\tau)} \, \mathrm{d}\tau, \qquad j \in \{1, \dots, k\}, \text{ a.e. } t \in \mathbb{T},$$

so that

$$\frac{\mathrm{d}c_j}{\mathrm{d}t}(t) = \frac{\mathrm{W}_{b,j}(\xi_1, \dots, \xi_k)(t)}{\mathrm{W}(\xi_1, \dots, \xi_k)(t)}, \qquad j \in \{1, \dots, k\}, \text{ a.e. } t \in \mathbb{T}.$$

Note that this is equivalent, by Cramer's Rule for linear systems of algebraic equations (Proposition I-5.3.12), to the set of equations

$$\begin{bmatrix} \xi_1(t) & \xi_2(t) & \cdots & \xi_k(t) \\ \frac{\mathrm{d}\xi_1}{\mathrm{d}t}(t) & \frac{\mathrm{d}\xi_2}{\mathrm{d}t}(t) & \cdots & \frac{\mathrm{d}\xi_k}{\mathrm{d}t}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{d}^{k-1}\xi_1}{\mathrm{d}t^{k-1}}(t) & \frac{\mathrm{d}^{k-1}\xi_2}{\mathrm{d}t^{k-1}}(t) & \cdots & \frac{\mathrm{d}^{k-1}\xi_k}{\mathrm{d}t^{k-1}}(t) \end{bmatrix} \begin{bmatrix} \frac{\mathrm{d}c_1}{\mathrm{d}t}(t) \\ \frac{\mathrm{d}c_2}{\mathrm{d}t}(t) \\ \vdots \\ \frac{\mathrm{d}c_k}{\mathrm{d}t}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ b(t) \end{bmatrix}, \qquad \text{a.e. } t \in \mathbb{T}. \tag{4.15}$$

Note that the proposition is then that

$$\xi_{\mathrm{p}}(t) = \sum_{j=1}^{k} c_j(t)\xi_j(t), \qquad \text{a.e. } t \in \mathbb{T},$$

defines a particular solution for F. This we shall prove by direct computation.

We compute

$$\frac{\mathrm{d}\xi_{\mathrm{p}}}{\mathrm{d}t}(t) = \sum_{j=1}^{k} \frac{\mathrm{d}c_j}{\mathrm{d}t}(t)\xi_j(t) + \sum_{j=1}^{k} c_j(t)\frac{\mathrm{d}\xi_j}{\mathrm{d}t}(t) = \sum_{j=1}^{k} c_j(t)\frac{\mathrm{d}\xi_j}{\mathrm{d}t}(t)$$

for $t \in \mathbb{T}$, using the first of equations (4.15). Repeatedly differentiating and using successive equations from (4.15), we deduce that

$$\frac{\mathrm{d}^l\xi_{\mathrm{p}}}{\mathrm{d}t^j}(t) = \sum_{j=1}^{k} c_j(t)\frac{\mathrm{d}^l\xi}{\mathrm{d}t^l}(t), \qquad l \in \{0, 1, \dots, k-1\}, \text{ a.e. } t \in \mathbb{T}.$$

We also have, using the last of equations (4.15),

$$\frac{\mathrm{d}^k\xi_{\mathrm{p}}}{\mathrm{d}t^k}(t) = \sum_{j=1}^{k} \frac{\mathrm{d}c_j}{\mathrm{d}t}(t)\frac{\mathrm{d}^{k-1}\xi_j}{\mathrm{d}t^{k-1}}(t) + \sum_{j=1}^{k} c_j(t)\frac{\mathrm{d}^k\xi_j}{\mathrm{d}t^k}(t) = \sum_{j=1}^{k} c_j(t)\frac{\mathrm{d}^k\xi_j}{\mathrm{d}t^k}(t) + b(t).$$

Therefore, combining these calculations,

$$\begin{aligned} &\frac{\mathrm{d}^k\xi_{\mathrm{p}}}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi_{\mathrm{p}}}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi_{\mathrm{p}}}{\mathrm{d}t}(t) + a_0(t)\xi(t) \\ &\quad = \sum_{j=1}^{k} c_j(t)\left(\frac{\mathrm{d}^k\xi_j}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi_j}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi_j}{\mathrm{d}t}(t) + a_0(t)\xi_j(t)\right) + b(t) = b(t), \end{aligned}$$

using the fact that $\xi_1, \dots, \xi_k$ are solutions for F_{h}. Thus ξ_{p} is indeed a particular solution. ■

Let us illustrate this result on an example.

4.3.6 Example (First-order scalar linear inhomogeneous ordinary differential equations) We consider here the first-order equation F with right-hand side

$$\widehat{F}(t,x) = -a(t)x + b(t)$$

for $a, b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathbb{R})$. We have seen in Example 4.2.5 that a fundamental set of solutions is given by $\{\xi_1(t)\}$, with

$$\xi_1(t) = \mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau}$$

for some $t_0 \in \mathbb{T}$. Therefore,

$$\mathrm{W}(\xi_1)(t) = \det\Big[\xi_1(t)\Big] = \xi_1(t), \quad \mathrm{W}(\xi_1)_{b,1} = \det\Big[b(t)\Big] = b(t).$$

Thus

$$\begin{aligned}\xi_{\mathrm{p}}(t) &= \xi_1(t)\left(\int_{t_0}^t \frac{b(\tau)}{\xi_1(\tau)}\,\mathrm{d}\tau\right)\\ &= \mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau}\int_{t_0}^t b(\tau)\mathrm{e}^{\int_{t_0}^\tau a(\sigma)\,\mathrm{d}\sigma}\mathrm{d}\tau\end{aligned}$$

defines a particular solution for F. Thus, as in Theorem 4.3.3, *any* solution for F has the form

$$\xi(t) = C\mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau} + \mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau}\int_{t_0}^t b(\tau)\mathrm{e}^{\int_{t_0}^\tau a(\sigma)\,\mathrm{d}\sigma}\mathrm{d}\tau$$

for some $C \in \mathbb{R}$. In we apply an initial condition $\xi(t_0) = x_0$, then we see that $C = x_0$. Therefore, finally, we have the solution

$$\xi(t) = x_0\mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau} + \mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau}\int_{t_0}^t b(\tau)\mathrm{e}^{\int_{t_0}^\tau a(\sigma)\,\mathrm{d}\sigma}\mathrm{d}\tau$$

to the initial value problem

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a(t)\xi(t) = b(t), \quad \xi(t_0) = x_0.$$

Because we have expressed the solution of a differential equation as an integral, we declare victory![3] •

4.3.1.3 The continuous-time Green's function

In this section we describe another means of determining a particular solution. In this case, what we arrive at is a description of a particular solution that allows for the inhomogeneous term "b" to be plugged into an integral. We shall see a close variant of this in Section 5.3 when we discuss linear inhomogeneous *systems* of equations.

The result is the following.

[3] Because victories are few and far between in the business of solving differential equations.

4.3.7 Theorem (Existence of, and properties of, the continuous-time Green's function) *Consider the linear homogeneous ordinary differential equation* F *with right-hand side equation (4.1) and suppose that* $a_0, a_1, \dots, a_{k-1} \in L^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. *Then there exists*

$$G_F \colon \mathbb{T} \times \mathbb{T} \to \mathbb{R}$$

with the following properties:

(i) $\dfrac{\partial^l G_F}{\partial t^l}$ *is continuous for* $l \in \{0, 1, \dots, k-2\}$;

(ii) $\dfrac{\partial^{k-1} G_F}{\partial t^{k-1}}$ *is continuous on*

$$\{(t, \tau) \in \mathbb{T} \times \mathbb{T} \mid t \neq \tau\};$$

(iii) for $\tau \in \mathbb{T}$, *we have*

$$\lim_{t \uparrow \tau} \frac{\partial^l G_F}{\partial t^l}(t, \tau) = 0, \quad \lim_{t \downarrow \tau} \frac{\partial^l G_F}{\partial t^l}(t, \tau) = 0, \qquad l \in \{0, 1, \dots, k-2\},$$

and

$$\lim_{t \uparrow \tau} \frac{\partial^{k-1} G_F}{\partial t^{k-1}}(t, \tau) = 0, \quad \lim_{t \downarrow \tau} \frac{\partial^{k-1} G_F}{\partial t^{k-1}}(t, \tau) = 1;$$

(iv) for $t \in \mathbb{T} \setminus \{\tau\}$ *we have*

$$\frac{\partial^k G_F}{\partial t^k}(t, \tau) + a_{k-1}(t) \frac{\partial^{k-1} G_F}{\partial t^{k-1}}(t, \tau) + \cdots + a_1(t) \frac{\partial G_F}{\partial t}(t, \tau) + a_0(t) G_F(t, \tau) = 0;$$

(v) if $b \in L^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$, *if* $t_0 \in \mathbb{T}$, *and if* $\xi_{p,b} \colon \mathbb{T}_{\geq t_0} \to \mathbb{R}$ *is given by*

$$\xi_{p,b}(t) = \int_{t_0}^{t} G_F(t, \tau) b(\tau) \, d\tau,$$

then $\xi_{p,b}$ *solves the initial value problem*

$$\frac{d^k \xi}{dt^k}(t) + a_{k-1}(t) \frac{d^{k-1} \xi}{dt^{k-1}}(t) + \cdots + a_1(t) \frac{d\xi}{dt}(t) + a_0(t) \xi(t) = b(t),$$
$$\xi(t_0) = \cdots = \frac{d^{k-1} \xi}{dt^{k-1}}(t_0) = 0.$$

Moreover, there is only one such function satisfying all of the above properties.

Proof For $\tau \in \mathbb{T}$, let $\xi_\tau \colon \mathbb{T} \to \mathbb{R}$ be the solution to the initial value problem

$$\frac{d^k \xi}{dt^k}(t) + a_{k-1}(t) \frac{d^{k-1} \xi}{dt^{k-1}}(t) + \cdots + a_1(t) \frac{d\xi}{dt}(t) + a_0(t) \xi(t) = 0,$$
$$\xi(\tau) = \cdots = \frac{d^{k-2} \xi}{dt^{k-2}}(\tau) = 0, \ \frac{d^{k-1} \xi}{dt^{k-1}}(\tau) = 1.$$

Let $\{\xi_1, \dots, \xi_k\}$ be a fundamental set of solutions and write

$$\xi_\tau(t) = \sum_{j=1}^{k} c_j(\tau)\xi_j(t), \qquad t \in \mathbb{T}_{\geq\tau}.$$

The specified initial conditions for ξ_τ can then be written in matrix form as

$$\begin{bmatrix} \xi_1(\tau) & \cdots & \xi_k(\tau) \\ \frac{\mathrm{d}\xi_1}{\mathrm{d}t}(\tau) & \cdots & \frac{\mathrm{d}\xi_k}{\mathrm{d}t}(\tau) \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}^{k-1}\xi_1}{\mathrm{d}t^{k-1}}(\tau) & \cdots & \frac{\mathrm{d}^{k-1}\xi_k}{\mathrm{d}t^{k-1}}(\tau) \end{bmatrix} \begin{bmatrix} c_1(\tau) \\ c_2(\tau) \\ \vdots \\ c_k(\tau) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Following the construction preceding the statement of Proposition 4.3.5, denote

$$\mathrm{W}_j(\xi_1, \dots, \xi_k)(t) = \det \begin{bmatrix} \xi_1(t) & \cdots & \xi_{j-1}(t) & 0 & \xi_{j+1}(t) & \cdots & \xi_k(t) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{d}^{k-2}\xi_1}{\mathrm{d}t^{k-2}}(t) & \cdots & \frac{\mathrm{d}^{k-2}\xi_{j-1}}{\mathrm{d}t^{k-2}}(t) & 0 & \frac{\mathrm{d}^{k-2}\xi_{j+1}}{\mathrm{d}t^{k-2}}(t) & \cdots & \frac{\mathrm{d}^{k-2}\xi_k}{\mathrm{d}t^{k-2}}(t) \\ \frac{\mathrm{d}^{k-1}\xi_1}{\mathrm{d}t^{k-1}}(t) & \cdots & \frac{\mathrm{d}^{k-1}\xi_{j-1}}{\mathrm{d}t^{k-1}}(t) & 1 & \frac{\mathrm{d}^{k-1}\xi_{j+1}}{\mathrm{d}t^{k-1}}(t) & \cdots & \frac{\mathrm{d}^{k-1}\xi_k}{\mathrm{d}t^{k-1}}(t) \end{bmatrix},$$

for $j \in \{1, \dots, k\}$. Then

$$c_j(\tau) = \frac{\mathrm{W}_j(\xi_1, \dots, \xi_k)(\tau)}{\mathrm{W}(\xi_1, \dots, \xi_k)(\tau)}.$$

This allows us to conclude that $c_1, \dots, c_k \in \mathsf{C}^0(\mathbb{T}; \mathbb{R})$.

Now define

$$G_F(t, \tau) = \begin{cases} \xi_\tau(t), & t \geq \tau, \\ 0, & t < \tau. \end{cases}.$$

With this definition of G_F, let us check off the assertions of the theorem.

First note that, since $\xi_\tau \in \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$, the partial derivatives $\frac{\partial^l G_F}{\partial t^l}$, $l \in \{0, 1, \dots, k-1\}$, are immediately continuous on

$$\{(t, \tau) \in \mathbb{T} \times \mathbb{T} \mid t \neq \tau\}.$$

Now let $\tau \in \mathbb{T}$ and let $((t_j, \tau_j))_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{T}\times\mathbb{T}$ converging to (τ, τ). If $t_j < \tau_j$, then

$$\frac{\partial^l G_F}{\partial t^l}(t_j, \tau_j) = 0, \qquad l \in \{0, 1, \dots, k-2\}.$$

Thus, without loss of generality, suppose that $t_j \geq \tau_j$, $j \in \mathbb{Z}_{>0}$. The initial conditions for ξ_τ then ensure that

$$\lim_{j\to\infty} \frac{\partial^l G_F}{\partial t^l}(t_j, \tau_j) = \lim_{j\to\infty} \sum_{j=1}^{k} c_j(\tau_j) \frac{\partial^l \xi_\tau}{\partial t^l}(t_j) = 0, \qquad l \in \{0, 1, \dots, k-2\}.$$

This gives continuity of $\frac{\partial^l G_F}{\partial t^l}$, $l \in \{0, 1, \dots, k-2\}$, on $\mathbb{T} \times \mathbb{T}$. This gives parts (i) and (ii). By definition, we have parts (iii) and (iv).

Let $b \in \mathsf{L}^1_{\text{loc}}(\mathbb{T}; \mathbb{R})$ and, for $t_0 \in \mathbb{T}$, define $\xi_{\text{p},b} \colon \mathbb{T}_{\geq t_0} \to \mathbb{R}$ by

$$\xi_{\text{p},b}(t) = \int_{t_0}^{t} G_F(t, \tau) b(\tau) \, \mathrm{d}\tau.$$

For part (v), we must show that $\xi_{\text{p},b}$ solves the initial value problem in the theorem statement. We can inductively compute

$$\frac{\mathrm{d}^l \xi_{\text{p},b}}{\mathrm{d}t^l}(t) = \lim_{\tau \uparrow t} \frac{\partial^{l-1} G_F}{\partial t^{l-1}}(t, \tau) b(t) + \int_{t_0}^{t} \frac{\partial^l G_F}{\partial t^l}(t, \tau) b(\tau) \, \mathrm{d}\tau, \qquad l \in \{0, 1, \dots, k\}.$$

Since

$$\lim_{\tau \uparrow t} \frac{\partial^{l-1} G_F}{\partial t^{l-1}}(t, \tau) = 0, \qquad l \in \{0, 1, \dots, k-1\},$$

by parts (i) and (iii), we have

$$\frac{\mathrm{d}^l \xi_{\text{p},b}}{\mathrm{d}t^l}(t) = \int_{t_0}^{t} \frac{\partial^l G_F}{\partial t^l}(t, \tau) b(\tau) \, \mathrm{d}\tau, \qquad l \in \{0, 1, \dots, k-1\}. \tag{4.16}$$

Similarly, by parts (i) and (iii), we have

$$\frac{\mathrm{d}^k \xi_{\text{p},b}}{\mathrm{d}t^k}(t) = b(t) + \int_{t_0}^{t} \frac{\partial^k G_F}{\partial t^k}(t, \tau) b(\tau) \, \mathrm{d}\tau. \tag{4.17}$$

Combining (4.16) and (4.17), and using part (iv), we have, for $t \in \mathbb{T}_{\geq t_0}$,

$$\frac{\partial^k \xi_{\text{p},b}}{\partial t^k}(t) + a_{k-1}(t) \frac{\mathrm{d}^{k-1} \xi_{\text{p},b}}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t) \frac{\mathrm{d}\xi_{\text{p},b}}{\mathrm{d}t}(t) + a_0(t) \xi_{\text{p},b}(t) = b(t),$$

giving (v).

The final uniqueness assertion of the theorem is obtained from the following observations:

1. for $t < \tau$, $t \mapsto G_F(t, \tau)$ is the unique element of $\mathrm{Sol}(F)$ with initial conditions

$$\lim_{t \uparrow \tau} \frac{\partial^l G_F}{\partial t^l}(t, \tau) = 0, \qquad l \in \{0, 1, \dots, k-1\};$$

2. for $t \geq \tau$, $t \mapsto G_F(t, \tau)$ is the unique element of $\mathrm{Sol}(F)$ with initial conditions

$$\frac{\partial^l G_F}{\partial t^l}(\tau, \tau) = 0, \qquad l \in \{0, 1, \dots, k-2\},$$
$$\frac{\partial^{k-1} G_F}{\partial t^{k-1}}(\tau, \tau) = 1.$$

These, combined with Proposition 4.3.1, give the theorem. ∎

Of course, we can give a name to the function G_F from the preceding theorem.

4.3.8 Definition (Continuous-time Green's function) Consider the linear homogeneous ordinary differential equation F with right-hand side equation (4.11) and suppose that $a_0, a_1, \dots, a_{k-1} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. The function G_F of Theorem 4.3.7 is the ***continuous-time Green's function*** for F. •

There are a few observations one can make about the continuous-time Green's function.

4.3.9 Remarks (Attributes of the continuous-time Green's function)

1. As we observed in Remark 4.3.4, the mapping

$$\begin{aligned} L_F \colon \mathsf{AC}^{k-1}_{\mathrm{loc}}(\mathbb{T}; \mathbb{R}) &\to \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R}) \\ \xi &\mapsto F_{\mathrm{h}}\left(t, \xi(t), \frac{\mathrm{d}\xi}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t)\right) \end{aligned}$$

   is surjective, and so, for any $b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$, there exists one (indeed, many by Theorem 4.3.3), solution of the differential equation with solutions

$$F_{\mathrm{h}}\left(t, \xi(t), \frac{\mathrm{d}\xi}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t)\right) = b(t).$$

   One can think of the mapping

$$b \mapsto \left(t \mapsto \int_{t_0}^{t} G_F(t, \tau) b(\tau)\, \mathrm{d}\tau\right) \tag{4.18}$$

   as prescribing a right-inverse of L_F. Of course, the prescription of a particular right-inverse amounts to a prescription for choosing initial conditions, since initial conditions are what distinguish elements of $\mathrm{Sol}(F)$. We refer the reader to Exercise 4.3.2 for just what initial condition choice is being made by the assignment (4.18).

2. There is also a physical interpretation of the mapping $t \mapsto G_F(t, \tau)$. For $t < \tau$, the solution is zero, until something happens at $t = \tau$. At $t = \tau$, we imagine the system being given an "impulse" i.e., a short duration, large magnitude input. If the area under the graph of this impulse is 1, this will give a jolt to the kth derivative of G_F at $t = \tau$. This discontinuity when integrated, will give an input of 1 to the $(k-1)$st derivative, resulting in the initial conditions of part (iii) of Theorem 4.3.7.

   This nonsense can be made precise using the theory of distributions, and we do this in Theorem 4.4.5 below. In system theory, this is connected to the "impulse response" which plays an important rôle, as we shall see in . • what

Let us give the simplest possible example to illustrate the use of the continuous-time Green's function.

4.3.10 Example (Continuous-time Green's function for first-order scalar linear ordinary differential equation) We consider the first order equation F with right-hand side

$$\widehat{F}(t, x) = -a(t)x.$$

Let us take $\mathbb{T}$ to be the time-domain for the equation. The way one determines the continuous-time Green's function is by first taking $\tau \in \mathbb{T}$ and then solving the initial value problem

$$\dot{\xi}(t) + a(t)\xi(t) = 0, \quad \xi(\tau) = 1,$$

just as prescribed in part (iii) of Theorem 4.3.7. However, in Example 4.2.5 we obtained the solution to this initial value problem as

$$\xi(t) = \mathrm{e}^{-\int_\tau^t a(s)\,\mathrm{d}s}.$$

Then the continuous-time Green's function is given by

$$G_F(t, \tau) = \begin{cases} 0, & t < \tau, \\ \mathrm{e}^{-\int_\tau^t a(s)\,\mathrm{d}s}, & t \geq \tau. \end{cases}$$

Therefore, given $b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$, the solution to the initial value problem

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a(t)\xi(t) = b(t), \quad \xi(t_0) = 0,$$

is given by

$$\xi_{\mathrm{p},b}(t) = \int_{t_0}^t \mathrm{e}^{-\int_\tau^t a(s)\,\mathrm{d}s} b(\tau)\,\mathrm{d}\tau.$$

Note that this is, in this first-order case, the same particular solution as in Example 4.3.6 using the Wronskian method of Proposition 4.3.5. This is simply because both solutions satisfy the same initial value problem. To rectify that the solutions are, in fact the same, can be done by a change of integration variable. •

We plot the graph of G_F in the case of $\mathbb{T} = [0, \infty)$ and $a(t) = 1$ in Figure 4.3. •

4.3.11 Remark (Continuous-time Green's function for constant coefficient equations and convolution) Suppose that F is a kth-order scalar linear inhomogeneous ordinary differential equation with constant coefficients, and take $\mathbb{T} = [0, \infty)$. As in the statement of Theorem 4.3.7, for each $\tau \in \mathbb{T}$, $t \mapsto G_F(t, \tau)$ is a solution for F satisfying the initial conditions

$$\begin{aligned} \frac{\partial^j G_F}{\partial t^j}(\tau, \tau) &= 0, \qquad j \in \{0, 1, \ldots, k-2\}, \\ \frac{\partial^{k-1} G_F}{\partial t^{k-1}}(\tau, \tau) &= 1. \end{aligned}$$

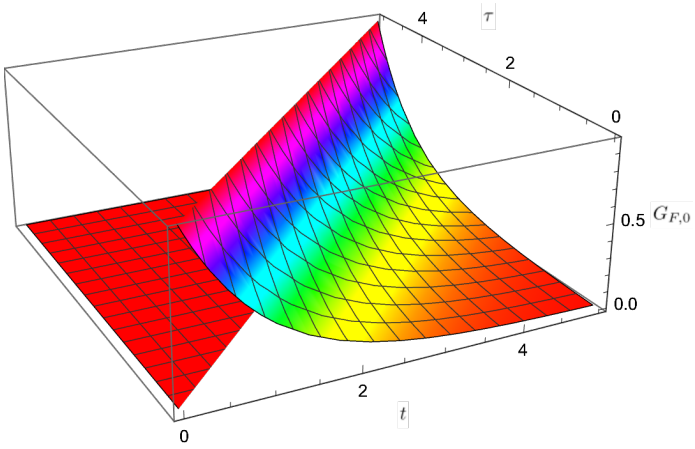


Figure 4.3 The continuous-time Green's function for a scalar linear ordinary differential equation with constant coefficients

Since F has constant coefficients, it is autonomous, and so by Exercise 3.1.19 there exists $H_F\colon \mathbb{T} \to \mathbb{R}$ such that $G_F(t,\tau) = H_F(t-\tau)$. Then, if we add an inhomogeneous term b to F, the particular solution of Theorem 4.3.7(v) is

$$\xi_{\mathrm{p},b}(t) = \int_0^t H_F(t-\tau)b(\tau)\,\mathrm{d}\tau.$$

Integrals of the type

$$\int f(t-\tau)g(\tau)\,\mathrm{d}\tau$$

are known as ***convolution integrals***. These arise in system theory, Fourier theory, and approximation theory, for example. We shall consider convolution in the context of transform theory in .

- better forward refs

4.3.2 Equations with constant coefficients

We now specialise the general discussion from the preceding section to equations with constant coefficients. Thus we are looking at scalar linear inhomogeneous ordinary differential equations with right-hand sides given by

$$\widehat{F}(t,x,x^{(1)},\dots,x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1x^{(1)} - a_0x + b(t) \tag{4.19}$$

for $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$ and $b\colon \mathbb{T} \to \mathbb{R}$. Thus a solution $t \mapsto \xi(t)$ satisfies the equation

$$\frac{\mathrm{d}^k\xi(t)}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0\xi(t) = b(t). \tag{4.20}$$

These equations are, of course, a special case of the equations considered in Section 4.3.1, and so all statements made about the general case of time-varying

coefficients hold in the special case of constant coefficients. In particular, Propositions 4.3.1 and 4.3.2, and Theorem 4.3.3 hold for equations of the form (4.20). However, for these constant coefficient equations, it is possible to say some things a little more explicitly, and this is what we undertake to do.

4.3.2.1 The "method of undetermined coefficients" We present in this section a so-called method for solving scalar linear inhomogeneous ordinary differential equations with constant coefficients. With this method, one guesses a form of particular solution based on the form of the function b, and then does algebra to determine the precise solution. The advantages to this method are

1. it does not require first finding a fundamental set of solutions, as in Proposition 4.3.5,
2. it is in principle possible for a brainless monkey to apply the method, and
3. it is an excellent source of mindless computations that students can be forced to do for marks in homework and on exams.

The disadvantages of the method are

1. it only works for *very* specific functions b, and so does not work most of the time,
2. even when it does work, it is tedious and likely to produce errors when used in the hands of most humans,
3. it is 2016, for crying out loud, and there are computer packages that do this sort of thing in their sleep!

What we shall do is (1) describe when the method applies, (2) describe how one uses the method, and (3) reiterate the silliness of the method at the end of the discussion.

First let us indicate the sorts of "b's" we allow.

4.3.12 Definition (Pretty uninteresting function) Let $\mathbb{T} \subseteq \mathbb{R}$ be an interval. A function $f : \mathbb{T} \to \mathbb{R}$ is ***pretty uninteresting*** if it has one of the following three forms:

(i) $f(t) = t^m e^{rt}$ for $m \in \mathbb{Z}_{\geq 0}$ and $r \in \mathbb{R}$;

(ii) $f(t) = t^m e^{\sigma t} \cos(\omega t)$ for $m \in \mathbb{Z}_{\geq 0}$, $\sigma \in \mathbb{R}$, and $\omega \in \mathbb{R}_{>0}$;

(iii) $f(t) = t^m e^{\sigma t} \sin(\omega t)$ for $m \in \mathbb{Z}_{\geq 0}$, $\sigma \in \mathbb{R}$, and $\omega \in \mathbb{R}_{>0}$.

The nonnegative integer m in the above forms is the ***order*** of f and is denoted by $o(f)$. If $f : \mathbb{T} \to \mathbb{R}$ has the form

$$f(t) = c_1 f_1(t) + \cdots + c_r f_r(t)$$

where $c_1, \ldots, c_r \in \mathbb{R}$ and each of $f_1, \ldots, f_r$ is pretty uninteresting, then f is ***also pretty uninteresting***. •

Here are some examples of useful pretty uninteresting functions.

4.3.13 Examples (Examples of interesting pretty uninteresting functions)

1. Consider the function $1_{\geq 0}\colon [0,\infty) \to \mathbb{R}$ defined by $1_{\geq 0}(t) = 1$ for all $t \in [0,\infty)$. This is a "step function" and is pretty uninteresting. Often it is taken to be defined on all of $\mathbb{R}$, and to be zero for negative times. The idea is that it gives an input to a differential equation that "switches on" at $t = 0$. Among the many particular solutions for a differential equation with $b = 1_{\geq 0}$, there is one that is known as the "step response," and it is determined by a specific choice of initial condition. Students going on to take a course in system theory will learn about this.
2. Next consider the function $H_\omega\colon [0,\infty) \to \mathbb{R}$ defined by $H_\omega(t) = \sin(\omega t)$ for $\omega \in \mathbb{R}_{>0}$. This is an example of an "harmonic" function, and specifically is a "sinusoid." In this case, one can think of prescribing a "b" of this form as "shaking" a differential equation. It can be interesting to know how the behaviour of the system will vary as we change ω. This gives rise in system theory to something called the "frequency response." •

We now state a few elementary properties of pretty uninteresting functions.

4.3.14 Lemma (Properties of pretty uninteresting functions)

Let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, let $f, f_1, \dots, f_r\colon \mathbb{T} \to \mathbb{R}$ be pretty uninteresting functions, and consider a scalar linear homogeneous ordinary differential equation F with constant coefficients with right-hand side of the form (4.19). Define normalised scalar linear inhomogeneous ordinary differential equations F_j, $j \in \{1,\dots,r\}$, by

$$F_j(t, x, x^{(1)}, \dots, x^{(k)}) = x^{(k)} - \widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) - f_j(t).$$

Then the following statements hold:

(i) there exists a unique normalised scalar linear homogeneous ordinary differential equation F_f of order $o(f)$ such that

$$F_f\left(t, f(t), \frac{df}{dt}(t), \dots, \frac{d^{o(f)}f}{dt^{o(f)}}(t)\right) = 0, \qquad t \in \mathbb{T};$$

(ii) if $\xi_j \in \mathrm{Sol}(F_j)$, $j \in \{1,\dots,r\}$, and if

$$g = c_1 f_1 + \cdots + c_r f_r$$

is also pretty uninteresting, then, if $\xi = c_1\xi_1 + \cdots + c_r\xi_r$, then $\xi \in \mathrm{Sol}(F_g)$, where

$$F_j(t, x, x^{(1)}, \dots, x^{(k)}) = x^{(k)} - \widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) - g(t).$$

Proof (i) An examination of Procedure 4.2.18 and the attendant Theorem 4.2.19 shows that F_f can be defined by defining their characteristic polynomials as follows:

1. $f(t) = t^m e^{rt}$: take
$$P_{F_f} = (X - r)^{m+1};$$

2. $f(t) = t^m e^{\sigma t}\cos(\omega t)$ or $f(t) = t^m e^{\sigma t}\sin(\omega t)$: take

$$P_{F_f} = ((X-\sigma)^2 + \omega^2)^{m+1}.$$

(ii) This is a mere verification, once one understands the symbols involved. ■

The differential equation F_f in the first part of the lemma we call the ***annihilator*** of the pretty uninteresting function f. The following examples illustrate how one finds the annihilator in practice, based on the proof of the first part of the lemma.

4.3.15 Examples (Annihilator)

1. Consider the function $f(t) = 1$. This is the pretty uninteresting function $t \mapsto t^k e^{rt}$ with $k = 0$ and $r = 0$. This corresponds, from Procedure 4.2.18, to a root $r = 0$ of a polynomial with multiplicity 1. Thus $P_{F_f} = X$, and so

$$F_f(t, x, x^{(1)}) = x^{(1)}.$$

2. Now consider $f(t) = e^{-2t}$. This is the pretty uninteresting function $t \mapsto t^k e^{\sigma t}\cos(\omega t)$ with $k = 0$, $\sigma = -2$, and $\omega = 0$. This corresponds to a root $r = -2$ of a polynomial with multiplicity 1. Thus $P_{F_f} = X + 2$ and so

$$F_f(t, x, x^{(1)}) = x^{(1)} + 2x.$$

3. Next we take $f(t) = 2e^{3t}\sin(2t) + t^2$. This is an also pretty uninteresting function, being a linear combination of $f_1(t) = e^{3t}\sin(2t)$ and $f_2(t) = t^2$.
   Note that f_1 is the pretty uninteresting function $t \mapsto t^k e^{\sigma t}\sin(\omega t)$ with $k = 0$, $\sigma = 3$, and $\omega = 2$. This function is associated, via Procedure 4.2.18, with a root $\rho = 3 + 2\mathrm{i}$ of a polynomial with multiplicity 1. Of course, we must also have the root $\bar{\rho} = 3 - 2\mathrm{i}$.
   Note that f_2 is the pretty uninteresting function $t \mapsto t^k e^{\sigma t}\cos(\omega t)$ with $k = 2$, $\sigma = 0$, and $\omega = 0$. This is associated with a root $r = 0$ with multiplicity 3.
   Putting this all together,

$$P_{F_f} = (X - (3 + 2\mathrm{i}))(X - (3 - 2\mathrm{i}))X^3 = X^5 - 6X^4 + 13X^3.$$ •

The second part of the lemma points out, in short, the obvious fact that if "b" is also pretty uninteresting, then one can obtain a particular solution by obtaining a particular solution for each of its pretty uninteresting components, and then summing these with the same coefficients as in the also pretty uninteresting function. The point of this is that, to obtain a particular solution for an also pretty uninteresting "b," it suffices to know how to do this for a pretty uninteresting b. Thus we deliver the following construction.

4.3.16 Procedure (Method of undetermined coefficients) We let F be a normalised scalar linear inhomogeneous ordinary differential equation with constant coefficients with right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x + f(t),$$

where f is pretty uninteresting. Do the following.

1. Let F_f be the annihilator of f.
2. Let G_f be the normalised scalar linear homogeneous ordinary differential equation whose characteristic polynomial is $P_{G_f} = P_{F_f} P_{F_h}$.
3. Using Procedure 4.2.18, find
   (a) pretty uninteresting functions $\xi_1, \dots, \xi_k$ for which $\{\xi_1, \dots, \xi_k\}$ is a fundamental set of solutions for F_h and
   (b) pretty uninteresting functions $\eta_1, \dots, \eta_{o(f)+1}$ for which $\{\xi_1, \dots, \xi_k, \eta_1, \dots, \eta_{o(f)+1}\}$ is a fundamental set of solutions for G_f.
4. For (as yet) undetermined coefficients $c_1, \dots, c_{o(f)+1} \in \mathbb{R}$, denote
$$\xi_p = c_1\eta_1 + \cdots + c_{o(f)+1}\eta_{o(f)+1}.$$
5. Determine $c_1, \dots, c_{o(f)+1}$ by demanding that ξ_p be a particular solution for F.

We shall show that this procedure makes sense and defines a particular solution for F. •

Let us verify that the preceding procedure gives what we want.

4.3.17 Proposition (Validity of the method of undetermined coefficients) *Let* F *be a normalised scalar linear inhomogeneous ordinary differential equation with constant coefficients with right-hand side*

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x + f(t),$$

where f *is pretty uninteresting. Then all steps in Procedure 4.3.16 are unambiguously defined, and the result is a particular solution for* F.

Proof In the proof we shall assume that $f(t) = t^{o(f)}e^{rt}$ for $r \in \mathbb{R}$. Entirely similar reasoning works for the other two sorts of pretty uninteresting functions.

From Procedure 4.2.18 we know that $P_{F_f} = (X - r)^{o(f)+1}$. Let us suppose that

$$P_{F_h} = (X - r)^{m(r)}P,$$

where P does not have r as a root. Therefore,

$$P_{G_f} = (X - r)^{m(r)+o(f)+1}P.$$

Then, according to Procedure 4.2.18, among the pretty uninteresting solutions for F_h are

$$t \mapsto t^j e^{rt}, \qquad j \in \{0, 1, \dots, m(r) - 1\}.$$

The rest of the pretty uninteresting solutions for F_h have nothing to do with the root "r" of the characteristic polynomial, and are not interesting to us here. Now the $o(f)+1$ pretty uninteresting solutions for G_f that are added to those for F_h are

$$t \mapsto t^j \mathrm{e}^{rt}, \qquad j \in \{m(r), \dots, m(r) + o(f)\},$$

again according to Procedure 4.2.18. This demonstrates the viability of the first three steps of Procedure 4.2.18. We now need to show that one can solve for the coefficients $c_1, \dots, c_{o(f)+1}$ to obtain a particular solution for F. If

$$\xi_\mathrm{p}(t) = c_1 t^{m(r)} \mathrm{e}^{rt} + \cdots + c_{o(f)+1} t^{m(r)+o(f)} \mathrm{e}^{rt},$$

then Lemma 1 from the proof of Theorem 4.2.19 shows that

$$\left(\frac{\mathrm{d}^{m(r)}}{\mathrm{d}t^{m(r)}} - r \right) \xi_\mathrm{p}(t)$$

is an also pretty uninteresting function associated with the root r whose highest order (as a pretty uninteresting function) term is of order $o(f)$. By Corollary 4.2.17, and since the derivative of a pretty uninteresting function of order m associated with the root r is an also pretty uninteresting function of order m associated with the root r (as can be verified by a direct computation), we have that

$$F_\mathrm{h}\left(t, \xi_\mathrm{p}(t), \frac{\mathrm{d}\xi_\mathrm{p}}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^k \xi_\mathrm{p}}{\mathrm{d}t^k}(t) \right)$$

is an also pretty uninteresting function of order $o(f)$. Therefore, we can use the equality

$$F_\mathrm{h}\left(t, \xi_\mathrm{p}(t), \frac{\mathrm{d}\xi_\mathrm{p}}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^k \xi_\mathrm{p}}{\mathrm{d}t^k}(t) \right) = f(t)$$

to solve for the coefficients $c_1, \dots, c_{o(f)+1}$, as asserted in Procedure 4.2.18. ■

While the preceding discussion does indeed provide a means of solving, in principle, scalar linear inhomogeneous ordinary differential equations with also pretty uninteresting "b's," it does tend to be a lot of work, cf. Example 4.3.18, and there are precisely zero equations that can be solved by this procedure that cannot far more easily be solved with a computer.

4.3.2.2 Some examples We carry on with the three examples of Section 4.2.2.4. Thus we first give an "academic" example to illustrate Procedure 4.3.16. Then we consider a first- and second-order system with specific "b's," in order to discuss some features of the solutions in these cases.

4.3.18 Example ("Academic" example) We continue the example of Example 4.2.20, now adding an inhomogeneous term. Specifically, we consider the 4th-order scalar linear homogeneous ordinary differential equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}, x^{(2)}, x^{(3)}) = -5x + 8x^{(1)} - 2x^{(2)} + te^t + 2\cos(2t).$$

Thus solutions $t \mapsto \xi(t)$ to this equation satisfy

$$\frac{d^4\xi}{dt^4}(t) + 2\frac{d^2\xi}{dt^2}(t) - 8\frac{d\xi}{dt}(t) + 5\xi(t) = te^t + 2cos(2t).$$

The right-hand side of this equation has the form $b(t) = f_1(t) + 2f_2(t)$ for the two pretty uninteresting functions

$$f_1(t) = te^t, \quad f_2(t) = \cos(2t).$$

We find two particular solutions $\xi_{p,1}$ and $\xi_{p,2}$, satisfying

$$\frac{d^4\xi_{p,1}}{dt^4}(t) + 2\frac{d^2\xi_{p,1}}{dt^2}(t) - 8\frac{d\xi_{p,1}}{dt}(t) + 5\xi_{p,1}(t) = te^t$$

and

$$\frac{d^4\xi_{p,2}}{dt^4}(t) + 2\frac{d^2\xi_{p,2}}{dt^2}(t) - 8\frac{d\xi_{p,2}}{dt}(t) + 5\xi_{p,2}(t) = \cos(2t),$$

and then, by Lemma 4.3.14(ii),

$$\xi_p = \xi_{p,1} + 2\xi_{p,2}$$

is a particular solution.

Let us find $\xi_{p,1}$ corresponding to $f_1(t) = te^t$. The annihilator F_{f_1} of f_1 has characteristic polynomial $P_{F_{f_1}} = (X-1)^2$. We have

$$P_{F_{f_1}} P_{F_h} = (X-1)^2(X-1)^2(X^2+2X+5) = (X-1)^4(X^2+2X+5)$$

as the characteristic polynomial of $F_{f_1} \circ F_h$. According to Procedure 4.2.18, a fundamental set of solutions, each of which is a pretty uninteresting function, is given by

$$e^{-t}\cos(2t),\ e^{-t}\sin(2t),\ e^t,\ te^t,\ t^2e^t,\ t^3e^t.$$

The first four of these are solutions for F_h. So we form our candidate particular solution from the last two functions:

$$\xi_{p,1}(t) = c_1t^2e^t + c_2t^3e^t.$$

To determine c_1 and c_2, we compute

$$\left(\frac{d^4}{dt^4} + 2\frac{d^2}{dt^2} - 8\frac{d}{dt} + 5\right)\xi_{p,1}(t) = (16c_1 + 24c_2)e^t + 48c_2te^t.$$

Thus we have

$$16c_1 + 24c_2 = 0,\ 48c_2 = 1 \quad \Longrightarrow \quad c_1 = -\frac{1}{32},\ c_2 = \frac{1}{48}.$$

Thus

$$\xi_{\mathrm{p},1}(t) = -\frac{t^2\mathrm{e}^t}{32} + \frac{t^3\mathrm{e}^t}{48}.$$

Now we find $\xi_{\mathrm{p},2}$ corresponding to $f_2 = \cos(2t)$. Here the annihilator F_{f_2} of f_2 has characteristic polynomial $P_{F_{f_2}} = X^2 + 4$. We have

$$P_{F_{f_2}} P_{F_{\mathrm{h}}} = (X^2 + 4)(X^4 + 2X^2 - 8X + 5).$$

Thus the fundamental set of solutions for $F_{f_2} \circ F_{\mathrm{h}}$ is given by

$$\mathrm{e}^{-t}\cos(2t),\ \mathrm{e}^{-t}\sin(2t),\ \mathrm{e}^t,\ t\mathrm{e}^t,\ \cos(2t),\ \sin(2t).$$

Since the first four of these are solutions for F_{h}, we have

$$\xi_{\mathrm{p},2}(t) = c_1\cos(2t) + c_2\sin(2t).$$

To determine c_1 and c_2 we compute

$$\left(\frac{\mathrm{d}^4}{\mathrm{d}t^4} + 2\frac{\mathrm{d}^2}{\mathrm{d}t^2} - 8\frac{\mathrm{d}}{\mathrm{d}t} + 5\right)\xi_{\mathrm{p},2}(t) = (13c_1 - 16c_2)\cos(2t) + (16c_1 + 13c_2)sin(2t).$$

Therefore,

$$13c_1 - 16c_2 = 1,\ 16c_1 + 13c_2 = 0 \quad \Longrightarrow \quad c_1 = \frac{13}{425},\ c_2 = \frac{16}{425}.$$

Thus

$$\xi_{\mathrm{p},2} = \frac{13}{425}\cos(2t) + \frac{16}{425}\sin(2t).$$

Finally, we have the particular

$$\xi_{\mathrm{p}}(t) = -\frac{t^2\mathrm{e}^t}{32} + \frac{t^3\mathrm{e}^t}{48} + \frac{13}{425}\cos(2t) + \frac{16}{425}\sin(2t).$$

Thus, as per Theorem 4.3.3, *any* solution ξ of F can be written we

$$\xi(t) = c_1\mathrm{e}^t + c_2t\mathrm{e}^t + c_3\mathrm{e}^{-t}\cos(2t) + c_4\mathrm{e}^{-t}\sin(2t) - \frac{t^2\mathrm{e}^t}{32} + \frac{t^3\mathrm{e}^t}{48} + \frac{26}{425}\cos(2t) + \frac{32}{425}\sin(2t).$$

To determine the constants c_1, c_2, c_3, c_4, we use the initial conditions

$$\xi(0) = x_0,\ \frac{\mathrm{d}\xi}{\mathrm{d}t}(0) = x + 0^{(1)},\ \frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}(0) = x_0^{(2)},\ \frac{\mathrm{d}^3\xi}{\mathrm{d}t}(0) = x_0^{(3)}.$$

These do *not* have the same solution as in Example 4.2.20 because of the presence of the particular solution. Some unpleasant computation gives the equations

$$\begin{aligned} c_1 + c_3 &= -\frac{26}{425} + x_0, \\ c_1 + c_2 - c_3 + 2c_4 &= -\frac{64}{425} + x_0^{(1)}, \\ c_1 + 2c_2 - 3c_3 - 4c_4 &= \frac{2089}{6800} + x_0^{(2)}, \\ c_1 + 3c_2 + 11c_3 - 2c_4 &= \frac{4521}{6800} + x_0^{(3)} \end{aligned}$$

that have to be solved. Here's what you get:

$$\begin{aligned} c_1 &= \frac{15}{16}x_0 + \frac{1}{16}x_0^{(1)} + \frac{1}{16}x_0^{(2)} - \frac{1}{16}x_0^{(3)} - \frac{303}{3400}, \\ c_2 &= -\frac{5}{8}x_0 + \frac{3}{8}x_0^{(1)} + \frac{1}{8}x_0^{(2)} + \frac{1}{8}x_0^{(3)} + \frac{2809}{27200}, \\ c_3 &= \frac{1}{16}x_0 - \frac{1}{16}x_0^{(1)} - \frac{1}{16}x_0^{(2)} + \frac{1}{16}x_0^{(3)} + \frac{19}{680}, \\ c_4 &= -\frac{1}{8}x_0 + \frac{1}{4}x_0^{(1)} - \frac{1}{8}x_0^{(2)} - \frac{3721}{54400}. \end{aligned}$$

Alternatively, one can use MATHEMATICA® as illustrated in Section 4.9.1. You will then get back a reliable answer after about 15 seconds of typing. You can decide which method you think is best in practice. •

The next two examples give an illustration of where pretty uninteresting functions are interesting in application.

4.3.19 Example (First-order system with step input) The differential equation we consider here is an inhomogeneous version of the equation considered in Example 4.2.21. We take the first-order scalar linear inhomogeneous ordinary differential equation F with constant coefficients and with right-hand side

$$\widehat{F}(t, x) = -\frac{x}{\tau} + 1.$$

Thus solutions $t \mapsto \xi(t)$ to this differential equation satisfy

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + \frac{1}{\tau}\xi(t) = 1.$$

We have already determined that a solution to the homogeneous equation will have the form $\xi(t) = c\mathrm{e}^{-t/\tau}$, taking the convention that $\frac{1}{\tau} = 0$ when "$\tau = \infty$."

So next we find a particular solution. The annihilator F_f of the pretty uninteresting function $f(t) = 1$ has characteristic polynomial $P_{F_f} = X$. The characteristic

polynomial for F_{h} is $P_{F_{\mathrm{h}}} = X + \frac{1}{\tau}$. Thus we must list the fundamental solutions for G_f, where

$$P_{G_f} = X(X + \tfrac{1}{\tau}).$$

There are two cases.

First, when $\tau \neq \infty$, the fundamental solutions are $t \mapsto \mathrm{e}^{-t/\tau}$ and $t \mapsto 1$, using Procedure 4.2.18. The first of these is a solution for the homogeneous equation, so we take a particular solution to be a multiple of the second: $\xi_{\mathrm{p}}(t) = c$. To find c we substitute into the differential equation:

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau}\right)\xi_{\mathrm{p}} = \frac{c}{\tau}.$$

To be a particular solution, we must have $\frac{c}{\tau} = 1$ and so $c = \tau$. Thus $\xi_{\mathrm{p}}(t) = \tau$.

The other case arises when $\tau = \infty$, and in this case the fundamental solutions for G_f are $t \mapsto 1$ and $t \mapsto t$, again using Procedure 4.2.18. In this case, the first of these functions is a solution for the homogeneous system, and so a multiple of the second will be a particular solution, i.e., $\xi_{\mathrm{p}}(t) = ct$. To determine c we require that ξ_{p} be a particular solution:

$$\frac{\mathrm{d}}{\mathrm{d}t}\xi_{\mathrm{p}}(t) = c,$$

from which we deduce that $c = 1$. Thus $\xi_{\mathrm{p}}(t) = t$.

In summary, a particular solution is

$$\xi_{\mathrm{p}}(t) = \begin{cases} \tau, & \tau \neq \infty, \\ t, & t = \infty. \end{cases}$$

Therefore, *any* solution has the form

$$\xi(t) = c\mathrm{e}^{-t/\tau} + \xi_{\mathrm{p}}(t).$$

In case $\tau \neq \infty$, one normally takes the initial condition $\xi(0) = 0$ to get $c = -\tau$ and so

$$\xi(t) = \tau(1 - \mathrm{e}^{-t/\tau}).$$

To allow a fruitful comparison of the effects of changing τ, let us normalise this solution by multiplying by $\frac{1}{\tau}$ to get the ***step response***

$$\mathsf{1}_F(t) = 1 - \mathrm{e}^{-t/\tau}.$$

In Figure 4.4 we graph this step response for varying values of $\tau \in \mathbb{R}_{>0}$. We note that as τ gets smaller, the step response rises more quickly, i.e., responds faster. •

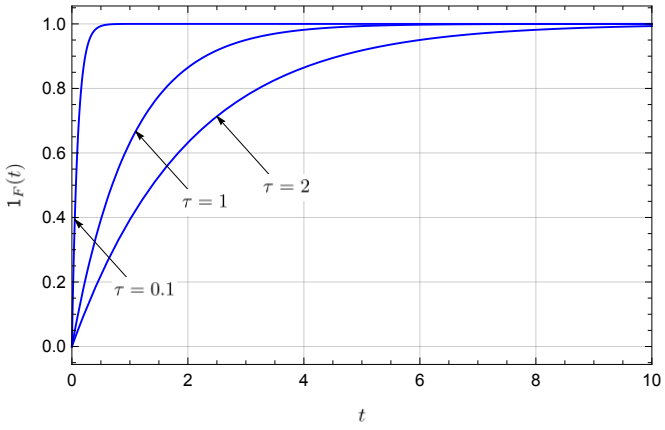


Figure 4.4 The step response of a first-order system

4.3.20 Example (Second-order system with sinusoidal input) Next we consider the second-order differential equation of Example 4.2.22, but with a sinusoidal inhomogeneous term. Thus we take the second-order scalar linear inhomogeneous ordinary differential equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}) = -\omega_0^2 x - 2\zeta\omega_0 x^{(1)} + A\sin(\omega t)$$

for $A, \omega \in \mathbb{R}_{>0}$. Solutions $t \mapsto \xi(t)$ then satisfy

$$\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}(t) + 2\zeta\omega_0\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + \omega_0^2\xi(t) = A\sin(\omega t).$$

In Example 4.2.22 we carefully and thoroughly investigated the nature of the solutions for the homogeneous system. There we saw, for example, that as long as $\zeta > 0$, solutions to the homogeneous equation decay to zero as $t \to \infty$. For $\zeta = 0$, solutions were periodic. Here we will thus focus on $\zeta \in \mathbb{R}_{\geq 0}$ and on the nature of the particular solution. When $\zeta \in \mathbb{R}_{>0}$, this means that we are looking at the "steady-state" behaviour of the system, i.e., what we see after a long time. When $\zeta = 0$, we do not have this steady-state interpretation, but nonetheless we will interpret these solutions in light of our understanding of what happens when $\zeta \in \mathbb{R}_{>0}$.

The annihilator F_f for the pretty uninteresting function $f(t) = A\sin(\omega t)$ has characteristic polynomial $P_{F_f} = X^2+\omega^2$. We have two cases to consider for particular solutions.

The first case is when $\zeta \in \mathbb{R}_{>0}$ or when $\zeta = 0$ and $\omega \neq \omega_0$. Here the characteristic polynomial for G_f in Procedure 4.3.16 is

$$P_{G_f} = (X^2 + \omega^2)(X^2 + 2\zeta\omega_0 X + \omega_0^2).$$

The fundamental solutions for G_f associated to this polynomial, according to Procedure 4.2.18, are

$$\xi_1(t),\ \xi_2(t),\ \cos(\omega t),\ \sin(\omega t),$$

where ξ_1 and ξ_2 are homogeneous solutions as determined in Example 4.2.22. Thus a particular solution will be of the form

$$\xi_{\mathrm{p}}(t) = c_1 \cos(\omega t) + c_2 \sin(\omega t).$$

To determine c_1 and c_2 we require that ξ_{p} be a particular solution. Thus we compute

$$\begin{aligned}\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2} + 2\zeta\omega_0\frac{\mathrm{d}}{\mathrm{d}t} + \omega_0^2\right)\xi_{\mathrm{p}}(t)&\\ = (c_1(\omega_0^2 - \omega^2) + c_2 2\zeta\omega_0\omega)\cos(\omega t) + (-c_2 2\zeta\omega_0\omega + c_2(\omega_0^2 - \omega^2))\sin(\omega t).&\end{aligned}$$

We must, therefore, have

$$\begin{aligned} c_1(\omega_0^2 - \omega^2) + c_2 2\zeta\omega_0\omega &= 0,\\ -c_2 2\zeta\omega_0\omega + c_2(\omega_0^2 - \omega^2) &= A,\end{aligned} \quad \Longrightarrow \quad \begin{aligned} c_1 &= \frac{2\zeta\omega_0\omega A}{\omega^4 + \omega_0^4 - 2\omega_0^2\omega^2(1 - 2\zeta^2)},\\ c_2 &= \frac{(\omega_0^2 - \omega^2)A}{\omega^4 + \omega_0^4 - 2\omega_0^2\omega^2(1 - 2\zeta^2)}\end{aligned}$$

Thus a particular solution is

$$\xi_{\mathrm{p}}(t) = \frac{2\zeta\omega_0\omega A}{\omega^4 + \omega_0^4 - 2\omega_0^2\omega^2(1 - 2\zeta^2)}\cos(\omega t) + \frac{(\omega_0^2 - \omega^2)A}{\omega^4 + \omega_0^4 - 2\omega_0^2\omega^2(1 - 2\zeta^2)}\sin(\omega t).$$

The other case is when $\zeta = 0$ and $\omega = \omega_0$. Here the characteristic polynomial for G_f in Procedure 4.3.16 is

$$P_{G_f} = (X^2 + \omega^2)^2$$

The fundamental solutions for G_f associated to this polynomial, according to Procedure 4.2.18, are

$$\xi_1(t),\ \xi_2(t),\ t\cos(\omega t),\ t\sin(\omega t),$$

where ξ_1 and ξ_2 are homogeneous solutions as determined in Example 4.2.22. Therefore, a particular solution will have the form

$$\xi_{\mathrm{p}}(t) = c_1 t\cos(\omega_0 t) + c_2 t\sin(\omega_0 t).$$

To determine c_1 and c_2 we ask that this be a particular solution. Thus we compute

$$\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2} + \omega_0^2\right)\xi_{\mathrm{p}}(t) = 2c_2\omega_0\cos(\omega_0 t) - 2c_1\omega_0\sin(\omega_0 t).$$

Therefore, we must have

$$2c_2\omega_0 = 0,\ 2c_1\omega_0 = A, \quad \Longrightarrow \quad c_1 = \frac{A}{2\omega_0},\ c_2 = 0,$$

and so the particular solution we obtain is

$$\xi_{\mathrm{p}}(t) = \frac{At}{2\omega_0}\cos(\omega_0 t).$$

Therefore, in summary, a particular solution is

$$\xi_{\mathrm{p}}(t) = \begin{cases} \frac{At}{2\omega_0}\cos(\omega_0 t), & \zeta = 0,\ \omega = \omega_0, \\ \frac{2\zeta\omega_0\omega A}{\omega^4+\omega_0^4-2\omega_0^2\omega^2(1-2\zeta^2)}\cos(\omega t) + \frac{(\omega_0^2-\omega^2)A}{\omega^4+\omega_0^4-2\omega_0^2\omega^2(1-2\zeta^2)}\sin(\omega t), & \text{otherwise.} \end{cases}$$

Any solution will be a sum of this solution, plus some solution to the homogeneous equation as determined in Example 4.2.22.

In Figure 4.5 we graph particular solutions for various ζ's and ω's, keeping A and ω_0 fixed. We make the following observations.

1. For small values of ω (compared to ω_0), the response $\xi_{\mathrm{p}}(t)$ is quite closely aligned in amplitude and phase with the input $f(t)$.
2. For small values of ζ, i.e., small damping, as $\omega \to \omega_0$ the response gets large in amplitude and the phase shift is about $\frac{1}{4}$ of a period.
3. For not so small values of ζ, the amplitude as $\omega \to \omega_0$ does not grow so much, but the phase still shifts by about $\frac{1}{4}$ of a period.
4. As the frequency ω gets large (compared to ω_0), the amplitude decays to zero, and the response and input are out of phase, i.e., the phase shift is about $\frac{1}{2}$ of a period.

One can see in the previous description the genesis of what happens when $\zeta = 0$, i.e., the response amplitude grows over time. This phenomenon is called "resonance," meaning that the excitation from the inhomogeneous term has the same frequency as the natural frequency of the system.

The matters touched above in the preceding discussion are captured in system theory by the notion of "frequency response." •

4.3.3 Notes

[Duffy 2015]

Exercises

4.3.1 Consider the ordinary differential equation F with right-hand side given by (4.11).

(a) Convert this to a first-order equation with k states, following Exercise 3.1.23.

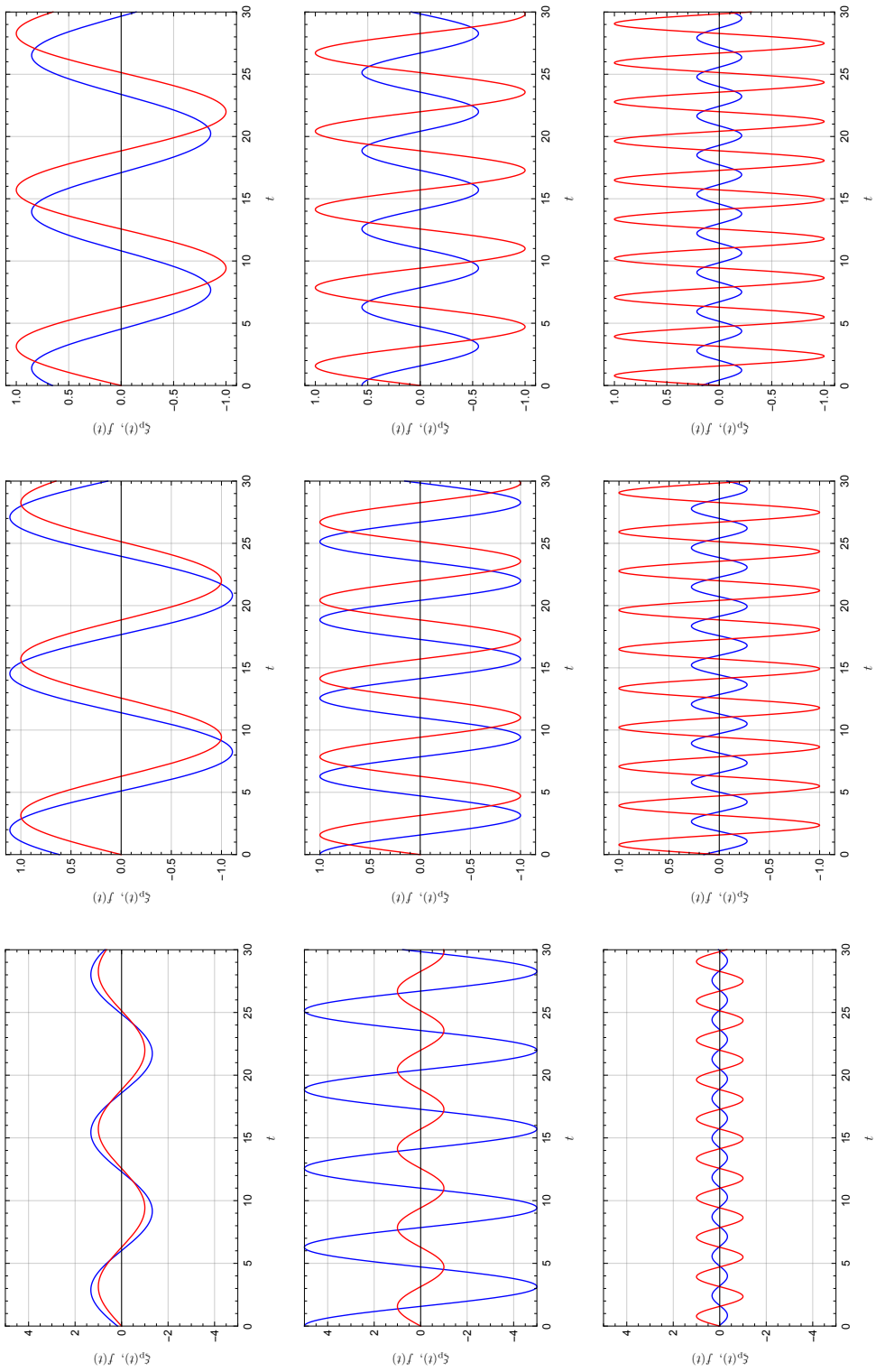


Figure 4.5 Response (in blue) of a second-order system with $\omega_0 = 1$ to a sinusoidal input with $A = 1$ (in red) for varying ζ and ω (left: $\zeta = 0.1$, $\omega \in \{0.5, 1, 2\}$; middle: $\zeta = 0.5$, $\omega \in \{0.5, 2, 1\}$; right: $\zeta = 0.9$, $\omega \in \{0.5, 1, 2\}$

(b) Show that, if $a_0, a_1, \dots, a_k, b \in \mathsf{L}^1_{\text{loc}}(\mathbb{T}; \mathbb{R})$, then the resulting first-order equation satisfies the conditions of Theorem 3.2.8 for existence of a unique solution $t \mapsto \xi(t)$ satisfying the initial conditions

$$\xi(t_0) = x_0, \xi(t_0 + h) = x_0^{(1)}, \dots, \xi(t_0 + (k-1)h) = x_0^{(k-1)}$$

at time $t_0 \in \mathbb{T}$.

4.3.2 Consider the ordinary differential equation F with right-hand side given by (4.11). Answer the following questions.

(a) Show that the particular particular solution

$$\xi_{\text{p},b}(t) = \int_{t_0}^{t} G_F(t, \tau) b(\tau) \, \mathrm{d}\tau$$

satisfies the initial value problem

$$\frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t) + a_{k-1}(t) \frac{\mathrm{d}^{k-1} \xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t) \frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t) \xi(t) = b(t),$$
$$\xi(t_0) = 0, \ \frac{\mathrm{d}\xi}{\mathrm{d}t}(t_0) = 0, \dots, \ \frac{\mathrm{d}^{k-1} \xi}{\mathrm{d}t^{k-1}}(t_0) = 0.$$

(b) Show that the solution to the initial value problem

$$\frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t) + a_{k-1}(t) \frac{\mathrm{d}^{k-1} \xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t) \frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t) \xi(t) = b(t),$$
$$\xi(t_0) = x_0, \ \frac{\mathrm{d}\xi}{\mathrm{d}t}(t_0) = x_0^{(1)}, \dots, \ \frac{\mathrm{d}^{k-1} \xi}{\mathrm{d}t^{k-1}}(t_0) = x_0^{(k-1)}$$

is given by $\xi(t) = \xi_{\text{h}} + \xi_{\text{p},b}$, where ξ_{h} is the solution to the homogeneous initial value problem

$$\frac{\mathrm{d}^k \xi_{\text{h}}}{\mathrm{d}t^k}(t) + a_{k-1}(t) \frac{\mathrm{d}^{k-1} \xi_{\text{h}}}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t) \frac{\mathrm{d}\xi_{\text{h}}}{\mathrm{d}t}(t) + a_0(t) \xi_{\text{h}}(t) = 0,$$
$$\xi_{\text{h}}(t_0) = x_0, \ \frac{\mathrm{d}\xi_{\text{h}}}{\mathrm{d}t}(t_0) = x_0^{(1)}, \dots, \ \frac{\mathrm{d}^{k-1} \xi_{\text{h}}}{\mathrm{d}t^{k-1}}(t_0) = x_0^{(k-1)}.$$

4.3.3 Find the annihilator for each of the following also pretty uninteresting functions f:

(a) $f(t) = 2t^2 + 3t - 5$;
(b) $f(t) = (t^2 + 2t + 1)\mathrm{e}^t$;
(c) $f(t) = t\mathrm{e}^{2t} \cos(t) + \mathrm{e}^{2t} \sin(t)$;
(d) $f(t) = t^3 \mathrm{e}^{-t} \sin(3t) + t^2 \mathrm{e}^{-t} \cos(3t)$.

4.3.4 For the following scalar linear inhomogeneous ordinary differential equations F, determine the general form of their solutions:

(a) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + 2x^{(1)} + x - 3\mathrm{e}^t$;
(b) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} - 5x^{(1)} + 6x - 2\mathrm{e}^{3t} - \cos(t)$;
(c) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} - 2x^{(1)} + 5x - t\mathrm{e}^t \sin(2t)$;
(d) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + 4x - t\cos(2t) + \sin(2t)$;
(e) $F(t, x, x^{(1)}, x^{(2)}, x^{(3)}) = x^{(3)} - x - t\mathrm{e}^t$;
(f) $F(t, x, x^{(1)}, \dots, x^{(4)}) = x^{(4)} + 4x^{(2)} + 4x - \cos(2t) - \sin(2t)$.

4.3.5 Solve the initial value problem for the following scalar linear inhomogeneous differential equations F with the stated initial conditions:
(a) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + 2x^{(1)} + x - 3\mathrm{e}^t$, and $\xi(0) = 1$, $\dot{\xi}(0) = 1$;
(b) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} - 5x^{(1)} + 6x - 2\mathrm{e}^{3t} - \cos(t)$, and $\xi(0) = 0$, $\dot{\xi}(0) = 1$;
(c) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} - 2x^{(1)} + 5x - t\mathrm{e}^t \sin(2t)$, and $\xi(0) = 1$, $\dot{\xi}(0) = 0$;
(d) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + 4x - t\cos(2t) + \sin(2t)$, and $\xi(0) = 2$, $\dot{\xi}(0) = 1$;
(e) $F(t, x^{(1)}, x^{(2)}, x^{(3)}) = x^{(3)} - x - t\mathrm{e}^t$, and $\xi(0) = 1$, $\dot{\xi}(0) = 1$, $\ddot{\xi}(0) = 1$;
(f) $F(t, x, x^{(1)}, \dots, x^{(4)}) = x^{(4)} + 4x^{(2)} + 4x - \cos(2t) - \sin(2t)$, and $\xi(0) = 0$, $\dot{\xi}(0) = 0$, $\ddot{\xi}(t) = 0$, $\dddot{\xi}(t) = 0$.

4.3.6 Suppose a mass m falls under the influence of gravity with gravitational acceleration a_g and suppose that the force due to air resistance is proportional to velocity, i.e., given by ρv, where v is the velocity.
(a) Use Newton's laws of force balance to write the equations governing the falling velocity of the mass.
(b) Obtain the solution to the differential equation from part (a), supposing the mass is at rest at $t = 0$.
(c) What is the terminal velocity of the mass?
(d) What are the units of m, a_g, and ρ, in terms of mass, length, and time units?
(e) Combine the physical constants m, a_g, and ρ in such a way that the units for the combined expression are "length/time," i.e., velocity. How does this constant compare to the terminal velocity you computed in part (c)?

4.3.7 Let $P \in \mathbb{R}[X]$ be given by

$$P = X^k + a_{k-1}X^{k-1} + \cdots + a_1 X + a_0,$$

and suppose that $r \in \mathbb{R}$ is not a root of P. Show that

$$\xi_{\mathrm{p}}(t) = \frac{\mathrm{e}^{rt}}{\widehat{P}(r)}$$

is a particular solution of the differential equation

$$F(t, x, x^{(1)}, \dots, x^{(k)}) = x^{(k)} + a_{k-1}x^{(k-1)} + \cdots + a_1 x^{(1)} + a_0 x - \mathrm{e}^{rt}.$$

4.3.8 For the following scalar linear homogeneous ordinary differential equations with time-domain $\mathbb{T} = [0, \infty)$, find their continuous-time Green's function:

(a) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + x^{(1)}$;
(b) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + \omega^2 x$, $\omega \in \mathbb{R}_{>0}$;
(c) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} - 2x^{(1)} + x$.

Section 4.4

Scalar linear inhomogeneous ordinary differential equations with distributions as right-hand side

In our development of continuous-time linear system theory, an important rôle will be played by ordinary differential equations involving distributions. In this section we shall undertake a systematic development of this theory for scalar equations. Our objectives will be twofold: (1) provide distributional characterisations of results already obtained; (2) provide new general results for differential equations with distributions as right-hand side.

Do I need to read this section? The results and techniques we employ in this section we be an important part of a full understanding of the theory of linear systems. •

4.4.1 Definitions and preliminary constructions

Since the differential equations we consider in this section do not fall precisely into any of the classes of equations thus far considered, let us define precisely the objects with which we are dealing in this section. We shall consider ordinary differential equations with time-domain $\mathbb{T} = \mathbb{R}$ for simplicity, since we are working with spaces of distributions, which we have considered to be defined on $\mathbb{R}$. This can be done without loss of generality since an ordinary differential equation with time-domain $\mathbb{T} \subset \mathbb{R}$ can be extended to one with time-domain $\mathbb{R}$ by taking the right-hand side to be zero for times outside $\mathbb{T}$.

Here is the class if differential equations we work with in this section.

4.4.1 Definition (Scalar linear ordinary differential equations with distribution forcing) A ***scalar linear ordinary differential equation with distribution forcing*** is a pair (F, β) where:

(i) F is a scalar linear homogeneous ordinary differential equation with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{R} \times \mathbb{R} \times \mathsf{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) &\to \mathbb{R}\\ (t, x, x^{(1)}, \dots, x^{(k-1)}) &\mapsto -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x,\end{aligned}$$

where $a_0, a_1, \dots, a_{k-1} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$;

(ii) $\beta \in \mathscr{D}'(\mathbb{R};\mathbb{R})$.

A ***solution*** for a scalar linear ordinary differential equation with distribution forcing (F, β) is a distribution $\theta \in \mathscr{D}'(\mathbb{R};\mathbb{R})$ such that, for each $j \in \{0, 1, \dots, k-1\}$, either

(iii) distribution $\theta^{(j)}$ has a well-defined value on $a_j\phi$ for $\phi \in \mathscr{D}(\mathbb{R};\mathbb{R})$, denoted by

$$\langle a_j\theta^{(j)};\phi\rangle \triangleq \langle \theta^{(j)}; a_j\phi\rangle,$$

or

(iv) $\theta^{(j)}$ is a regular distribution associated with some $f_{\theta^{(j)}} \in \mathsf{C}^0(\mathbb{R};\mathbb{R})$ and the regular distribution $\theta_{a_j\phi}$ has a well-defined value on $f_{\theta^{(j)}}$, denoted by

$$\langle a_j\theta^{(j)};\phi\rangle \triangleq \langle \theta_{a_j\phi}; f_{\theta^{(j)}}\rangle,$$

and the equation

$$\langle \theta^{(k)};\phi\rangle + \langle a_{k-1}\theta^{(k-1)};\phi\rangle + \cdots + \langle a_1\theta^{(1)};\phi\rangle + \langle a_0\theta;\phi\rangle = \langle \beta;\phi\rangle, \qquad \phi \in \mathscr{D}(\mathbb{R};\mathbb{R}),$$

is satisfied. •

The dichotomous meaning of the symbol $\langle a_j\theta^{(j)};\phi\rangle$ employed in the definition of a solution is convenient and gives precise meaning to the expression

$$\theta^{(k)} + a_{k-1}\theta^{(k-1)} + \cdots + a_1\theta^{(1)} + a_0\theta = \beta.$$

Let us consider a few cases where we might employ this notation.

1. If $a_j \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$ and if $\theta = \theta_\xi$ is a regular distribution for which $\xi \in \mathsf{AC}^{k-1}(\mathbb{R};\mathbb{R})$, then, for $j \in \{0,1,\ldots,k-1\}$, $\theta_\xi^{(j)}$ is a regular distribution associated with a continuous signal, and so

$$\langle a_j\theta^{(j)};\phi\rangle = \langle \theta_{a_j\phi}; \xi^{(j)}\rangle = \int_{\mathbb{R}} a_j(t)\xi^{(j)}(t)\phi(t)\,\mathrm{d}t$$

makes sense.

2. If $a_j \in \mathsf{C}^r(\mathbb{R};\mathbb{R})$ and if θ is a distribution of order $r+j$ for some $r \in \mathbb{Z}_{>0}$, then

$$\langle a_j\theta^{(j)};\phi\rangle = \langle \theta^{(j)}; a_j\phi\rangle$$

makes sense since $\theta^{(j)}$ has order r, cf. Proposition IV-3.2.49.

3. Note that both of the above situations might arise in the same equation.

Another construction of which we shall make use is that of the formal adjoint of an ordinary differential equation. We will define this for a differential equation with infinitely differentiable time-varying coefficients.

4.4.2 Definition (Formal adjoint of a scalar linear ordinary differential equation) Let F be an homogeneous scalar linear ordinary differential equation with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{R}\times\mathbb{R}\times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) &\to \mathbb{R}\\ (t,x,x^{(1)},\dots,x^{(k-1)}) &\mapsto -a_{k-1}(t)x^{(k-1)}-\cdots-a_1(t)x^{(1)}-a_0(t)x,\end{aligned}$$

where $a_0, a_1, \dots, a_{k-1} \in \mathrm{C}^\infty(\mathbb{R};\mathbb{R})$. The ***formal adjoint*** of F is the homogeneous scalar linear ordinary differential equation F^* with right-hand side

$$\begin{aligned}\widehat{F^*}\colon \mathbb{R}\times\mathbb{R}\times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) &\to \mathbb{R}\\ (t,x,x^{(1)},\dots,x^{(k-1)}) &\mapsto -a^*_{k-1}(t)x^{(k-1)}-\cdots-a^*_1(t)x^{(1)}-a^*_0(t)x,\end{aligned}$$

where $a^*_0, a^*_1, \dots, a^*_{k-1} \in \mathrm{C}^\infty(\mathbb{R};\mathbb{R})$ satisfy

$$a^*_j = \sum_{l=0}^{k-j}(-1)^{l+j-k+1}\binom{l+j}{l}a^{(l)}_{l+j}, \qquad j\in\{0,1,\dots,k-1\}. \qquad \bullet$$

We note that the restriction to the equations with infinitely differentiable coefficients is a significant specialisation of the general cases considered in Sections 4.2.1 and 4.3.1. The reason that we make this simplification in the cases that we do is that we wish to be able to give these equations a general distribution as an argument, and so need a time-domain suited to distributions and we need to be able to multiply the distribution by the coefficient functions, cf. Example IV-3.2.11–2. Indeed, by assuming that these coefficients are smooth, we can make use of the definition of a solution above with

$$\langle a_j\theta^{(j)};\phi\rangle = \langle \theta^{(j)}; a_j\phi\rangle,$$

and allowing an arbitrary distribution θ as a possible solution. This then makes the following mapping sensible for a scalar linear homogeneous ordinary differential equation with smooth time-varying coefficients $a_0, a_1, \dots, a_{k-1} \in \mathrm{C}^\infty(\mathbb{R};\mathbb{R})$:

$$\begin{aligned}L_F\colon \mathscr{D}'(\mathbb{R};\mathbb{R}) &\to \mathscr{D}'(\mathbb{R};\mathbb{R})\\ \theta &\mapsto \theta^{(k)}+a_{k-1}\theta^{(k-1)}+\cdots+a_1\theta^{(1)}+a_0\theta.\end{aligned}$$

Making use of this notation, the principal value of the formal adjoint, for our purposes, is the following result.

4.4.3 Lemma (Characterisation of the formal adjoint) *Let* F *be an homogeneous scalar linear ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{\mathrm{F}}\colon \mathbb{R}\times\mathbb{R}\times \mathrm{L}^{\leq \mathrm{k}-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) &\to \mathbb{R}\\ (\mathrm{t},\mathrm{x},\mathrm{x}^{(1)},\dots,\mathrm{x}^{(\mathrm{k}-1)}) &\mapsto -\mathrm{a}_{\mathrm{k}-1}(\mathrm{t})\mathrm{x}^{(\mathrm{k}-1)}-\cdots-\mathrm{a}_1(\mathrm{t})\mathrm{x}^{(1)}-\mathrm{a}_0(\mathrm{t})\mathrm{x},\end{aligned}$$

where $\mathrm{a}_0, \mathrm{a}_1, \dots, \mathrm{a}_{\mathrm{k}-1} \in \mathrm{C}^\infty(\mathbb{R};\mathbb{R})$. *Then the formal adjoint* F^* *satisfies*

$$\langle \mathrm{L}_{\mathrm{F}}(\theta);\phi\rangle = (-1)^{\mathrm{k}}\langle \theta; \mathrm{L}_{\mathrm{F}^*}(\phi)\rangle, \qquad \theta\in\mathscr{D}'(\mathbb{R};\mathbb{R}),\ \phi\in\mathscr{D}(\mathbb{R};\mathbb{R}).$$

Proof We have, using the higher-order Leibniz rule (Proposition I-3.2.11),

$$\begin{aligned}\langle L_F(\theta);\phi\rangle &= \sum_{j=0}^{k}\langle\theta^{(j)};a_j\phi\rangle = \sum_{j=0}^{k}(-1)^j\langle\theta;(a_j\phi)^{(j)}\rangle\\ &= \sum_{j=0}^{k}\sum_{r=0}^{j}(-1)^j\binom{j}{r}\langle\theta;a_j^{(j-r)}\phi^{(r)}\rangle = \sum_{r=0}^{k}\sum_{j=r}^{k}(-1)^j\binom{j}{r}\langle\theta;a_j^{(j-r)}\phi^{(r)}\rangle\\ &= \sum_{r=0}^{k}\sum_{j=0}^{k-r}(-1)^{j+r}\binom{j+r}{r}\langle\theta;a_{j+r}^{(j)}\phi^{(r)}\rangle.\end{aligned}$$

(We use the convention that $a_k(t) = 1$, $t \in \mathbb{R}$.) Noting that the term involving $r = k$ is

$$(-1)^k\langle\theta;\phi^{(k)}\rangle,$$

this can be written as

$$\langle L_F(\theta);\phi\rangle = (-1)^k\left\langle\theta;\phi^{(k)} + \sum_{r=0}^{k-1}\sum_{j=0}^{k-r}(-1)^{j+r-k}\binom{j+r}{j}a_{j+r}^{(j)}\phi^{(r)}\right\rangle,$$

which gives the desired result. ∎

4.4.2 Equations with time-varying coefficients

In this section we consider a few results that are valid for scalar linear ordinary differential equations with time-varying coefficients. There is no completely general theory here, in part because the notion of a solution for such equations involves some sort of compatibility between the regularity of the time-varying coefficients and the regularity of the solution, as evidenced by the definition of solution in Definition 4.4.1. Thus we shall consider a few partial results. First we consider the case of infinitely differentiable time-varying coefficients, since in this case there will be no restrictions on distributional solutions, cf. the comments following Definition 4.4.2. Next we work in the general setting of equations with locally integrable time-varying coefficients, and give a distributional interpretation of the continuous-time Green's function of Definition 4.3.8.

4.4.2.1 Solutions and their properties In this section we consider scalar linear homogeneous ordinary differential equations F with right-hand side

$$\begin{aligned}\widehat{F}\colon\ &\mathbb{R}\times\mathbb{R}\times\mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R})\to\mathbb{R}\\ &(t,x,x^{(1)},\dots,x^{(k-1)})\mapsto -a_{k-1}(t)x^{(k-1)}-\cdots-a_1(t)x^{(1)}-a_0x,\end{aligned}\tag{4.21}$$

where $a_0, a_1, \dots, a_{k-1} \in \mathsf{C}^\infty(\mathbb{R};\mathbb{R})$. We then have the corresponding mappings

$$L_F\colon \mathscr{D}(\mathbb{R};\mathbb{R})\to\mathscr{D}(\mathbb{R};\mathbb{R})$$

and

$$L_F \colon \mathscr{D}'(\mathbb{R};\mathbb{R}) \to \mathscr{D}'(\mathbb{R};\mathbb{R})$$

defined by

$$L_F(\phi)(t) = \frac{\mathrm{d}^k\phi}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\phi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\phi}{\mathrm{d}t}(t) + a_0(t)\phi(t), \qquad \phi \in \mathscr{D}(\mathbb{R};\mathbb{R}),$$

and

$$L_F(\theta) = \theta^{(k)} + a_k\theta^{(k-1)} + \cdots + a_1\theta^{(1)} + a_0\theta, \qquad \theta \in \mathscr{D}'(\mathbb{R};\mathbb{R}).$$

For $\beta \in \mathscr{D}'(\mathbb{R};\mathbb{R})$, we denote

$$\mathrm{Sol}(F,\beta) = \{\theta \in \mathscr{D}'(\mathbb{R};\mathbb{R}) \mid L_F(\theta) = \beta\}.$$

We have the following analogue of Theorem 4.3.3.

4.4.4 Theorem (Affine space structure for sets of solutions) *Let* F *be a scalar linear homogeneous ordinary differential equation with right-hand side (4.21) for* $\mathrm{a}_0, \mathrm{a}_1, \ldots, \mathrm{a}_{k-1} \in \mathsf{C}^\infty(\mathbb{R};\mathbb{R})$. *If, for* $\beta \in \mathscr{D}'(\mathbb{R};\mathbb{R})$, *the set* $\mathrm{Sol}(\mathrm{F},\beta)$ *is nonempty, then it is an affine subspace of* $\mathscr{D}'(\mathbb{R};\mathbb{R})$.

Proof This is a direct consequence of Proposition I-5.4.48. ∎

There is an important point of difference here with Theorem 4.3.3, namely that we do not assert that $\mathrm{Sol}(F,\beta)$ is nonempty. To this end, we next establish some situations when $\mathrm{Sol}(F,\beta)$ is indeed nonempty. In doing so, we make use of the formal adjoint L_F^* from Definition 4.4.2.

4.4.2.2 A distributional interpretation of the continuous-time Green's function An interesting connection can be made between the continuous-time Green's function of Theorem 4.3.7 and the inhomogeneous equation with an appropriate delta-function as right-hand side. The following theorem gives the desired result.

4.4.5 Theorem (The continuous-time Green's function as a solution to a distributional differential equation) *Let* F *be a scalar linear homogeneous ordinary differential equation with* $\mathbb{T} = \mathbb{R}$, *right-hand side (4.1), and suppose that* $\mathrm{a}_0, \mathrm{a}_1, \ldots, \mathrm{a}_{k-1} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$. *For* $\mathrm{s} \in \mathbb{T}$, *let* θ_s *be the regular distribution corresponding to the locally integrable function* $\mathrm{t} \mapsto \mathrm{G_F}(\mathrm{t},\mathrm{s})$. *Then* θ_s *is a solution to* $(\mathrm{F}, \tau_\mathrm{s}^*\delta)$.

Proof As in the proof of Theorem 4.3.7, let $\xi_s \colon \mathbb{T} \to \mathbb{R}$ be the solution to the initial value problem

$$\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t) + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t)\xi(t) = 0,$$
$$\xi(s) = \cdots = \frac{\mathrm{d}^{k-2}\xi}{\mathrm{d}t^{k-2}}(s) = 0, \ \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(s) = 1,$$

so that θ_s is the distribution associated with the locally integrable function $\xi_s\tau_s^*\mathbf{1}_{\geq 0}$. Using a simple induction, we then have, for $l \in \{0, 1, \dots, k\}$,

$$(\xi_s\tau_s^*\mathbf{1}_{\geq 0})^{(l)} = \sum_{j=0}^{l-1} \xi_s^{(j)}(s)(\tau_s^*\delta)^{(l-j-1)} + \xi_s^{(l)}\tau_s^*\mathbf{1}_{\geq 0}.$$

(All products of distributions with functions are to be interpreted in the sense of Proposition IV-3.2.49 and Corollary IV-3.7.28.) By using the initial conditions for ξ_s, we have

$$(\xi_s\tau_s^*\mathbf{1}_{\geq 0})^{(k)} = \xi_s^{(k)}\tau_s^*\mathbf{1}_{\geq 0} + \tau_s^*\delta, \quad (\xi_s\tau_s^*\mathbf{1}_{\geq 0})^{(j)} = \xi_s^{(j)}\tau_s^*\mathbf{1}_{\geq 0}, \qquad j \in \{0, 1, \dots, k-1\}.$$

Referring to the discussion following Definition 4.4.1, we have, for $\phi \in \mathscr{D}(\mathbb{R}; \mathbb{R})$,

$$\langle a_j\theta_s^{(j)}; \phi\rangle = \int_{\mathbb{R}} a_j(t)\mathbf{1}_{\geq 0}(t-s)\xi_s(t)\phi(r)\,\mathrm{d}t, \qquad j \in \{0, 1, \dots, k-1\},$$

and

$$\langle a_k\theta_s^{(k)}; \phi\rangle = \langle \theta^{(1)}_{\xi_s(\tau_s^*\mathbf{1}_{\geq 0})^{(k-1)}}; \phi\rangle,$$

keeping in mind that $a_k = 1$. Thus

$$\theta_s^{(k)} + a_{k-1}\theta_s^{(k-1)} + \cdots + a_1\theta_s^{(1)} + a_0\theta_s$$
$$= (\xi_s^{(k)} + a_{k-1}\xi_s^{(k-1)} + \cdots + a_1\xi_s^{(1)} + a_0\xi_s)\tau_s^*\mathbf{1}_{\geq 0} + \tau_s^*\delta = \tau_s^*\delta,$$

as claimed. ■

The theorem makes precise our informal discussion of "pulses" applied at initial times in Remark 4.3.9–2. Let us undertake an informal discussion of this so as to get some insight into how one might thing about this. The differential equation we are considering is

$$\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k} + a_{k-1}(t)\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}} + \cdots + \frac{\mathrm{d}\xi}{\mathrm{d}t} + a_0(t)\xi = \delta(t-s),$$

and as initial conditions we take

$$\lim_{t\uparrow s} \frac{\mathrm{d}^j\xi}{\mathrm{d}t^j}(t) = 0, \qquad j \in \{0, 1, \dots, k-1\}.$$

We then apply Theorem 4.3.7(v) in a direct (and not entirely precise) way. Thus the solution to the initial value problem is

$$\xi(t) = \int_s^t G_F(t, \tau)\delta(t-s)\,\mathrm{d}\tau = G_F(t, s). \tag{4.22}$$

Note that there are some ways in which this formula does not quite jive with our intuition about solutions to initial value problems. First of all, $\frac{\mathrm{d}^{k-1}G_F(t,s)}{\mathrm{d}t^{k-1}}(s)$ has two different values, 0 for the limit from the left and 1 for the limit from the right. This inconsistency is a result of the fact that the delta-function is not an. . . er. . . function. Thus the integral in (4.22) does not have the usual continuity properties with respect to the limits of integration. However, if one closes ones eyes to this, then this gives a nice interpretation of the continuous-time Green's function.

4.4.3 Equations with constant coefficients

Next we turn to the consideration of scalar linear ordinary differential equations with distribution forcing (F, β), where F is a constant coefficient equation. In this case there is a great deal more that one can say about solutions of these equations, especially if one restricts β and solutions to lie in particular subspaces of distributions. Of particular interest are the subspaces $\mathscr{D}'_+(\mathbb{R}; \mathbb{F})$ and $\mathscr{D}'_-(\mathbb{R}; \mathbb{R})$ of causal and acausal distributions (see Definition IV-3.2.17). We shall see that, for constant coefficient equations, an important rôle is played by convolution.

4.4.3.1 Solutions and their properties We state the basic existence and uniqueness results for constant coefficient equations. The character of the results is similar to, but not identical to, results for non-distributional equations. Essential differences arise as a result of the fact that one cannot prescribe initial conditions for distributions as one does to obtain uniqueness of solutions in the usual case.

The following simple result illustrates one of the special features of constant coefficient equations. In stating the result, we recall from that distributions with compact support can be convolved with arbitrary distributions.

what

4.4.6 Lemma (Convolution and scalar linear ordinary differential equations with constant coefficients) *Let* F *be a scalar linear homogeneous ordinary differential equation with constant coefficients. If* $\theta \in \mathscr{D}'(\mathbb{R}; \mathbb{R})$*, then*

$$\mathrm{L}_{\mathrm{F}}(\theta) = \mathrm{L}_{\mathrm{F}}(\delta) * \theta.$$

Proof We suppose that

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x.$$

We then compute

$$\begin{aligned} L_F(\theta) = L_F(\delta * \theta) &= (\delta * \theta)^{(k)} + a_{k-1}(\delta * \theta)^{(k-1)} + \cdots + a_1(\delta * \theta)^{(1)} + a_0(\delta * \theta) \\ &= (\delta^{(k)} + a_{k-1}\delta^{(k-1)} + \cdots + a_1\delta^{(1)} + a_0\delta) * \theta, \end{aligned}$$

convolution with delta

derivative of convolution

using and . ∎

With the lemma at hand, we can easily prove the basic existence and uniqueness theorem for (F, β), where F has constant coefficients.

4.4.7 Theorem (Existence and uniqueness of solutions for constant coefficient scalar linear inhomogeneous ordinary differential equations with distribution forcing) *Let* F *be a scalar linear homogeneous ordinary differential equation with constant coefficients. Then the following statements hold:*

(i) *if* $\beta \in \mathscr{D}'(\mathbb{R}; \mathbb{R})$*, then* $\mathrm{card}(\mathrm{Sol}(\mathrm{F}, \beta)) \geq 2$*;*

(ii) *if* $\beta \in \mathscr{D}'_+(\mathbb{R}; \mathbb{R})$*, then* $\mathrm{card}(\mathrm{Sol}(\mathrm{F}, \beta) \cap \mathscr{D}'_+(\mathbb{R}; \mathbb{R})) = 1$*;*

(iii) *if* $\beta \in \mathscr{D}'_-(\mathbb{R}; \mathbb{R}))$*, then* $\mathrm{card}(\mathrm{Sol}(\mathrm{F}, \beta) \cap \mathscr{D}'_-(\mathbb{R}; \mathbb{R})) = 1$*.*

Proof Let us first establish the existence of two distribution solutions θ to the equation $L_F(\theta) = \delta$. We suppose that

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x.$$

Let $\xi_0 \in \mathsf{C}^\infty(\mathbb{R}; \mathbb{R})$ be the solution to the initial value problem

$$\frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1 \frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0 \xi(t) = 0,$$
$$\xi(0) = \cdots = \frac{\mathrm{d}^{k-2}\xi}{\mathrm{d}t^{k-1}}(0) = 0, \ \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(0) = 1,$$

and denote $\xi_+ = \mathsf{1}_{\geq 0}\xi_0$ and $\xi_- = -\sigma^*\mathsf{1}_{\geq 0}\xi_0$. Note that $\theta_{\xi_+} \in \mathscr{D}'_+(\mathbb{R}; \mathbb{R})$ and $\theta_{\xi_-} \in \mathscr{D}'_-(\mathbb{R}; \mathbb{R})$.

From Theorem 4.4.5 we have $L_F(\theta_{\xi_+}) = \delta$. We claim that $L_F(\theta_{\xi_-}) = \delta$. Since $(\sigma^*\mathsf{1}_{\geq 0})^{(1)} = -\delta$, we have

$$(\xi_0 \sigma^* \mathsf{1}_{\geq 0})^{(l)} = -\sum_{j=0}^{l-1} \xi_0^{(j)}(0)(\sigma^*\delta)^{(l-j-1)} + \xi_0^{(l)}\sigma^*\mathsf{1}_{\geq 0},$$

similarly to what we saw in the proof of Theorem 4.4.5. Thus

$$(\xi_0\sigma^*\mathsf{1}_{\geq 0})^{(k)} = \xi_0^{(k)}\sigma^*\mathsf{1}_{\geq 0} - \sigma^*\delta, \quad (\xi_0\sigma^*\mathsf{1}_{\geq 0})^{(j)} = \xi_0^{(j)}\sigma^*\mathsf{1}_{\geq 0}, \qquad j \in \{0, 1, \dots, k-1\}.$$

Thus we have

$$\theta_{\xi_-}^{(k)} + a_{k-1}\theta_{\xi_-}^{(k-1)} + \cdots + a_1\theta_{\xi_-}^{(1)} + a_0\theta_{\xi_-}$$
$$= (-\xi_0^{(k)} - a_{k-1}\xi_0^{(k-1)} - \cdots - a_1\xi^{(1)} - a_0\xi)\sigma^*\mathsf{1}_{\geq 0} + \sigma^*\delta = +\delta,$$

noting that $\sigma^*\delta = \delta$. Thus $L_F(\theta_{\xi_-}) = \delta$, as claimed.

Note that this immediately gives

$$L_F(\delta) * \theta_{\xi_+} = L_F(\delta * \theta_{\xi_+}) = L_F(\theta_{\xi_+}) = \delta.$$

Similarly,

$$L_F(\delta) * \theta_{\xi_-} = \delta,$$

showing that both θ_{ξ_+} and θ_{ξ_-} are multiplicative inverses of $L_F(\delta)$ in the ring $\mathscr{D}'(\mathbb{R}; \mathbb{R})$ with the convolution product.

Now we proceed with the proof of the theorem, using the notation just introduced.

(i) We have, using Lemma 4.4.6 and the computations above,

$$L_F(\theta_{\xi_+} * \beta) = L_F(\delta) * (\theta_{\xi_+} * \beta) = (L_F(\delta) * \theta_{\xi_+}) * \beta = \delta * \beta = \beta,$$

and so $\theta_{\xi_+} * \beta \in \mathrm{Sol}(F, \beta)$. We similarly have $\theta_{\xi_-} * \beta \in \mathrm{Sol}(F, \beta)$.

(ii) Suppose that $\theta_1, \theta_2 \in \mathrm{Sol}(F, \beta) \cap \mathscr{D}'_+(\mathbb{R}; \mathbb{R})$. Then

$$\begin{aligned} & L_F(\theta_1) = \beta,\ L_F(\theta_2) = \beta \\ \implies & L_F(\delta) * \theta_1 = \beta,\ L_F(\delta) * \theta_2 = \beta \\ \implies & \theta_1 = \theta_{\xi_+} * \beta,\ \theta_2 = \theta_{\xi_+} * \beta \\ \implies & \theta_1 = \theta_2. \end{aligned}$$

Here we use the fact that θ_{ξ_+} is the *unique* inverse of $L_F(\delta)$ in $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$, by . what

(iii) This follows in the same manner as the previous part of the theorem. ■

In summary, distributional equations for constant coefficient equations always have solutions, and if we look for solutions in $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$ (resp. $\mathscr{D}'_-(\mathbb{R}; \mathbb{R})$) for equations where the forcing is in $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$ (resp. $\mathscr{D}'_-(\mathbb{R}; \mathbb{R})$), then solutions are unique. Moreover, the proof of the theorem furnishes formulae for the unique solutions in $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$ and $\mathscr{D}'_-(\mathbb{R}; \mathbb{R})$. Let us present this, outside the stodgy confines of a proof, in the case of $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$.

4.4.8 Corollary (Solutions to distributional differential equations in $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$) *For a scalar linear homogeneous ordinary differential equation* F *with constant coefficients and with right-hand side*

$$\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}, \mathrm{x}^{(1)}, \dots, \mathrm{x}^{(\mathrm{k}-1)}) = -\mathrm{a}_{\mathrm{k}-1}\mathrm{x}^{(\mathrm{k}-1)} - \cdots - \mathrm{a}_1\mathrm{x}^{(1)} - \mathrm{a}_0\mathrm{x},$$

let $\xi_+(\mathrm{t}) = \mathrm{G}_{\mathrm{F}}(\mathrm{t}, 0)$*, noting that, for* $\mathrm{t} \geq 0$*,* ξ_+ *is the solution to the initial value problem*

$$\frac{\mathrm{d}^{\mathrm{k}}\xi}{\mathrm{dt}^{\mathrm{k}}}(\mathrm{t}) + \mathrm{a}_{\mathrm{k}-1}\frac{\mathrm{d}^{\mathrm{k}-1}\xi}{\mathrm{dt}^{\mathrm{k}-1}}(\mathrm{t}) + \cdots + \mathrm{a}_1\frac{\mathrm{d}\xi}{\mathrm{dt}}(\mathrm{t}) + \mathrm{a}_0\xi(\mathrm{t}) = 0,$$
$$\xi(0) = \cdots = \frac{\mathrm{d}^{\mathrm{k}-2}\xi}{\mathrm{dt}^{\mathrm{k}-1}}(0) = 0,\ \frac{\mathrm{d}^{\mathrm{k}-1}\xi}{\mathrm{dt}^{\mathrm{k}-1}}(0) = 1,$$

while $\xi_+(\mathrm{t}) = 0$ *for* $\mathrm{t} < 0$*. Then the unique solution in* $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$ *to* (F, β) *for* $\beta \in \mathscr{D}'_+(\mathbb{R}; \mathbb{R})$ *is* $\theta_{\xi_+} * \beta$*.*

Note that the uniqueness of solutions in $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$ and $\mathscr{D}'_-(\mathbb{R}; \mathbb{R})$ are in contrast to the situation in Proposition 4.3.2 where, to achieve uniqueness, one needs to also prescribe initial conditions. One might then wonder whether the rôle of initial conditions can be mimicked for distributional differential equations. This is possible, and is presented in Proposition 4.4.11 below.

4.4.3.2 Distributional solutions of equations non-distributional equations

In this section we further connect solutions to distributional equations to their non-distributional counterparts by constructing the distributional solution to a non-distributional equation, including initial conditions.

Let us get started by noting that, if F is a scalar linear homogeneous ordinary differential equation with constant coefficients and if $b \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathbb{R})$ satisfies

$\inf\operatorname{supp}(b) > -\infty$, then there is a unique solution $\theta \in \mathscr{D}'_+(\mathbb{R};\mathbb{R})$ to (F, θ_b). This is a consequence of Theorem 4.4.7(ii). One might then wonder whether there are other distributional solutions, not in $\mathscr{D}'_+(\mathbb{R};\mathbb{R})$, to (F, θ_b). The following result indicates the constraints on other such solutions.

4.4.9 Proposition (Uniqueness of distributional solutions to non-distributional equations) *Let* F *be a scalar linear homogeneous ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{F}\colon \mathbb{R}\times\mathbb{R}\times \mathsf{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) &\to \mathbb{R}\\ (t, x, x^{(1)}, \dots, x^{(k-1)}) &\mapsto -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x,\end{aligned}$$

and let $b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$ *satisfy* $\inf\operatorname{supp}(b) > -\infty$. *Then, if* $\theta \in \operatorname{Sol}(F, \theta_b)$, *we have*

$$\langle \theta; \phi\rangle = \langle \theta_\xi; \phi\rangle, \qquad \phi \in \mathscr{D}((\inf\operatorname{supp}(b), \infty);\mathbb{R}),$$

where ξ *satisfies*

$$\frac{d^k\xi}{dt^k}(t) + a_{k-1}\frac{d^{k-1}\xi}{dt^{k-1}}(t) + \cdots + a_1\frac{d\xi}{dt}(t) + a_0\xi(t) = b(t), \qquad t \in (\inf\operatorname{supp}(b), \infty).$$

In particular,

$$\operatorname{Sol}(F, 0) = \{\theta_\xi \mid \xi \in \operatorname{Sol}(F)\},$$

i.e., solutions to the homogeneous equation in $\mathscr{D}'(\mathbb{R};\mathbb{R})$ *are exactly the regular distributions associated to the usual solutions of the homogeneous equation.*

Proof Let us first extend the notion of the order of a distribution given in Definition IV-3.2.47. Let $\mathbb{T} \subseteq \mathbb{R}$ be a compact interval. The $\mathbb{T}$-order of a distribution $\theta \in \mathscr{D}'(\mathbb{R};\mathbb{R})$ is the smallest $r \in \mathbb{Z}_{\ge 0}$ for which there exists $f \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$ such that

$$\langle \theta; \phi\rangle = \langle \theta_f^{(r+1)}; \phi\rangle, \qquad \phi \in \mathscr{D}(\mathbb{T};\mathbb{R}).$$

Thus it makes sense to say that a distribution of $\mathbb{T}$-order -1 is one that agrees with a regular distribution θ_f on $\mathbb{T}$ where $f \notin \mathsf{AC}_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$. Similarly, we shall say that θ has $\mathbb{T}$-order -2 if it agrees on $\mathbb{T}$ with a regular distribution θ_f for a signal $f \in \mathsf{AC}_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$, but $f \notin \mathsf{AC}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$. More generally, we shall see that θ has $\mathbb{T}$-order $-r$ if θ agrees on $\mathbb{T}$ with a regular distribution θ_f for a signal $f \in \mathsf{AC}^{r-2}_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$, but $f \notin \mathsf{AC}^{r-1}_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$. A distribution of $\mathbb{T}$-order $-\infty$ is then one that agrees on $\mathbb{T}$ with the regular distribution associated with an infinitely differentiable function.

Now suppose that $\operatorname{supp}(\beta) \subseteq [t_0, \infty)$ and let $\mathbb{T} \subseteq (t_0, \infty)$ be compact. If $\theta \in \mathscr{D}'(\mathbb{R};\mathbb{R})$ is in $\operatorname{Sol}(F, \beta)$, then we have

$$\theta^{(k)} = \theta_b - a_{k-1}\theta^{(k-1)} - \cdots - a_1\theta^{(1)} - a_0\theta. \tag{4.23}$$

Suppose that g has $\mathbb{T}$-order q and let r be the $\mathbb{T}$-order of $\theta^{(k)}$. We claim that $r = q$. Indeed, suppose that $r > q$. The orders of $\theta, \theta^{(1)}, \dots, \theta^{(k-1)}$ are then less than or equal

to $r-1$. This means that the order of the right-hand side of (4.23) is less than or equal to $r-1$, whereas the order of the left-hand side is r. This is a contradiction of the assumption that $r > q$. Similar, the assumption that $r < q$ leads to a contradiction. Therefore, $r = q$, meaning that the $\mathbb{T}$-order of $\theta^{(k)}$ is the same as that of θ_b. Thus $\theta^{(k)}$, and hence θ, must agree with a regular distribution on (t_0, ∞). Thus $\theta = \theta_\xi$ where ξ is a solution of the inhomogeneous equation with right-hand side b.

The final assertion of the proposition is a mere specialisation of the first part of the result, noting that $\operatorname{supp}(0) = \varnothing$. ■

An important consequence of the preceding result is the following complete characterisation of $\operatorname{Sol}(F, \beta)$ for $\beta \in \mathscr{D}'_+(\mathbb{R}; \mathbb{R})$. Of course, a similar result holds for $\mathscr{D}'_-(\mathbb{R}; \mathbb{R})$.

4.4.10 Corollary (Characterisation of $\operatorname{Sol}(F, \beta)$ for $\beta \in \mathscr{D}'_+(\mathbb{R}; \mathbb{R})$) *Let* F *be a scalar linear homogeneous ordinary differential equation with constant coefficients, let* $\beta \in \mathscr{D}'_+(\mathbb{R}; \mathbb{R})$, *and let* θ_0 *be the unique solution to* (F, β) *in* $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$, *as in Theorem 4.4.7(ii). Then*

$$\operatorname{Sol}(F, \beta) = \{\theta_0 + \theta_\xi \mid \xi \in \operatorname{Sol}(F)\}.$$

Proof If $\theta \in \operatorname{Sol}(F, \beta)$ then $L_F(\theta - \theta_0) = 0$. By Proposition 4.4.11, this means that $\theta - \theta_0$ is a regular distribution associated to a solution of the homogeneous equation F. ■

Now let us see how we can resolve the seeming paradox of the uniqueness of solutions asserted in Proposition 4.4.9 with the non-uniqueness arising from Proposition 4.3.2 (due to dependence on initial conditions). We do this by conjuring a distribution as right-hand side that incorporates the initial conditions.

4.4.11 Proposition (Distributional solutions of non-distributional equations with initial conditions) *For a scalar linear homogeneous ordinary differential equation* F *with constant coefficients and with right-hand side*

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x,$$

for $b \in L^1_{\mathrm{loc}}(\mathbb{R}; \mathbb{R})$, *and for* $t_0 \in \mathbb{R}$, *the following statements are equivalent for* $\xi \colon \mathbb{R} \to \mathbb{R}$:

(i) $\xi = \tau^*_{t_0} 1_{\geq 0} \xi_{t_0}$, *where* ξ_{t_0} *satisfies the initial value problem*

$$\frac{d^k \xi}{dt^k}(t) + a_{k-1}\frac{d^{k-1}\xi}{dt^{k-1}}(t) + \cdots + a_1 \frac{d\xi}{dt}(t) + a_0 \xi(t) = b(t),$$
$$\xi(t_0) = x_0, \ \frac{d\xi}{dt}(t_0) = x_1, \dots, \frac{d^{k-1}\xi}{dt^{k-1}}(t_0) = x_{k-1};$$

(ii) the distribution θ_ξ *is the unique solution in* $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$ *to* (F, β), *where*

$$\beta = \theta_{b\tau^*_{t_0} 1_{\geq 0}} + \sum_{j=0}^{k-1} \sum_{l=0}^{k-j-1} a_{j+l+1} x_l \tau^*_{t_0} \delta^{(j)}$$

Moreover, θ_ξ is the unique solution in $\mathscr{D}'(\mathbb{R};\mathbb{R})$ to (F,β).

Proof Suppose that ξ is as in part (i). As in the proof of Theorem 4.4.5, we have

$$(\xi_{t_0}\tau^*_{t_0}\mathbf{1}_{\geq 0})^{(l)} = \sum_{j=0}^{l-1} \xi^{(j)}_{t_0}(t_0)(\tau^*_{t_0}\delta)^{(l-j-1)} + \xi^{(l)}_{t_0}\tau^*_{t_0}\mathbf{1}_{\geq 0}.$$

Therefore,

$$\begin{aligned}
L_F(\theta_\xi) &= \sum_{l=0}^{k} a_l(\xi_{t_0}\tau^*_{t_0}\mathbf{1}_{\geq 0})^{(l)} \\
&= a_0\xi_{t_0}\tau^*_{t_0}\mathbf{1}_{\geq 0} + \sum_{l=1}^{k} a_l\left(\sum_{j=0}^{l-1} x_j(\tau^*_{t_0}\delta)^{(l-j-1)} + \xi^{(l)}_{t_0}\tau^*_{t_0}\mathbf{1}_{\geq 0}\right) \\
&= L_F(\theta_{\xi_{t_0}})\tau^*_{t_0}\mathbf{1}_{\geq 0} + \sum_{l=0}^{k-1}\sum_{j=0}^{l} a_{l+1}x_j\tau^*_{t_0}\delta^{(l-j)} \\
&= \theta_{b\tau^*_{t_0}\mathbf{1}_{\geq 0}} + \sum_{j=0}^{k-1}\sum_{j=l}^{k-1} a_{l+1}x_j\tau^*_{t_0}\delta^{(l-j)} \\
&= \theta_{b\tau^*_{t_0}\mathbf{1}_{\geq 0}} + \sum_{j=0}^{k-1}\sum_{l=0}^{k-j-1} a_{j+l+1}x_l\tau^*_{t_0}\delta^{(j)},
\end{aligned}$$

which is exactly condition (ii). It is evident from the preceding calculations that the conditions are, in fact, equivalent. ■

4.4.4 Notes

[Gates, Jr 1956]

Exercises

4.4.1 Let F be a scalar linear homogeneous ordinary differential equation with constant coefficients and let $\beta \in \mathscr{D}'_+(\mathbb{R};\mathbb{R})$. As in Theorem 4.4.7(ii), let θ_0 be the unique solution to (F,β) in $\mathscr{D}'_+(\mathbb{R};\mathbb{R})$. Show that $\operatorname{supp}(\theta_0) \subseteq [t_0,\infty)$ if $\operatorname{supp}(\beta) \subseteq [t_0,\infty)$.

4.4.2 Let F be a scalar linear ordinary differential equation with with constant coefficients given by

$$F(t,x,x^{(1)},\dots,x^{(k)}) = x^{(k)} + a_{k-1}x^{(k-1)} + \cdots + a_1x^{(1)} + a_0x.$$

Show, by direct computation, that the unique solution in $\mathscr{D}'_+(\mathbb{R};\mathbb{R})$ to the equation

$$\theta^{(k)} + a_{k-1}\theta^{(k-1)} + \cdots + a_1\theta^{(1)} + a_0\theta = \delta$$

is given by $\theta = \theta_{1_{\geq 0}\xi}$, where ξ is the solution to the initial value problem

$$\frac{\mathrm{d}^k\xi}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1(t)\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0(t)\xi(t) = 0,$$
$$\xi(0) = \cdots = \frac{\mathrm{d}^{k-2}\xi}{\mathrm{d}t^{k-2}}(0) = 0,\ \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(0) = 1.$$

Demonstrate that you understand each part of the computation by pointing to the place in the text where your assertion is defined or shown to make sense.

Section 4.5

Laplace transform methods for scalar ordinary differential equations

Laplace transforms can be used to study various sorts of differential equations, both partial and ordinary. In this section, we will stick to considering the application of Laplace transform techniques to the study of scalar linear ordinary differential equations with constant coefficients. We shall consider systems of equations in Section 5.4. The techniques we illustrate here can be thought of as the prototypical application of transform methods in the theory of differential equations and, moreover, is one of the more elementary applications of transform theory. Thus this section can be seen as having a twofold purpose: (1) to demonstrate the basic philosophy of transform analysis in the study of differential equations; (2) to develop fully an application of the Laplace transform to ordinary differential equations. To both ends, the emphasis will be on seeing how transforms can be helpful in understanding differential equations, rather than in solving differential equations (although we shall see that the latter is a part of the story).

Do I need to read this section? This is a section that can, maybe, be skipped. It will have its best context in the setting of transfer functions in Chapter 7. •

4.5.1 Scalar homogeneous equations

We begin our discussion with scalar linear homogeneous ordinary differential equations with constant coefficients, first considered in detail in Section 4.2.2. Thus, as in that section we are working with differential equations

$$F\colon \mathbb{R}_{\geq 0} \times \mathbb{R} \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$

with right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x \tag{4.24}$$

for $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$. Given Propositions IV-9.1.19 and IV-9.1.20, the causal CLT is particularly well suited for working with ordinary differential equations with initial conditions. Thus we shall consider the initial value problem

$$\begin{aligned} \frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0\xi(t) = 0,& \\ \xi(0) = x_0,\ \frac{\mathrm{d}\xi}{\mathrm{d}t}(0) = x_0^{(1)},\ \cdots,\ \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(0) = x_0^{(k-1)}.& \end{aligned} \tag{4.25}$$

We shall now take the causal CLT of this initial value problem. To do so, it is tacitly assumed that all members of $\mathrm{Sol}(F)$ and their derivatives are in $\mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; \mathbb{C})$ so

that we may use the derivative rule of Proposition IV-9.1.19 or IV-9.1.20. This is true, however, since all members of $\mathrm{Sol}(F)$ are also pretty uninteresting functions, and so are in $\mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0};\mathbb{C})$, when restricted to the domain $\mathbb{R}_{\geq 0}$, as we saw in Example IV-9.1.9–6. Another way to think of taking the causal CLT of the equation, were one to not know *a priori* that solutions were Laplace transformable, would be to go ahead and take the transform assuming this is so, and then see if the assumption is valid by seeing if the equation can be solved (or by some other means). In any case, the following result records what happens when we take the causal CLT of the initial value problem.

4.5.1 Proposition (Causal CLT of scalar homogeneous equation) *The causal CLT of the initial value problem (4.25) has the solution*

$$\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) = \frac{\sum_{j=0}^{k}\sum_{l=0}^{j-1} a_j z^l \xi^{(j-l-1)}(0)}{z^k + a_{k-1}z^{k-1} + \cdots + a_1 z + a_0},$$

with the convention that $a_k = 1$.

Proof By Corollary IV-9.1.22 we have

$$\mathscr{L}^{\infty}_{\mathrm{C}}\left(\frac{\mathrm{d}^j\xi}{\mathrm{d}t^j}\right)(z) = z^j\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) - \sum_{l=0}^{j-1} z^l\xi^{(j-l-1)}(0), \qquad j \in \{0,1,\ldots,k\}.$$

Therefore, with the stated convention that $a_k = 1$,

$$\mathscr{L}^{\infty}_{\mathrm{C}}\left(\sum_{j=0}^{k} a_j\frac{\mathrm{d}^j\xi}{\mathrm{d}t^j}\right) = \sum_{j=0}^{k} a_j\left(z^j\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) - \sum_{l=0}^{j-1} z^l\xi^{(j-l-1)}(0)\right),$$

and solving this equation for $\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z)$ gives the asserted conclusion. ■

To obtain the solution to the initial value problem in the time-domain, we should apply the inverse transform to the expression from the proposition. To do this, one could, in principle, apply the definition of the inverse causal CLT using the Fourier–Mellin integral (Definition IV-9.1.14). However, in cases where one can actually compute the inverse transform, it is not typically done in this way. Indeed, typically one "looks up" the answer. However, to do this requires a manipulation of the form of the expression from the proposition, and we outline this in the following procedure.

4.5.2 Procedure (Partial fraction expansion) While we shall apply the procedure to a $\mathbb{C}$-valued function of a complex variable (namely, the causal CLT of something), the construction is best explained in algebraic terms, so we present it in this way. Algebraically, the problem we are considering is a way of expressing a rational function, i.e., a quotient $R_{N,D} = \frac{N}{D}$ of polynomials N and D, in a manner where the roots of D and their multiplicities are accounted for properly.

Given two polynomials $N, D \in \mathbb{R}[X]$ with real coefficients, with D monic, with no common roots, and with $\deg(N) < \deg(D)$, do the following.

1. Find all roots of D and their multiplicities. Let the real roots be denoted by $r_1, \dots, r_l$ and let $m(r_j)$, $j \in \{1, \dots, l\}$, be the multiplicity of the root r_j. Let the complex roots be denote by $\rho_j = \sigma_j + \mathrm{i}\omega_j$, $\sigma_j \in \mathbb{R}$, $\omega_j \in \mathbb{R}_{>0}$, $j \in \{1, \dots, p\}$ (along with the complex conjugate roots $\sigma_j - \mathrm{i}\omega_j$) and let $m(\rho_j)$, $j \in \{1, \dots, p\}$, be the multiplicity of the root ρ_j.
2. Write
$$R_{N,D} = \sum_{j=1}^{l} \sum_{k=1}^{m(r_j)} \frac{a_{j,k}}{(X - r_j)^k} + \sum_{j=1}^{p} \sum_{k=1}^{m(\rho_j)} \frac{\alpha_{j,k} X + \beta_{j,k}}{((X - \sigma_j)^2 + \omega_j^2)^k} \tag{4.26}$$
for constants $a_{j,k} \in \mathbb{R}$, $j \in \{1, \dots, l\}$, $k \in \{1, \dots, m(r_j)\}$, and $\alpha_{j,k}, \beta_{j,k} \in \mathbb{R}$, $j \in \{1, \dots, p\}$, $k \in \{1, \dots, m(\rho_j)\}$, that are to be determined.
3. Express the right-hand side of (4.26) in the form
$$\frac{P}{(X - r_1)^{m(r_1)} \cdots (X - r_l)^{m(r_l)} ((X - \sigma_1)^2 + \omega_1^2)^{m(\rho_1)} \cdots ((X - \sigma_p)^2 + \omega_p^2)^{m(\rho_p)}},$$
for some polynomial $P \in \mathbb{R}[X]$.
4. By matching coefficients of powers of the indeterminate X, arrive at a set of linear algebraic equations for the constants $a_{j,k} \in \mathbb{R}$, $j \in \{1, \dots, l\}$, $k \in \{1, \dots, m(r_j)\}$, and $\alpha_{j,k}, \beta_{j,k} \in \mathbb{R}$, $j \in \{1, \dots, p\}$, $k \in \{1, \dots, m(\rho_j)\}$. It is a fact that these linear algebraic equations have a unique solution.
5. The ***partial fraction expansion*** of $R_{N,D}$ is then the right-hand side of the expression (4.26) with the constants as computed in the previous step. •

The idea of a partial fraction expansion in practice is straightforward, albeit quite tedious.

4.5.3 Examples (Partial fraction expansion)

1. We take $N = 5X + 4$ and $D = X^2 + X - 2$ so that
$$R_{N,D} = \frac{5X + 4}{X^2 + X - 2}.$$
We determine the roots of D to be $r_1 = 1$ and $r_2 = -2$, with $m(r_1) = m(r_2) = 1$. We then write
$$\frac{5X + 4}{X^2 + X - 2} = \frac{a_{1,1}}{X - 1} + \frac{a_{2,1}}{X + 2} = \frac{(a_{1,1} + a_{2,1})X + 2a_{1,1} - a_{2,1}}{(X - 1)(X + 2)}.$$
Thus, matching coefficients of powers of X in the numerator, we must have
$$a_{1,1} + a_{2,1} = 5, \quad 2a_{1,1} - a_{2,1} = 4 \quad \Longrightarrow \quad a_{1,1} = 3, \ a_{2,1} = 2.$$
Thus the partial fraction expansion is
$$R_{N,D} = \frac{3}{X - 1} + \frac{2}{X + 2}.$$

2. We take $N = -3X^2 + 5X + 2$ and $D = X^3 - 3X^2 + X - 3$, so that

$$R = \frac{-3X^2 + 5X + 2}{X^3 - 3X^2 + X - 3}.$$

The roots of the denominator polynomial are $r_1 = 3$, $\rho_1 = \mathrm{i}$, and $\bar{\rho}_1 = -\mathrm{i}$. We then write

$$\begin{aligned}\frac{-3X^2 + 5X + 2}{X^3 - 3X^2 + X - 3} &= \frac{a_{1,1}}{X - 3} + \frac{\alpha_{1,1}X + \beta_{1,1}}{(X - 0)^2 + 1} \\ &= \frac{(a_{1,1} + \alpha_{1,1})X^2 + (\beta_{1,1} - 3\alpha_{1,1})X + a_{1,1} - 3\beta_{1,1}}{(X - 3)(X^2 + 1)}.\end{aligned}$$

Matching coefficients of powers of X in the numerator, we must have

$$\begin{aligned}a_{1,1} + \alpha_{1,1} = -3,\ \ \beta_{1,1} - 3\alpha_{1,1} = 5,\ \ a_{1,1} - 3\beta_{1,1} = 2 \\ \Longrightarrow \quad a_{1,1} = -1,\ \alpha_{1,1} = -2,\ \beta_{1,1} = -1.\end{aligned}$$

The partial fraction expansion is

$$R_{N,D} = -\frac{1}{X - 3} - \frac{2X + 1}{X^2 + 1}.$$

3. We take $N = 2X^2 + 1$ and $D = X^3 + 3X^2 + 3X + 1$ so that

$$R_{N,D} = \frac{2X^2 + 1}{X^3 + 3X^2 + 3X + 1}.$$

The denominator polynomial has a single root $r_1 = -1$ which has multiplicity $m(r_1) = 3$. We write

$$\begin{aligned}\frac{2X^2 + 1}{X^3 + 3X^2 + 3X + 1} &= \frac{a_{1,1}}{X + 1} + \frac{a_{1,2}}{(X + 1)^2} + \frac{a_{1,3}}{(X + 1)^3} \\ &= \frac{a_{1,1}X^2 + (2a_{1,1} + a_{1,2})X + a_{1,1} + a_{1,2} + a_{1,3}}{(X + 1)^3}.\end{aligned}$$

Thus, matching coefficients of powers of X in the numerator,

$$\begin{aligned}a_{1,1} = 2,\ \ 2a_{1,1} + a_{1,2} = 0,\ \ a_{1,1} + a_{1,2} + a_{1,3} = 1 \\ \Longrightarrow \quad a_{1,1} = 2,\ a_{1,2} = -4,\ a_{1,3} = 3.\end{aligned}$$

Thus the partial fraction expansion is

$$R_{N,D} = \frac{2}{X + 1} - \frac{4}{(X + 1)^2} + \frac{3}{(X + 1)^3}.$$

•

There are complex function methods for computing the coefficients in a partial fraction decomposition, but we shall not present this here, mainly because this method for solving initial value problems offers very little in terms of insight, and nothing over the methods we learned in Procedure 4.2.18 for solving scalar linear homogeneous ordinary differential equations with constant coefficients. So presenting multiple methods for computing partial fraction expansions seems a little silly.

Now let us see how one uses the partial fraction expansion to compute the inverse causal CLT of the expression from Proposition 4.5.1. This is most easily done via examples.

4.5.4 Examples (Solving scalar homogeneous equations using the causal CLT)

1. Consider the initial value problem

$$\ddot{\xi}(t) + \dot{\xi}(t) - 2\xi(t) = 0, \quad \xi(0) = 5,\ \dot{\xi}(0) = -1.$$

Taking the causal CLT of the initial value problem gives

$$z^2\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) - z\xi(0) - \dot{\xi}(0) + z\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) - \xi(0) - 2\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) = 0$$
$$\implies \quad \mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) = \frac{5z+4}{z^2+z-2}.$$

Borrowing our partial fraction expansion from Example 4.5.3–1 we have

$$\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) = \frac{3}{z-1} + \frac{2}{z+2}.$$

Thus, referring to Example IV-9.1.15–2,

$$\xi(t) = 3\mathrm{e}^{t} + 2\mathrm{e}^{-2t}.$$

2. Consider the initial value problem

$$\dddot{\xi}(t) - 3\ddot{\xi}(t) + \dot{\xi}(t) - 3\xi(t) = 0, \quad \xi(0) = -3,\ \dot{\xi}(0) = -4,\ \ddot{\xi}(0) = -7.$$

Taking the causal CLT of the initial value problem gives

$$z^3\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) - z^2\xi(0) - z\dot{\xi}(0) - \ddot{\xi}(0) - 3z^2\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) + 3z\xi(0)$$
$$+ 3\dot{\xi}(0) + z\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) - \xi(0) - 3\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) = 0$$
$$\implies \quad \mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) = \frac{-3z^2+5z+2}{z^3-3z^2+z-3}.$$

Borrowing our partial fraction expansion from Example 4.5.3–1 we have

$$\mathscr{L}_{\mathrm{C}}^{\infty}(\xi)(z) = -\frac{1}{z-3} - \frac{2z+1}{z^2+1}.$$

Thus, referring to Example IV-9.1.15–2 and Example IV-9.1.15–4,

$$\xi(t) = -\mathrm{e}^{3t} - 2\cos(t) - \sin(t).$$

3. Consider the initial value problem

$$\dddot{\xi}(t) + 3\ddot{\xi}(t) + 3\dot{\xi}(t) + \xi(t) = 0, \quad \xi(0) = 2,\ \dot{\xi}(0) = -6,\ \ddot{\xi}(0) = 13.$$

Taking the causal CLT of the initial value problem gives

$$\begin{aligned} z^3\mathscr{L}^\infty_{\mathrm{C}}(\xi)(z) - z^2\xi(0) - z\dot{\xi}(0) - \ddot{\xi}(0) + 3z^2\mathscr{L}^\infty_{\mathrm{C}}(\xi)(z) - 3z\xi(0) - 3\dot{\xi}(0) \\ + 3z\mathscr{L}^\infty_{\mathrm{C}}(\xi)(z) - 3\xi(0) + \mathscr{L}^\infty_{\mathrm{C}}(\xi)(z) = 0 \\ \implies \quad \mathscr{L}^\infty_{\mathrm{C}}(\xi)(z) = \frac{2z^2+1}{z^3+3z^2+3z+1}. \end{aligned}$$

Borrowing our partial fraction expansion from Example 4.5.3–1 we have

$$\mathscr{L}^\infty_{\mathrm{C}}(\xi)(z) = \frac{2}{z+1} - \frac{4}{(z+1)^2} + \frac{3}{(z+1)^3}.$$

Thus, referring to Example IV-9.1.15–2,

$$\xi(t) = 2\mathrm{e}^{-t} - 4t\mathrm{e}^{-t} + \frac{3}{2}t^2\mathrm{e}^{-t}. \qquad \bullet$$

The above business about partial fraction expansions gives a reader who likes doing algorithmic computations a venue to exercise this skill. However, it is not really the point of the causal CLT. The really useful feature of the causal CLT for linear differential equations, and not just those equations that are scalar and homogeneous, is that initial value problems are converted into algebraic expressions. The use of partial fraction expansions to determine the inverse causal CLT of these algebraic expressions is something of a novelty act.

4.5.2 Scalar inhomogeneous equations

We next consider scalar linear inhomogeneous ordinary differential equations, first considered in Section 4.3.2. Thus we are working with scalar ordinary differential equations with right-hand sides given by

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1x^{(1)} - a_0x + b(t) \tag{4.27}$$

for $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$ and $b\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$. The initial value problem we consider is then

$$\begin{aligned} \frac{\mathrm{d}^k\xi(t)}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) + a_0\xi(t) = b(t), \\ \xi(0) = x_0,\ \frac{\mathrm{d}\xi}{\mathrm{d}t}(0) = x_0^{(1)},\ \cdots,\ \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(0) = x_0^{(k-1)}. \end{aligned} \tag{4.28}$$

If $b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$, then the solution to the initial value problem (4.28) is given by $\xi(t) = \xi_{\mathrm{h}}(t) + H_F * b(t)$, where ξ_{h} satisfies the homogeneous initial value problem

$$\frac{\mathrm{d}^k \xi_{\mathrm{h}}(t)}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}\xi_{\mathrm{h}}}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}\xi_{\mathrm{h}}}{\mathrm{d}t}(t) + a_0\xi_{\mathrm{h}}(t) = 0,$$
$$\xi_{\mathrm{h}}(0) = x_0, \ \frac{\mathrm{d}\xi_{\mathrm{h}}}{\mathrm{d}t}(0) = x_0^{(1)}, \ \cdots, \ \frac{\mathrm{d}^{k-1}\xi_{\mathrm{h}}}{\mathrm{d}t^{k-1}}(0) = x_0^{(k-1)},$$

where $H_F(t-\tau) = G_F(t,\tau)$ and where G_F is the Green's function from Section 4.3.1.3. This follows from Remark 4.3.11 and Exercise 4.3.2.

We wish to provide an interpretation of this strategy using the causal CLT and the connection of the transform and convolution from Proposition IV-9.1.10. As with inhomogeneous equations above, we take the causal CLT of the equation (4.28). However, unlike in the homogeneous case, here taking the transform is not generally valid; indeed, it is valid if and only if $b \in \mathsf{LT}^{1,+}(\mathbb{R}_{\geq 0}; \mathbb{R})$. To apply the convolution result from Proposition IV-9.1.10, we further assume that $b \in \mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; \mathbb{R})$.

First let us give determine the causal CLT of the Green's function in this case.

4.5.5 Proposition (Causal CLT and the Green's function) *Consider the scalar linear homogeneous ordinary differential equation* F *with right-hand side (4.24). Let* $\mathrm{G_F}$ *be the Green's function and denote* $\mathrm{H_F}(\mathrm{t}-\tau) = \mathrm{G_F}(\mathrm{t},\tau)$. *Then the causal CLT of* $\mathrm{H_F}$ *is given by*

$$\mathscr{L}^{\infty}_{\mathrm{C}}(\mathrm{H_F})(\mathrm{z}) = \frac{1}{\mathrm{z^k} + \mathrm{a_{k-1}z^{k-1}} + \cdots + \mathrm{a_z z} + \mathrm{a_0}}.$$

Proof According to Remark 4.3.11, $G_F(t,\tau) = H_F(t-\tau)$, where H_F satisfies the initial value problem

$$\frac{\mathrm{d}^k H_F}{\mathrm{d}t^k}(t) + a_{k-1}\frac{\mathrm{d}^{k-1}H_F}{\mathrm{d}t^{k-1}}(t) + \cdots + a_1\frac{\mathrm{d}H_F}{\mathrm{d}t}(t) + a_0 H_F(t) = 0,$$
$$H_F(0) = 0, \ \frac{\mathrm{d}H_K}{\mathrm{d}t}(0) = 0, \ldots, \ \frac{\mathrm{d}^{k-2}H_F}{\mathrm{d}t^{k-2}}(0) = 0, \ \frac{\mathrm{d}^{k-1}H_F}{\mathrm{d}t^{k-1}}(0) = 1.$$

Therefore, according to Proposition 4.5.1,

$$\mathscr{L}_{\mathrm{C}}(H_F)(z) = \frac{1}{z^k + a_{k-1}z^{k-1} + \cdots + a_z z + a_0},$$

as claimed. ∎

By combining Proposition 4.5.1 with the preceding result and the convolution solution $\xi(t) = \xi_{\mathrm{h}}(t) + H_F * b(t)$ of the initial value problem (4.28), we obtain the following result.

4.5.6 Proposition (Causal CLT of scalar inhomogeneous equation) *Consider the scalar ordinary differential equation with right-hand side (4.27), and suppose that* $b \in \mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; \mathbb{R})$. *The causal CLT of the solution of the initial value problem (4.28) is given by*

$$\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) = \frac{\sum_{j=0}^{k} \sum_{l=0}^{j-1} a_j z^l \xi^{(j-l-1)}(0) + \mathscr{L}^{\infty}_{\mathrm{C}}(b)(z)}{z^k + a_{k-1} z^{k-1} + \cdots + a_1 z + a_0},$$

with the convention that $a_k = 1$.

There are two ways in which the proposition has value. One is theoretical and one is that it provides another tedious algorithmic procedure—augmenting the "method of undetermined coefficients"—for computing solutions when the inhomogeneous term is an also pretty uninteresting function. Let us consider these in turn.

Next let us turn to a less interesting but somehow more concrete application of the causal CLT in the study of scalar linear inhomogeneous ordinary differential equations. Specifically, we consider such an equation F with right-hand side (4.27), and where b is an also pretty uninteresting function. In this case, as we see from Example IV-9.1.9, the causal CLT $\mathscr{L}^{\infty}_{\mathrm{C}}(b)$ of b will be a rational function of the complex variable z whose numerator polynomial has degree strictly less than that of the denominator polynomial. Therefore, as per Proposition 4.5.6, the causal CLT $\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)$ of the solution ξ of the initial value problem (4.28) will itself be such a rational function of z. Thus we can perform a partial fraction expansion of $\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)$ as per Procedure 4.5.2, and then perform the inversion of the causal CLT as per Example 4.5.4 to obtain the solution. This is not something to be belaboured—not least because we already have the often easier "method of undetermined coefficients" for such situations–and we content ourselves with an illustration via a example.

4.5.7 Example (Solving scalar inhomogeneous equations using the causal CLT) We consider the initial value problem

$$\ddot{\xi}(t) + \omega^2 \xi(t) = \sin(\omega t), \quad \xi(0) = x_0, \ \dot{\xi}(0) = x_0^{(1)},$$

for $\omega \in \mathbb{R}_{>0}$. Using Example IV-9.1.9–4 we compute the causal CLT of this initial value problem:

$$z^2 \mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) - z x_0 - x_0^{(1)} + \omega^2 \mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) = \frac{\omega}{z^2 + \omega^2}$$

$$\implies \mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) = \frac{\omega}{(z^2 + \omega^2)^2} + \frac{z x_0 + x_0^{(1)}}{z^2 + \omega^2}.$$

Using Example IV-9.1.15–4 and Example IV-9.1.15–5 we have

$$\xi(t) = x_0 \cos(\omega t) + \frac{x_0^{(1)}}{\omega} \sin(\omega t) - \frac{t}{2\omega} \cos(\omega t) + \frac{1}{2\omega^2} \sin(\omega t). \qquad \bullet$$

Exercises

4.5.1 Determine the Laplace transform of the solution of the following initial value problems:

(a) $\dot{\xi}(t) + 3\xi(t) = 0$, $\xi(0) = 4$;

(b) $\ddot{\xi}(t) - 4\dot{\xi}(t) + 4\xi(t) = 0$, $\xi(0) = 0$, $\dot{\xi}(0) = 1$;

(c) $\ddot{\xi}(t) - 4\dot{\xi}(t) - 4\xi(t) = 0$, $\xi(0) = 1$, $\dot{\xi}(0) = 1$;

(d) $\dddot{\xi}(t) - 7\ddot{\xi}(t) + 15\dot{\xi}(t) - 9\xi(t) = 0$, $\xi(0) = 1$, $\dot{\xi}(0) = 1$, $\ddot{\xi}(0) = 1$;

(e) $\dddot{\xi}(t) + 3\ddot{\xi}(t) + 4\dot{\xi}(t) + 2\xi(t) = 0$, $\xi(0) = 0$, $\dot{\xi}(0) = 1$, $\ddot{\xi}(0) = 2$;

(f) $\ddddot{\xi}(t) + \dddot{\xi}(t) + \ddot{\xi}(t) + \dot{\xi}(t) + \xi(t) = 0$, $\xi(0) = 0$, $\dot{\xi}(0) = 0$, $\ddot{\xi}(0) = 0$, $\dddot{\xi}(0) = 0$.

NB. These are the same initial value problems you worked out in Exercise 4.2.10.

4.5.2 Using partial fraction expansion, determine the solution to the initial value problems from Exercise 4.5.1.

4.5.3 Determine the Laplace transform of the solution for the following scalar linear inhomogeneous differential equations F with the stated initial conditions:

(a) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + 2x^{(1)} + x - 3\mathrm{e}^t$, and $\xi(0) = 1$, $\dot{\xi}(0) = 1$;

(b) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} - 5x^{(1)} + 6x - 2\mathrm{e}^{3t} - \cos(t)$, and $\xi(0) = 0$, $\dot{\xi}(0) = 1$;

(c) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} - 2x^{(1)} + 5x - t\mathrm{e}^t \sin(2t)$, and $\xi(0) = 1$, $\dot{\xi}(0) = 0$;

(d) $F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + 4x - t\cos(2t) + \sin(2t)$, and $\xi(0) = 2$, $\dot{\xi}(0) = 1$;

(e) $F(t, x^{(1)}, x^{(2)}, x^{(3)}) = x^{(3)} - x - t\mathrm{e}^t$, and $\xi(0) = 1$, $\dot{\xi}(0) = 1$, $\ddot{\xi}(0) = 1$;

(f) $F(t, x, x^{(1)}, \dots, x^{(4)}) = x^{(4)} + 4x^{(2)} + 4x - \cos(2t) - \sin(2t)$, and $\xi(0) = 0$, $\dot{\xi}(0) = 0$, $\ddot{\xi}(t) = 0$, $\dddot{\xi}(t) = 0$.

NB. These are the same initial value problems you worked out in Exercise 4.3.5.

4.5.4 Using partial fraction expansion, determine the solution to the initial value problems from Exercise 4.5.3.

Section 4.6

Scalar linear homogeneous ordinary difference equations

In this section we shall mirror the results given in Section 4.2 for differential equations for difference equations. Here the equations have time-domain $\mathbb{T} = I \cap \mathbb{Z}(h)$ for an interval I and state space $U = \mathbb{R}$. The right-hand sides we consider are then of the form

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x \tag{4.29}$$

for functions $a_0, a_1, \dots, a_{k-1} \colon \mathbb{T} \to \mathbb{R}$. Thus solutions $t \mapsto \xi(t)$ satisfy

$$\xi(t + kh) + a_{k-1}(t)\xi(t + (k-1)h) + \cdots + a_1(t)\xi(t + h) + a_0(t)\xi(t) = 0.$$

We recall that the free domain is

$$\mathbb{T}_F = \{t \in \mathbb{T} \mid \ t + kh \in \mathbb{T}\}.$$

This means that we only use the values of the coefficients on $\mathbb{T}_F$, although we will not be fussy about this. We shall (1) examine the character of solution, (2) examine the set of all solutions, and (3) provide an in-principle procedure for solving these equations in the constant coefficient case.

Do I need to read this section? This section contains tools that are standard for anyone claiming to know something about ordinary difference equations. •

4.6.1 Equations with time-varying coefficients

We start by a consideration of the general situation where the coefficients $a_0, a_1, \dots, a_{k-1}$ depend on time. We shall, as we did with the corresponding differential equations, consider the properties of solutions and sets of solutions. We shall also introduce the discrete-time analogue of the Wronskian, the so-called "Casoratian."

4.6.1.1 Solutions and their properties

Let us state the adaptation to our current setting of the existence and uniqueness results of Section 3.4.

4.6.1 Proposition (Existence and uniqueness of solutions for scalar linear homogeneous ordinary difference equations) *Consider the linear homogeneous ordinary difference equation* F *with right-hand side (4.29). Let*

$$(\mathrm{t}_0, \mathrm{x}_0, \mathrm{x}_0^{(1)}, \dots, \mathrm{x}_0^{(\mathrm{k}-1)}) \in \mathbb{T}_{\mathrm{F}} \times \mathbb{R} \oplus \mathrm{L}_{\mathrm{sym}}^{\leq \mathrm{k}-1}(\mathbb{R}; \mathbb{R}).$$

Then there exists a unique $\xi\colon \mathbb{T}_{\geq t_0} \to \mathbb{R}$ *that is a solution for* F *and which satisfies*

$$\xi(t_0) = x_0, \xi(t_0 + h) = x_0^{(1)}, \dots, \xi(t_0 + (k-1)h) = x_0^{(k-1)}. \tag{4.30}$$

If F *is invertible, then there exists a unique* $\xi\colon \mathbb{T} \to \mathbb{R}$ *that is a solution for* F *and which satisfies (4.30).*

Proof Since the state space is $U = \mathbb{R}$, it follows that F is complete and so the first assertion follows from Theorem 3.4.2. The second assertion follows from Theorem 3.4.6. ■

There are some important differences between the preceding theorem and its counterpart Proposition 4.2.2 for differential equations.

1. There are no regularity requirements in Proposition 4.6.1 on the coefficients of the difference equation, nor any regularity conclusions for solutions. This is a consequence of the discreteness of the time-domain.
2. Generally, one can only assert existence forward in time. For solutions to exist backwards in time also requires invertibility of the difference equation. This is in contrast to differential equations in Proposition 4.2.2, where one always has solutions for all forward and all backward times.
3. For kth-order difference equations, one specifies k initial conditions at k different times. This is in contrast to kth-order differential equations where, in Proposition 4.2.2, one prescribes k initial conditions at the same time.

The second of the above points leads us to understand the character of invertible scalar linear ordinary difference equations, and the following result gives this.

4.6.2 Proposition (Invertible scalar linear homogeneous ordinary difference equations) *A scalar linear homogeneous ordinary difference equation* F *with right-hand side*

$$\widehat{\mathrm{F}}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x,$$

where $a_0, a_1, \dots, a_{k-1}\colon \mathbb{T} \to \mathbb{R}$*, is invertible if and only if* $a_0(t) \neq 0$ *for every* $t \in \mathbb{T}_F$.

Proof This is most easily proved by writing the kth-order scalar equation as a first-order vector equation in k variables, as in Exercise 3.3.7. Upon doing so, one can use Proposition 5.6.2 to get the desired result. ■

As with differential equations, we can consider the space of solutions. Thus we consider a scalar linear homogeneous ordinary difference equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x,$$

where $a_0, a_1, \dots, a_{k-1}\colon \mathbb{T} \to \mathbb{R}$. For $t_0 \in \mathbb{T}_F$, let us denote by

$$\mathrm{Sol}_{t_0}(F) = \Big\{\xi \in \mathbb{R}^{\mathbb{T}_{\geq t_0}} \Big| \; \xi(t + kh) + a_{k-1}(t)\xi(t + (k-1)h) + \cdots + a_1(t)\xi(t + h) + a_0(t)\xi(t) = 0,\ t \in \mathbb{T}_{F, \geq t_0}\Big\}$$

the set of solutions for F starting from t_0. In case F is invertible, we can define

$$\mathrm{Sol}(F) = \Big\{\xi \in \mathbb{R}^{\mathbb{T}} \Big| \\ \xi(t+kh) + a_{k-1}(t)\xi(t+(k-1)h) + \cdots + a_1(t)\xi(t+h) + a_0(t)\xi(t) = 0,\ t \in \mathbb{T}_F\Big\}.$$

The following result is then the main structural result for the class of differential equations we are considering in this section.

4.6.3 Theorem (Vector space structure of sets of solutions) *Consider the linear homogeneous ordinary differential equation* F *with right-hand side (4.29). Then, for* $\mathrm{t}_0 \in \mathbb{T}_{\mathrm{F}}$, $\mathrm{Sol}_{\mathrm{t}_0}(\mathrm{F})$ *is a* k*-dimensional subspace of* $\mathbb{R}^{\mathbb{T}_{\geq \mathrm{t}_0}}$. *If* F *is additionally invertible, then* Sol(F) *is a* k*-dimensional subspace of* $\mathbb{R}^{\mathbb{T}}$.

Proof That $\mathrm{Sol}_{t_0}(F)$ and $\mathrm{Sol}(F)$ are subspaces is easily shown, rather similarly to the proof of Theorem 4.2.3. To prove that the subspaces are of dimension k, we shall consider $\mathrm{Sol}_{t_0}(F)$ with the case of $\mathrm{Sol}(F)$ following similarly.

We show that the mapping

$$\begin{aligned}\sigma_{t_0} &: \mathrm{Sol}_{t_0}(F) \to \mathbb{R}^k \\ &\quad \xi \mapsto (\xi(t_0), \xi(t_0+h), \dots, \xi(t_0+(k-1)h))\end{aligned}$$

is an isomorphism of $\mathbb{R}$-vector spaces. The map is surjective by the existence part of Proposition 4.6.1. Linearity of σ_{t_0} is easily shown, cf. the proof of Theorem 4.2.3. Given this linearity, to show injectivity of σ_{t_0} it suffices to show that $\ker(\sigma_{t_0}) = \{0\}$ by Exercise I-4.5.23. Suppose, then, that $\sigma_{t_0}(\xi) = \mathbf{0}$ so that

$$\xi(t_0) = 0,\ \xi(t_0+h) = 0, \dots, \xi(t_0+(k-1)h) = 0.$$

It then follows directly by induction that $\xi(t_0 + jh) = 0$ for all $j \in \mathbb{Z}_{\geq 0}$, and so $\xi = 0$. ■

4.6.4 Definition (Fundamental set of solutions) Consider the linear homogeneous ordinary difference equation F with right-hand side (4.29).

(i) A set $\{\xi_1, \dots, \xi_k\}$ of linearly independent elements of $\mathrm{Sol}_{t_0}(F)$ is a ***fundamental set of solutions*** for F from t_0.

(ii) If F is invertible, then a set $\{\xi_1, \dots, \xi_k\}$ of linearly independent elements of $\mathrm{Sol}_{t_0}(F)$ is a ***fundamental set of solutions*** for F. •

As with differential equations, there is not much one can say in general about time-varying linear equations. We, therefore, content ourselves by considering a simple example in detail.

4.6.5 Example (First-order scalar linear homogeneous equations) We consider a first-order scalar linear homogeneous ordinary difference equation given by

$$\begin{aligned}F &: \mathbb{T} \times \mathbb{R} \times \mathrm{L}^1_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R} \\ &\quad (t, x, x^{(1)}) \mapsto x^{(1)} + a(t)x,\end{aligned}$$

for some $a\colon \mathbb{T} \to \mathbb{R}$. We take $\mathbb{T} \subseteq \mathbb{Z}(h)$. Solutions then satisfy an initial value problem

$$\xi(t+h) + a(t)\xi(t) = 0, \quad \xi(t_0) = x_0. \tag{4.31}$$

We shall write points in $\mathbb{T}$ as jh for $j \in \mathbb{Z}$, and we write $t_0 = j_0 h$. We claim that the solution to this equation is

$$\xi(j_0 h) = x_0, \ \ \xi(jh) = (-1)^{j-j_0} \prod_{l=j_0}^{j-1} a(lh) x_0, \qquad j \in \mathbb{T}_{>j_0 h}.$$

We shall simply verify that this does indeed satisfy the initial value problem. First of all, the initial condition is satisfied by declaration. To see that ξ satisfies the difference equation, we use induction to compute

$$\xi((j_0+1)h) = (-1)^{(j_0+1)-j_0} \left(\prod_{l=j_0}^{j_0} a(lh) \right) x_0 = -a(j_0 h)\xi(j_0 h)$$

and then, for $j > j_0$,

$$\begin{aligned} \xi((j+1)h) = -a(jh)\xi(jh) = \ & - a(jh)(-1)^{j-j_0} \left(\prod_{l=j_0}^{j-1} a(lh) \right) x_0 \\ & = (-1)^{(j+1)-j_0} \left(\prod_{l=j_0}^{j} a(lh) \right) x_0, \end{aligned}$$

as claimed.

We note that F is invertible if and only if $a(t) \neq 0$ for every $t \in \mathbb{T}$. Indeed, if $a(t) \neq 0$ for every $t \in \mathbb{T}$, we have the inverse difference equation

$$\begin{aligned} F^{-1}\colon \mathbb{T} \times \mathbb{R} \times \mathrm{L}^{\leq 1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) &\to \mathbb{R} \\ (t, x, x^{(1)}) &\mapsto x^{(1)} - a(t+h)^{-1} x. \end{aligned}$$

In this case, for $j < j_0$, we can define the solution to the initial value problem (4.31) by

$$\xi(jh) = (-1)^{j_0 - j} \left(\prod_{l=j}^{j_0-1} a(lh)^{-1} \right) x_0.$$

We leave to the reader the simple verification that the resulting function

$$\xi(jh) = \begin{cases} (-1)^{j_0-j} \left(\prod_{l=j}^{j_0-1} a(lh)^{-1} \right) x_0, & j < j_0, \\ x_0, & j = j_0, \\ (-1)^{j-j_0} \left(\prod_{l=j_0}^{j-1} a(lh) \right) x_0, & j > j_0, \end{cases}$$

is indeed a solution to the initial value problem (4.31) when F is invertible. •

We encourage the reader to compare the solution to this difference equation to the solution to the corresponding differential equation given in Example 4.2.5, and to come to peace with the idea that they are one and the same thing, *mutatis mutandis*.[4]

4.6.1.2 The Casoratian, and its properties and uses We give in this section the analogue for difference equations of the Wronskian introduced in Section 4.2.1.2.

4.6.6 Definition (Casoratian[5]) Let $\mathbb{T} \subseteq \mathbb{Z}(h)$ be a discrete time-domain and let $f_1, \dots, f_k \colon \mathbb{T} \to \mathbb{R}$. Denote

$$\mathbb{T}_k = \{t \in \mathbb{T} \mid t + (k-1)h \in \mathbb{T}\}.$$

The ***Casoratian*** for the functions $f_1, \dots, f_k$ is the function $\mathrm{C}(f_1, \dots, f_k) \colon \mathbb{T}_k \to \mathbb{R}$ defined by

$$\mathrm{C}(f_1, \dots, f_k)(t) = \det \begin{bmatrix} f_1(t) & f_2(t) & \cdots & f_k(t) \\ f_1(t+h) & f_2(t+h) & \cdots & f_k(t+h) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(t+(k-1)h) & f_2(t+(k-1)h) & \cdots & f_k(t+(k-1)h) \end{bmatrix}.$$

If $\mathbb{T}_k = \emptyset$, we take the convention that $\mathrm{C}(f_1, \dots, f_k)(t) = 0$ for $t \in \mathbb{T}$. •

An essential feature of the Casoratian is that it gives a sufficient condition for measuring the linear independence of finite sets of functions in the space of functions. More precisely, we have the following result, which again is not *a priori* related to differential equations.

4.6.7 Proposition (The Casoratian and linear independence) *Let $\mathbb{T} \subseteq \mathbb{Z}(\mathrm{h})$ be a discrete time-domain and let $\mathrm{f}_1, \dots, \mathrm{f}_\mathrm{k} \colon \mathbb{T} \to \mathbb{R}$. If $\mathrm{C}(\mathrm{f}_1, \dots, \mathrm{f}_\mathrm{k})(\mathrm{t}) \neq 0$ for some $\mathrm{t} \in \mathbb{T}_\mathrm{k}$, then the set $\{\mathrm{f}_1, \dots, \mathrm{f}_\mathrm{k}\}$ is linearly independent in $\mathbb{R}^{\mathbb{T}}$.*

Proof We prove the contrapositive, i.e., that, if the functions $\{f_1, \dots, f_k\}$ are linearly dependent, then $\mathrm{C}(f_1, \dots, f_k)(t) = 0$ for all $t \in \mathbb{T}_k$.

So suppose that $\{f_1, \dots, f_k\}$ is linearly dependent, and let $c_1, \dots, c_k \in \mathbb{R}$, not all zero, be such that

$$c_1 f_1 + \cdots + c_k f_k = 0.$$

Then, for any $j \in \{1, \dots, k-1\}$ and $t \in \mathbb{T}_k$,

$$c_1 f_1(t + jh) + \cdots + c_n f_n(t + jh) = 0.$$

[4] A key part of this, as shall become apparent after awhile, is that integrals for differential equations get replaced by sums for difference equations (as expected), while exponentials for differential equations get replaced by products (perhaps less expected initially).

[5] After Felice Casorati (1835–1890), an Italian mathematician who made contributions in the areas of differential equations and complex analysis.

Assembling these relationships for $j \in \{0, 1, \dots, k-1\}$ gives the single equation

$$\begin{bmatrix} f_1(t) & f_2(t) & \cdots & f_k(t) \\ f_1(t+h) & f_2(t+h) & \cdots & f_k(t+h) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(t+(k-1)h) & f_2(t+(k-1)h) & \cdots & f_k(t+(k-1)h) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

This means that the matrix on the left has a nontrivial kernel (since this kernel contains $(c_1, \dots, c_k)$) and so must have zero determinant. ■

Note that the converse of the preceding result is not generally true, as demonstrated by the following example.

4.6.8 Example (The Casoratian is not adequate to characterise linear independence) Let $\mathbb{T} = \mathbb{Z}$ and consider the two functions $f_1, f_2 \colon \mathbb{Z} \to \mathbb{R}$ defined by

$$f_1(t) = \begin{cases} 1, & t = 0, \\ 0, & \text{otherwise}, \end{cases} \qquad f_1(t) = \begin{cases} 1, & t = 2, \\ 0, & \text{otherwise}. \end{cases}$$

We then directly verify that $\mathrm{C}(f_1, f_2)(t) = 0$ for all $t \in \mathbb{Z}$. However, f_1 and f_2 are linearly independent. Indeed, suppose that $c_1, c_2 \in \mathbb{R}$ satisfy

$$c_1 f_1(t) + c_2 f_2(t) = 0, \qquad t \in \mathbb{Z}.$$

Then, taking $t = 0$, we get $c_1 = 0$ and taking $t = 2$ we get $c_2 = 0$. •

Thus the Casoratian is not quite the thing for precisely characterising the linear independence of general sets of functions. By examining the previous example, as astute reader may be able to guess the correct condition for the linear independence of discrete-time signals. We refer to Exercise 4.6.1 for a spelling out of these conditions. As with the Wronskian and differential equations, the Casoratian is just the thing when the set of functions under consideration are solutions to a scalar linear homogeneous ordinary difference equation, especially when that difference equation is invertible.

4.6.9 Proposition (Casoratians and linear independence in Sol(F)) *Consider the linear homogeneous ordinary difference equation* F *with right-hand side (4.29). Then the following statements are equivalent for* $\mathrm{t}_0 \in \mathbb{T}_{\mathrm{F}}$ *and for* $\xi_1, \dots, \xi_{\mathrm{k}} \in \mathrm{Sol}_{\mathrm{t}_0}(\mathrm{F})$*:*

(i) $\{\xi_1, \dots, \xi_{\mathrm{k}}\}$ *is linearly independent;*

(ii) $\mathrm{C}(\xi_1, \dots, \xi_{\mathrm{k}})(\mathrm{t}) \neq 0$ *for some* $\mathrm{t} \in \mathbb{T}_{\mathrm{F},\geq \mathrm{t}_0}$.

If, additionally, F *is invertible, then the preceding statements are equivalent to the following:*

(iii) $\mathrm{C}(\xi_1, \dots, \xi_{\mathrm{k}})(\mathrm{t}) \neq 0$ *for all* $\mathrm{t} \in \mathbb{T}_{\mathrm{F},\geq \mathrm{t}_0}$.

Proof (i) $\implies$ (ii) We prove the contrapositive, i.e., we prove that, if $\mathrm{C}(\xi_1, \dots, \xi_k)(t) = 0$ for all $t \in \mathbb{T}_{F,\geq t_0}$, then $\{\xi_1, \dots, \xi_k\}$ is linearly dependent.

So suppose that $\mathrm{C}(\xi_1, \dots, \xi_k)(t) = 0$ for all $t \in \mathbb{T}_{F,\geq t_0}$, which means that there exists $c_1, \dots, c_k \in \mathbb{R}$, not all zero, such that

$$\begin{bmatrix} \xi_1(t) & \xi_2(t) & \cdots & \xi_k(t) \\ \xi_1(t+h) & \xi_2(t+h) & \cdots & \xi_k(t+h) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1(t+(k-1)h) & \xi_2(t+(k-1)h) & \cdots & \xi_k(t+(k-1)h) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

for all $t \in \mathbb{T}_{F,\geq t_0}$. If we simply expand this out, we see that it is equivalent to

$$c_1 \sigma_{t_0}(\xi_1) + \cdots + c_k \sigma_{t_0}(\xi_k) = 0,$$

where $\sigma_{t_0} \colon \mathrm{Sol}_{t_0}(F) \to \mathbb{R}^k$ is the isomorphism defined by

$$\sigma_{t_0}(\xi) = (\xi(t_0), \xi(t_0 + h), \dots, \xi(t_0 + (k-1)h),$$

cf. the proof of Theorem 4.6.3. Since σ_{t_0} is linear, this gives

$$\sigma_{t_0}(c_1\xi_1 + \cdots + c_k\xi_k) = 0, \qquad t \in \mathbb{T}_F.$$

Injectivity of σ_{t_0} then gives

$$c_1\xi_1 + \cdots + c_k\xi_k = 0,$$

showing linear dependence of $\{\xi_1, \dots, \xi_k\}$.

(ii) $\implies$ (i) This follows from Proposition 4.6.7.

(ii) $\implies$ (iii) From Proposition 4.6.7 the assumption of (ii) implies that $\{\xi_1, \dots, \xi_k\}$ is linearly independent. Suppose now that there exists $t \in \mathbb{T}_F$ such that $\mathrm{C}(\xi_1, \dots, \xi_k)(t) = 0$. Then there exists $c_1, \dots, c_k \in \mathbb{R}$, not all zero, such that

$$\begin{bmatrix} \xi_1(t) & \xi_2(t) & \cdots & \xi_k(t) \\ \xi_1(t+h) & \xi_2(t+h) & \cdots & \xi_k(t+h) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1(t+(k-1)h) & \xi_2(t+(k-1)h) & \cdots & \xi_k(t+(k-1)h) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{4.32}$$

Now, define $\xi \colon \mathbb{T}_{\geq t_0} \to \mathbb{R}$ by

$$\xi = c_1\xi_1 + \cdots + c_k\xi_k.$$

By Theorem 4.6.3, $\xi \in \mathrm{Sol}(F)$. Moreover, the equation (4.32) gives

$$\xi(t) = 0,\ \xi(t+h) = 0, \dots, \xi(t+(k-1)h) = 0.$$

By Proposition 4.6.1 in the case that F is invertible, we conclude that $\xi(t) = 0$ for all $t \in \mathbb{T}$. This contradicts the linear independence of $\{\xi_1, \dots, \xi_k\}$.

It is evident that (iii) $\implies$ (ii). ∎

The following result gives an interesting characterisation of the Casoratian.

4.6.10 Proposition (Abel's formula) *Consider the scalar linear homogeneous ordinary difference equation* F *with right-hand side (4.29). If* $\{\xi_1, \dots, \xi_k\}$ *are linearly independent, then, for any* $t_0 \in \mathbb{T}_F$ *and* $j \in \mathbb{Z}_{\geq 0}$ *such that* $t_0 + jh \in \mathbb{T}$,

$$C(\xi_1, \dots, \xi_k)(t_0 + jh) = C(\xi_1, \dots, \xi_k)(t_0)(-1)^{kj} \left(\prod_{l=1}^{j-1} a_0(t_0 + lh) \right)$$

Proof We have

$$C(\xi_1, \dots, \xi_k)((j+1)h) = \det \begin{bmatrix} \xi_1((j+1)h) & \xi_2((j+1)h) & \cdots & \xi_k((j+1)h) \\ \xi_1((j+2)h) & \xi_2((j+2)h) & \cdots & \xi_k((j+2)h) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1((j+k)h) & \xi_2((j+k)h) & \cdots & \xi_k((j+k)h) \end{bmatrix}.$$

By the satisfaction of the difference equation, we have

$$\xi_l((j+k)h) = -a_0(jh)\xi_l(jh) - \sum_{r=1}^{k-1} a_r(jh)\xi_l((j+r)h), \qquad l \in \{1, \dots, k\}.$$

This means that the last row of $C(\xi_1, \dots, \xi_k)((j+1)h)$ is of the form

$$\begin{bmatrix} -a_0(jh)\xi_1(jh) + *_1 & -a_0(jh)\xi_2(jh) + *_2 & \cdots & -a_0(jh)\xi_k(jh) + *_k \end{bmatrix},$$

where $*_l$ means a linear combination of

$$\xi_l((j+1)h), \dots, \xi_l((j+k-1)h), \qquad l \in \{1, \dots, k\}.$$

Thus the last row of $C(\xi_1, \dots, \xi_k)((j+1)h)$ is

$$\begin{bmatrix} -a_0(jh)\xi_1(jh) & -a_0(jh)\xi_2(jh) & \cdots & -a_0(jh)\xi_k(jh) \end{bmatrix},$$

plus a linear combination of the first $k-1$ rows. By properties of determinants and elementary row operations, this gives

$$C(\xi_1, \dots, \xi_k)((j+1)h) = \det \begin{bmatrix} \xi_1((j+1)h) & \xi_2((j+1)h) & \cdots & \xi_k((j+1)h) \\ \xi_1((j+2)h) & \xi_2((j+2)h) & \cdots & \xi_k((j+2)h) \\ \vdots & \vdots & \ddots & \vdots \\ -a_0(jh)\xi_1(jh) & -a_0(jh)\xi_2(jh) & \cdots & -a_0(jh)\xi_k(jh) \end{bmatrix}$$

cf. Exercise I-5.3.1. Again using the properties of determinants and row operations as in Exercise I-5.3.1, we have

$$C(\xi_1, \dots, \xi_k)((j+1)h) = -a_0(jh) \det \begin{bmatrix} \xi_1((j+1)h) & \xi_2((j+1)h) & \cdots & \xi_k((j+1)h) \\ \xi_1((j+2)h) & \xi_2((j+2)h) & \cdots & \xi_k((j+2)h) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1(jh) & \xi_2(jh) & \cdots & \xi_k(jh) \end{bmatrix}.$$

By $k-1$ row permutations, we can move the last row of the matrix on the right to the first row. By one final use of the properties of determinants and row operations as in Exercise I-5.3.1, we have

$$C(\xi_1,\dots,\xi_k)((j+1)h) = (-1)^k a_0(jh) C(\xi_1,\dots,\xi_k)(jh).$$

We can now apply Example 4.6.5 to get the result. ∎

As with its Wronskian brother, one of the sort of peculiar features of the Casoratian is that it can be used to actually write down a difference equation.

4.6.11 Proposition (A Casoratian representation of a difference equation) *Consider the scalar linear homogeneous ordinary differential equation* F *with right-hand side (4.29). Let* $\{\xi_1,\dots,\xi_k\}$ *be a fundamental set of solutions for* F *and let* $t_0 \in \mathbb{T}_F$. *If* F *is invertible then, for* $\xi \in \mathbb{R}^{\mathbb{T}_{\geq t_0}}$ *and* $t \in \mathbb{T}_{F,\geq t_0}$,

$$\xi(t+kh) + a_{k-1}(t)\xi(t+(k-1)h) + \cdots + a_1(t)\xi(t+h) + a_0\xi(t) = \frac{C(\xi_1,\dots,\xi_k,\xi)(t)}{C(\xi_1,\dots,\xi_k)(t)}.$$

In particular,

$$\mathrm{Sol}_{t_0}(F) = \left\{ \xi \in \mathbb{R}^{\mathbb{T}_{\geq t_0}} \;\middle|\; \frac{C(\xi_1,\dots,\xi_k,\xi)(t)}{C(\xi_1,\dots,\xi_k)(t)} = 0,\ t \in \mathbb{T}_{F,\geq t_0} \right\}.$$

Proof First of all, note by Proposition 4.6.9 that $C(\xi_1,\dots,\xi_k)(t)$ is never zero, so this is valid to appear in denominators, as in the statement of the proposition.

We shall prove the last assertion first. First suppose that $\xi \in \mathrm{Sol}_{t_0}(F)$, then

$$\xi = c_1\xi_1 + \cdots + c_k\xi_k$$

for some (unique) constants $c_1,\dots,c_k \in \mathbb{R}$. Therefore, the functions $\{\xi,\xi_1,\dots,\xi_k\}$ are linearly dependent, cf.

$$-c_1\xi_1 - \cdots - c_k\xi_k + 1\xi = 0.$$

Therefore,

$$\begin{bmatrix} \xi_1(t) & \xi_2(t) & \cdots & \xi_k(t) & \xi(t) \\ \xi_1(t+h) & \xi_2(t+h) & \cdots & \xi_k(t+h) & \xi(t+h) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \xi_1(t+(k-1)h) & \xi_2(t+(k-1)h) & \cdots & \xi_k(t+(k-1)h) & \xi(t+(k-1)h) \\ \xi_1(t+kh) & \xi_2(t+kh) & \cdots & \xi_k(t+kh) & \xi(t+kh) \end{bmatrix} \begin{bmatrix} -c_1 \\ -c_2 \\ \vdots \\ -c_k \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

for all $t \in \mathbb{T}_{F,\geq t_0}$. From this we immediately conclude that $C(\xi_1,\dots,\xi_k,\xi)(t) = 0$ for all $t \in \mathbb{T}_{F,\geq t_0}$, and so

$$\xi \in \left\{ \tilde{\xi} \in \mathbb{R}^{\mathbb{T}_{\geq t_0}} \;\middle|\; \frac{C(\xi_1,\dots,\xi_k,\tilde{\xi})(t)}{C(\xi_1,\dots,\xi_k)(t)} = 0,\ t \in \mathbb{T}_{F,\geq t_0} \right\}.$$

Now note that, if we expand the determinant $\mathrm{C}(\xi_1, \dots, \xi_k, \xi)$ about the last column, we get an expression of the form

$$\begin{aligned}&\mathrm{C}(\xi_1, \dots, \xi_k, \xi)(t)\\ &\quad = \mathrm{C}(\xi_1, \dots, \xi_k)(t)\xi(t + kh) + b_{k-1}(t)\xi(t + (k-1)h) + \cdots + b_1(t)\xi(t + h) + b_0(t)\xi(t)\end{aligned}$$

for some functions $b_0, b_1, \dots, b_{k-1} \colon \mathbb{T} \to \mathbb{R}$. By Proposition 4.6.9 it follows that

$$\left\{\xi \in \mathbb{R}^{\mathbb{T}_{\geq t_0}} \;\middle|\; \frac{\mathrm{C}(\xi_1, \dots, \xi_k, \xi)(t)}{\mathrm{C}(\xi_1, \dots, \xi_k)(t)} = 0,\ t \in \mathbb{T}_{\geq t_0}\right\}$$

is the set of solutions to a kth-order scalar linear homogeneous ordinary difference equation. Moreover, since we clearly have $\mathrm{C}(\xi_1, \dots, \xi_k, \xi_j) = 0$ for every $j \in \{1, \dots, k\}$, (it is the determinant of a $(k+1) \times (k+1)$ matrix with two equal columns), it follows that $\{\xi_1, \dots, \xi_k\}$ is a fundamental set of solutions for this differential equation. Thus we have shown that

$$\mathrm{Sol}(F) = \left\{\xi \in \mathbb{R}^{\mathbb{T}_{\geq t_0}} \;\middle|\; \frac{\mathrm{C}(\xi_1, \dots, \xi_k, \xi)(t)}{\mathrm{C}(\xi_1, \dots, \xi_k)(t)} = 0,\ t \in \mathbb{T}_{F, \geq t_0}\right\}.$$

To prove the first assertion, we shall show that the set of solutions for a kth-order scalar linear homogeneous ordinary difference equation uniquely determines its coefficients. That is, we show that if two such equations F and G with right-hand sides

$$\begin{aligned}\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) &= -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x,\\ \widehat{G}(t, x, x^{(1)}, \dots, x^{(k-1)}) &= -b_{k-1}(t)x^{(k-1)} - \cdots - b_1(t)x^{(1)} - b_0(t)x\end{aligned}$$

satisfy $\mathrm{Sol}_{t_0}(F) = \mathrm{Sol}_{t_0}(G)$, then $a_j(t) = b_j(t)$, $t \in \mathbb{T}_{F, \geq t_0}$, $j \in \{0, 1, \dots, k-1\}$. Let us consider initial conditions

$$\xi(t) = c_0,\ \xi(t + h) = c_1, \dots, \xi(t + (k-1)h) = c_{k-1}.$$

The equality of the sets of solutions implies that

$$\xi(t + kh) + a_{k-1}(t)c_{k-1} + \cdots + a_1(t)c_1 + a_0(t)c_0 = \xi(t + kh) + b_{k-1}(t)c_{k-1} + \cdots + b_1(t)c_1 + b_0(t)c_0.$$

This immediately gives

$$(a_{k-1}(t) - b_{k-1}(t))c_{k-1} + \cdots + (a_1(t) - b_1(t))c_1 + (a_0(t) - b_0(t))c_0 = 0$$

for every $(c_0, c_1, \dots, c_{k-1}) \in \mathbb{R}^k$. This gives $a_j(t) = b_j(t)$, $t \in \mathbb{T}_{F, \geq t_0}$, $j \in \{0, 1, \dots, k-1\}$, as claimed. ∎

4.6.2 Equations with constant coefficients

Having said about as much as one can say, in general, about the scalar homogeneous linear ordinary difference equations with time-varying coefficients, we

now turn to the case of constant coefficient scalar linear homogeneous ordinary difference equations. If

$$F\colon \mathbb{T} \times \mathbb{R} \oplus \mathrm{L}^{\le k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$

is such an equation, then its right-hand side must be given by

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x \tag{4.33}$$

for $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$. Thus a solution $t \mapsto \xi(t)$ satisfies the equation

$$\xi(t + kh) + a_{k-1}\xi(t + (k-1)h) + \cdots + a_1\xi(t + h) + a_0\xi(t) = 0. \tag{4.34}$$

These equations are, of course, a special case of the equations considered in Section 4.6.1, and so all statements made about the general case of time-varying coefficients hold in the special case of constant coefficients. In particular, Proposition 4.6.1 and Theorem 4.6.3 hold for equations of the form (4.34). However, for these constant coefficient equations, it is possible to explicitly describe the character of the solutions, and this is what we undertake to do.

The trick, motivated to some extent by our approach to scalar linear ordinary differential equations, is to *assume* a solution of the form $\xi(jh) = ar^j$ for $a, r \in \mathbb{R}$, and see what happens. First of all, we note that

$$\xi((j + l)h) = ar^{j+l} = ar^l r^j.$$

Thus, a direct substitution into the equation (4.34) shows that, with ξ in this assumed form,

$$ar^{j+k} + a_{k-1}(ar^{j+k-1}) + \cdots + a_1(ar^{j+1}) + a_0(ar^j) = ar^j(r^k + a_{k-1}r^{k-1} + \cdots + a_1 r + a_0) = 0.$$

Since we are looking for nontrivial solutions, we suppose that $a \neq 0$, in which case $\xi(jh) = ar^j$ is a solution for F if and only if

$$r^k + a_{k-1}r^{k-1} + \cdots + a_1 r + a_0 = 0.$$

With this as backdrop, we make the following definition.

4.6.12 Definition (Characteristic polynomial of a scalar linear homogeneous difference equation with constant coefficients) Consider the linear homogeneous ordinary difference equation F with constant coefficients and with right-hand side (4.33). The ***characteristic polynomial*** of F is

$$P_F = X^k + a_{k-1}X^{k-1} + \cdots + a_1 X + a_0 \in \mathbb{R}[X]. \qquad \bullet$$

Note that, if r is a root of the characteristic polynomial, the corresponding solution for F is $t \mapsto r^{t/h}$. If $r = 0$, then the corresponding solution is zero, a possibility that does not arise for constant coefficient ordinary differential equations.

Now we systematically develop the methodology for solving scalar linear homogeneous ordinary differential equations with constant coefficients.

4.6.2.1 Complexification of scalar linear ordinary difference equations It turns out that to solve constant coefficient linear ordinary difference equations, one needs to work with complex numbers. To do this systematically, we introduce the notion of "complexification," by which a real equation is converted into a complex one. This is rather elementary in this setting, but will be less elementary in Section 5.6.2. Thus it will do not harm, and maybe do some good, to treat this systematically here.

First let us understand the notation for forward differences of $\mathbb{C}$-valued functions of a single discrete real variable, i.e., functions of discrete time. Let $\mathbb{T} \subseteq \mathbb{Z}(h)$ be a discrete time-domain and suppose that we have a mapping $\zeta\colon \mathbb{T} \to \mathbb{C}$. If we write ζ as a sum of its real and imaginary parts, $\zeta(t) = \xi(t) + \mathrm{i}\eta(t)$, then we have

$$\zeta(t + jh) = \xi(t + jh) + \mathrm{i}\eta(t + jh).$$

Thus forward differences of order j are just $\mathbb{C}$-valued functions of t. Thus we can follow the same line of reasoning as Remark 3.1.5 and make the identification $\mathrm{L}^k_{\mathrm{sym}}(\mathbb{R};\mathbb{C}) \simeq \mathbb{C}$.

Here is the basic and quite elementary construction.

4.6.13 Definition (Complexification of scalar linear ordinary difference equation) Consider the scalar linear homogeneous ordinary difference equation F with constant coefficients and with right-hand side (4.33). The ***complexification*** of F is the mapping

$$\begin{aligned} F^{\mathbb{C}}\colon \mathbb{T} \times \mathbb{C} \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R};\mathbb{C}) &\to \mathbb{C} \\ (t, z, z^{(1)}, \dots, z^{(k)}) &\mapsto z^{(k)} + a_{k-1}z^{(k-1)} + \cdots + a_1 z^{(1)} + a_0 z. \end{aligned}$$

A ***solution*** for $F^{\mathbb{C}}$ from t_0 is $\zeta \in \mathbb{C}^{\mathbb{T}_{\geq t_0}}$ that satisfies

$$\zeta(t + kh) + a_{k-1}\zeta(t + (k-1)h) + \cdots + a_1\zeta(t + h) + a_0\zeta(t) = 0.$$

By $\mathrm{Sol}_{t_0}(F^{\mathbb{C}})$ we denote the set of solutions for $F^{\mathbb{C}}$ from t_0. If F is invertible, then we denote by $\mathrm{Sol}(F^{\mathbb{C}})$ the set of solutions for $F^{\mathbb{C}}$. •

Everything we said in Section 4.6.1 about scalar linear homogeneous ordinary difference equations holds in the case of the complex differential equation $F^{\mathbb{C}}$, even when the coefficients are not constant. In particular, Proposition 4.6.1 and Theorem 4.6.3 hold in this case to give us the basic attributes of the complex differential equation, merely by replacing the appropriate occurrences of the symbol "$\mathbb{R}$" with the symbol "$\mathbb{C}$." In particular, $\mathrm{Sol}_{t_0}(F^{\mathbb{C}})$ and $\mathrm{Sol}(F^{\mathbb{C}})$ are k-dimensional $\mathbb{C}$-vector spaces if F has order k.

An essential result for returning to "reality" after complexification is the following simple result.

4.6.14 Lemma (Real and imaginary parts of complex solutions are solutions) *Consider the linear homogeneous ordinary difference equation* F *with constant coefficients, with right-hand side (4.33) and with complexification* $\mathrm{F}^{\mathbb{C}}$. *Let* $\mathrm{t}_0 \in \mathbb{T}_{\mathrm{F}}$. *If* $\zeta\colon \mathbb{T}_{\geq \mathrm{t}_0} \to \mathbb{C}$ *is a solution for* $\mathrm{F}^{\mathbb{C}}$, *then* $\mathrm{Re}(\zeta)$ *and* $\mathrm{Im}(\zeta)$ *are solutions for* F.

Proof This is elementary, rather like the proof of Lemma 4.2.14. ■

4.6.2.2 Difference operator calculus

We introduce a simple object that will be used to say a few simple things about our constant coefficient ordinary difference equations.

4.6.15 Definition (Scalar forward difference operator with constant coefficients) Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let $\mathbb{T} \subseteq \mathbb{Z}(h)$ be a discrete time-domain, and let $k \in \mathbb{Z}_{\geq 0}$. Let

$$\mathbb{T}' \subseteq \{t \in \mathbb{T} \mid t + kh \in \mathbb{T}\}.$$

A ***k*th-order scalar difference operator with constant coefficients in** $\mathbb{F}$ is a mapping

$$D\colon \mathbb{F}^{\mathbb{T}} \to \mathbb{F}^{\mathbb{T}'}$$

of the form

$$D(f)(t) = d_k f(t + kh) + d_{k-1} f(t + (k-1)h) + \cdots + d_1 f(t + h) + d_0 f(t),$$

for $d_0, d_1, \ldots, d_k \in \mathbb{F}$ with $d_k \neq 0$. The ***symbol*** for such an object is

$$\sigma(D) = d_k X^k + d_{k-1} X^{k-1} + \cdots + d_1 X + d_0 \in \mathbb{F}[X]. \quad \bullet$$

Note that the codomain of D has to be restricted so that the forward differences make sense. Since we will be composing forward difference operators, we do not demand that D be defined on the largest possible domain where it makes sense. Indeed, forward difference operators of the sort we are talking about have a product given by composition. Thus, if D_1 and D_1 are k_1th- and k_2th-order scalar differential operators with constant coefficients, then we define a $(k_1 + k_2)$th-order scalar differential operator $D_1 D_2$ with constant coefficients by $D_1 D_2(f) = D_1(D_2(f))$. The domain of $D_1 D_2$ must be such that the operator makes sense.

A simplifying observation about scalar forward difference operators with constant coefficients is the following.

4.6.16 Proposition (The symbol of a product is the product of the symbols) *Let* $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, *let* $\mathbb{T} \subseteq \mathbb{Z}(\mathrm{h})$ *be a discrete time-domain, let* $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$. *If* D_1 *and* D_1 *are* k_1*th- and* k_2*th-order scalar difference operators with constant coefficients, then* $\sigma(\mathrm{D}_1\mathrm{D}_2) = \sigma(\mathrm{D}_1)\sigma(\mathrm{D}_2)$.

Proof This follows rather in the manner of Proposition 4.2.16. ■

4.6.17 Corollary (The product for forward difference operators is commutative) *Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let $\mathbb{T} \subseteq \mathbb{Z}(\mathrm{h})$ be a discrete time-domain, let $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$. If D_1 and D_1 are k_1th- and k_2th-order scalar differential operators with constant coefficients, then $\mathrm{D}_1\mathrm{D}_2 = \mathrm{D}_2\mathrm{D}_1$.*

Proof The follows as does Corollary 4.2.17. ■

4.6.2.3 Bases of solutions Now we construct a family of solutions for a scalar linear homogeneous ordinary difference equation. We do this via a procedure.

4.6.18 Procedure (Basis of solutions for scalar linear homogeneous ordinary difference equations with constant coefficients) Given a scalar linear homogeneous ordinary differential equation

$$F\colon \mathbb{T} \times \mathbb{R} \oplus \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}$$

with discrete time-domain $\mathbb{T} \subseteq \mathbb{Z}(h)$ and right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x,$$

and $t_0 \in \mathbb{T}_F$, do the following.

1. Let $F^{\mathbb{C}}$ be the complexification of F,
2. Consider the kth-order scalar differential operator D_F with constant coefficients in $\mathbb{C}$ defined by
$$\sigma(D_{F^{\mathbb{C}}}) = X^k + a_{k-1}X^k + \cdots + a_1 X + a_0.$$
3. Let $r_1, \dots, r_s$ be the distinct roots of $\sigma(D_F)$ and let $m(r_j)$, $j \in \{1, \dots, s\}$, be the multiplicity of the root r_j. Thus
$$\sigma(D_{F^{\mathbb{C}}}) = (X - r_1)^{m(r_1)} \cdots (X - r_s)^{m(r_s)}.$$
4. Fix $j \in \{1, \dots, s\}$ and consider the following cases.
   (a) $r_j = 0$: Define functions $\xi_{0,l}\colon \mathbb{T}_{\geq t_0} \to \mathbb{R}$, $l \in \{1, \dots, m(0)\}$ by
   $$\xi_{0,l}(t) = \begin{cases} 1, & t = t_0 + (l-1)h, \\ 0, & \text{otherwise.} \end{cases}$$
   (b) $r_j \in \mathbb{R} \setminus \{0\}$: Define functions $\xi_{r_j,l}\colon \mathbb{T} \to \mathbb{R}$, $l \in \{1, \dots, m(r_j)\}$, by
   $$\xi_{r_j,l}(t) = t^{l-1} r_j^{t/h}, \qquad l \in \{1, \dots, m(r_j)\}.$$
   (c) $r_j \in \mathbb{C} \setminus \mathbb{R}$: Note that, since r_j is complex and not real, $\bar{r}_j$ is also a root of $\sigma(D_{F^{\mathbb{C}}})$. We will work only with one of these roots, so we write $r_j = \rho_j \mathrm{e}^{\mathrm{i}\theta_j}$ with $\rho_j \in \mathbb{R}_{>0}$ and $\theta_j \in (0, \pi)$. Define functions $\mu_{r_j,l}, \nu_{r_j,l}\colon \mathbb{T} \to \mathbb{R}$ by
   $$\mu_{r_j,l}(t) = t^l \rho_j^{t/h} \cos(\theta_j \tfrac{t}{h}), \quad \nu_{r_j,l}(t) = t^l \rho_j^{t/h} \sin(\theta_j \tfrac{t}{h}), \qquad l \in \{0, 1, \dots, m(r_j) - 1\}.$$
5. Note that the result of the above steps is k functions. We will show that these functions form a basis for $\mathrm{Sol}(F)$. •

4.6.19 Theorem (Basis of solutions for scalar linear homogeneous ordinary difference equations with constant coefficients) *Given a scalar linear homogeneous ordinary difference equation with constant coefficients*

$$\mathrm{F}\colon \mathbb{T} \times \mathbb{R} \oplus \mathrm{L}^{\le k-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) \to \mathbb{R}$$

with discrete time-domain $\mathbb{T} \subseteq \mathbb{Z}(\mathrm{h})$ *and with right-hand side*

$$\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}, \mathrm{x}^{(1)}, \dots, \mathrm{x}^{(k-1)}) = -\mathrm{a}_{k-1}\mathrm{x}^{(k-1)} - \cdots - \mathrm{a}_1\mathrm{x}^{(1)} - \mathrm{a}_0\mathrm{x},$$

define k *functions as in Procedure 4.6.18. If* $\mathrm{t}_0 \in \mathbb{T}_{\mathrm{F}}$*, then these functions, restricted to* $\mathbb{T}_{\ge \mathrm{t}_0}$*, form a basis for* $\mathrm{Sol}_{\mathrm{t}_0}(\mathrm{F})$*. If* F *is invertible, then the functions from parts 4(b) and 4(c) of Procedure 4.6.18 form a basis for* Sol(F)*.*

Proof First we show that each of the functions defined in Procedure 4.6.18 is a solution for F.

First we consider the functions $\xi_{0,l}$, $l \in \{1, \dots, m(0)\}$, associated to a zero root of the characteristic polynomial. Since the zero root has multiplicity $m(0)$, solutions to the difference equation we are considering satisfy

$$\xi(t + kh) + a_{k-1}\xi(t + (k-1)h) + \cdots + a_{m(0)}\xi(t + m(0)h) = 0. \tag{4.35}$$

Therefore, for any initial conditions of the form

$$\xi(t_0) = x_0, \dots, \xi(t_0 + (m(0)-1)h) = x_{m(0)-1},\ \xi(t_0 + m(0)h) = \cdots = \xi(t_0 + (k-1)h) = 0,$$

all solutions will satisfy $\xi(t_0 + jh) = 0$ for $j \ge m(0)$. In particular, $\xi_{0,1}, \dots, \xi_{0,m(0)}$ are solutions.

Next we consider the functions $\xi_{r_j,l}(t) = t^l r_j^{t/h}$, $l \in \{0, 1, \dots, m(r_j) - 1\}$, associated with a real root r_j of the characteristic polynomial for F. Since

$$\sigma(D_{F^{\mathbb{C}}}) = (X - r_1)^{m(r_1)} \cdots (X - r_s)^{m(r_s)},$$

by Corollary 4.6.17 we can write

$$\sigma(D_{F^{\mathbb{C}}}) = P(X - r_j)^{m(r_j)}$$

for some $P \in \mathbb{C}[X]$. Let us denote the forward difference operator

$$D_r(f)(t) = f(t + h) - rf(t).$$

It suffices to show that, for $r \in \mathbb{R}$ and for $m, l \in \mathbb{Z}_{\ge 0}$ with $m \in \mathbb{Z}_{>0}$ and $l < m$, we have

$$D_r^m(P(t)r^{t/h}) = 0, \tag{4.36}$$

where P is any polynomial function of degree $l \in \{0, 1, \dots, m-1\}$. To prove (4.36), we first prove a simple lemma.

1 Lemma *Let* $\mathrm{m} \in \mathbb{Z}_{>0}$ *and* $\mathrm{r} \in \mathbb{C}$. *If* $\xi \in \mathbb{C}^{\mathbb{T}}$ *then*

$$\mathrm{D}_{\mathrm{r}}^{\mathrm{m}}(\xi(\mathrm{t})\mathrm{r}^{\mathrm{t}/\mathrm{h}}) = \mathrm{h}^{\mathrm{m}}\mathrm{r}^{\mathrm{t}/\mathrm{h}+\mathrm{m}}\Delta^{\mathrm{m},+}\xi(\mathrm{t}).$$

Proof We prove this by induction on m. For $m = 1$ we have

$$D_r(\xi(t)r^{t/h}) = \xi(t+h)r^{t/h+1} - r\xi(t)r^{t/h} = hr^{t/h+1}\Delta^{1,+}(\xi)(t),$$

giving the lemma when $m = 1$. Now suppose that the lemma holds when $m = k$. Then

$$\begin{aligned} D_r^{k+1}(\xi(t)r^{t/h}) &= D_rD_r^k(\xi(t)t^{r/h}) = D_r(h^kr^{t/h+k}\Delta^{k,+}\xi(t)) \\ &= h^kr^{t/h+k+1}\Delta^{k,+}\xi(t+h) - rh^kr^{t/h+k}\Delta^{k,+}\xi(t) \\ &= h^kr^{t/h+k+1}(\Delta^{k,+}\xi(t+h) - \Delta^{k,+}\xi(t)) \\ &= h^{k+1}kr^{t/h+k+1}\Delta^{k+1,+}\xi(t), \end{aligned}$$

as desired. ▼

Now, if P is a polynomial function of degree $l \in \{0, 1, \dots, m\}$, by the Lemma 1 we have

$$D_r^m(P(t)r^{t/h}) = h^mr^{t/h+m}\Delta^{m,+}P(t).$$

By Exercise 3.3.4(d), $\Delta^{1,+}P(t)$ is a polynomial of degree $l-1$, just as we have for derivatives. Therefore, $\Delta^{m,+}P(t) = 0$. Thus shows that the functions $\xi_{r_j,l}(t) = t^l r_j^{t/h}$, $l \in \{0, 1, \dots, m(r_j) - 1\}$, are solutions for F.

Next we consider the functions

$$\mu_{r_j,l} = t^l\rho_j^{t/h}\cos(\theta_j\tfrac{t}{h}), \quad \nu_{r_j,l} = t^l\rho_j^{t/h}\sin(\theta_j\tfrac{t}{h}), \qquad l \in \{0, 1, \dots, m(r_j) - 1\},$$

corresponding to a complex root $r_j = \rho_j e^{i\theta_j}$, $\rho_j > 0$, $\theta_j \in (0, \pi)$, of the characteristic polynomial of F. In this case, we argue, exactly as in the case of a real root above, that the $\mathbb{C}$-valued functions $\zeta_{r_j,l}(t) = t^l r_j^{t/h}$, $l \in \{0, 1, \dots, m(r_j) - 1\}$, are solutions for $F^{\mathbb{C}}$. Then, by Lemma 4.6.14, we have that

$$\begin{aligned} \mu_{r_j,l}(t) &= t^l\rho_j^{t/h}\cos(\theta_j\tfrac{t}{h}) \\ &= \mathrm{Re}(t^l\rho_j^{t/h}(\cos(\theta_j\tfrac{t}{h}) + i\sin(\theta_j\tfrac{t}{h})) \\ &= \mathrm{Re}(t^l\rho_j^{t/h}e^{i\theta_j t/h}) = \mathrm{Re}(\zeta_{r_j,l}(t)) \end{aligned}$$

and, similarly,

$$\nu_{r_j,l} = t^l\rho_j^{t/h}\sin(\theta_j\tfrac{t}{h}) = \mathrm{Im}(\zeta_{r_j,l}(t))$$

are solutions for F for $l \in \{0, 1, \dots, m(r_j) - 1\}$.

Our above arguments show that the functions produced in Procedure 4.6.18 are solutions. Moreover, since Procedure 4.6.18 produces k solutions for F, by Theorem 4.2.3 it suffices to show that these solutions are linearly independent to show that they form a basis for $\mathrm{Sol}(F)$.

To undertake this, let us first dispense with the solutions $\xi_{0,l}$, $l \in \{1, \dots, m(0)\}$, corresponding to the zero eigenvalue. First of all, these solutions are linearly independent, and so yield $m(0)$ linearly independent solutions. Let us show that none of these $m(0)$ solutions are contained in the subspace spanned by the remaining $k - m(0)$ solutions. To see this, let ξ be a solution in the subspace spanned by the remaining $k - m(0)$ solutions. By the equation (4.35) that must be satisfied by solutions, we must have $\xi(t_0 + jh) \neq 0$ for some $j \geq m(0)$, and this precludes ξ from being in the subspace spanned by $\xi_{0,l}$, $l \in \{1, \dots, m(0)\}$. Thus the solutions for the zero root span a subspace complementary to that spanned by the remaining solutions.

Now we must show that the remaining solutions are linearly independent. We achieve this with the aid of the following lemma.

2 Lemma *Let $\mathbb{T} \subseteq \mathbb{R}$ be an interval containing more than one point. Let $\mathrm{r}_1, \dots, \mathrm{r}_\mathrm{s} \in \mathbb{R} \setminus 0$ be distinct and let $\mathrm{P}_1, \dots, \mathrm{P}_\mathrm{s}$ be $\mathbb{C}$-valued polynomial functions on $\mathbb{T}$. If*

$$\mathrm{P}_1(\mathrm{t})\mathrm{r}_1^{\mathrm{t}/\mathrm{h}} + \cdots + \mathrm{P}_\mathrm{s}(\mathrm{t})\mathrm{r}_\mathrm{s}^{\mathrm{t}/\mathrm{h}} = 0, \qquad \mathrm{t} \in \mathbb{T},$$

then $\mathrm{P}_\mathrm{j}(\mathrm{t}) = 0$ for all $\mathrm{j} \in \{1, \dots, \mathrm{s}\}$ and $\mathrm{t} \in \mathbb{T}$.

Proof We prove the lemma by induction on s. For $s = 1$ we have, for $r_1 \in \mathbb{R}$ and a polynomial function P_1,

$$\begin{aligned} & P_1(s)r_1^{t/h} = 0, \qquad t \in \mathbb{T}, \\ \implies \quad & P_1(t) = 0, \qquad t \in \mathbb{T}, \end{aligned}$$

giving the result in this case. Now suppose that the lemma is true for $s = k$ and suppose that

$$P_1(t)r_1^{t/h} + \cdots + P_k(t)r_k^{t/h} + P_{k+1}(t)r_{k+1}^{t/h} = 0, \qquad t \in \mathbb{T},$$

for distinct $r_1, \dots, r_k, r_{k+1} \in \mathbb{R} \setminus \{0\}$ and for polynomial functions $P_1, \dots, P_k, P_{k+1}$. Then

$$P_1(t)\left(\frac{r_1}{r_{k+1}}\right)^{t/h} + \cdots + P_k(t)\left(\frac{r_k}{r_{k+1}}\right)^{t/h} + P_{k+1}(t) = 0, \qquad t \in \mathbb{T}. \tag{4.37}$$

Now let us apply the iterated forward difference operator $\Delta^{m,+}$, using the Leibniz Rule for higher-order forward differences stated as Exercise 3.3.5. After applying $\Delta^{m,+}$ we get

$$P_1^m(t)\left(\frac{r_1}{r_{k+1}}\right)^{t/h} + \cdots + P_k^m(t)\left(\frac{r_k}{r_{k+1}}\right)^{t/h} + \Delta^{m,+}P_{k+1}(t) = 0, \qquad t \in \mathbb{T},$$

where

$$P_j^m(t) = \sum_{l=0}^{m} \frac{1}{h^{m-l}}\left(\left(\frac{r_j}{r_{k+1}}\right)^h - 1\right)^{m-l}\left(\frac{r_j}{r_{k+1}}\right)^{lh}\binom{m}{l}\Delta^{l,+}P_j(t), \tag{4.38}$$

after also making use of Exercise 3.3.4(c). Since $r_j \neq r_{k+1}$, P_j^m is a polynomial function whose degree is the same as the degree of P_j. Now, for m sufficiently large (larger than the degree of P_{k+1}, to be precise), $\Delta^{m,+}P_{k+1} = 0$. With m so chosen, we have

$$P_1^m(t)\left(\frac{r_1}{r_{k+1}}\right)^{t/h} + \cdots + P_k^m(t)\left(\frac{r_k}{r_{k+1}}\right)^{t/h} = 0, \qquad t \in \mathbb{T},$$

By the induction hypothesis, $P_j^m(t) = 0$ for $j \in \{1, \dots, k\}$ and $t \in \mathbb{T}$. Now, in the expression (4.38) for P_j^m, note that the highest polynomial degree term in t in the sum occurs when $l = 0$, and this term is

$$\frac{1}{h^m}\left(\left(\frac{r_j}{r_{k+1}}\right)^h - 1\right)^m P_j(t).$$

For the polynomial P_j^m to vanish, this term in the sum must vanish, i.e., $P_j(t) = 0$ for every $j \in \{1, \dots, k\}$ and $t \in \mathbb{T}$. Finally, (4.37) then gives $P_{k+1}(t) = 0$ for all $t \in \mathbb{T}$, giving the result. ▼

Now we can show that the solutions produced by Procedure 4.6.18 and associated with the nonzero roots are linearly independent. Suppose that there are s_1 distinct nonzero real roots with $\frac{1}{h}$th-powers, $r_1, \dots, r_{s_1}$, and s_2 distinct complex roots with $\frac{1}{h}$th-powers

$$\rho_1 \mathrm{e}^{\mathrm{i}\theta_1}, \dots, \rho_{s_2} \mathrm{e}^{\mathrm{i}\theta_{s_2}},$$

with $\rho_1, \dots, \rho_{s_2} > 0$ and $\theta_1, \dots, \theta_{s_2} \in (0, \pi)$, for the characteristic polynomial of F. Thus $s_1 + 2s_2 = k - m(0)$. Suppose that we have $k - m(0)$ scalars

$$c_{j,l}, \qquad j \in \{1, \dots, s_1\},\ l \in \{0, 1, \dots, m(r_j) - 1\}, \tag{4.39}$$

and

$$a_{j,l},\ b_{j,l}, \qquad j \in \{1, \dots, s_2\},\ l \in \{0, 1, \dots, m(\rho_j) - 1\}, \tag{4.40}$$

satisfying

$$\begin{aligned}
&(c_{1,0} + c_{1,1}t + \cdots + c_{1,m(r_1)-1}t^{m(r_1)-1})r_1^{t/h} + \dots \\
&\qquad + (c_{s_1,0} + c_{s_1,1}t + \cdots + c_{s_1,m(r_{s_1})-1}t^{m(r_{s_1})-1})r_1^{t/h} \\
&\qquad + (a_{1,0} + a_{1,1}t + \cdots + a_{1,m(\rho_1)-1}t^{m(\rho_1)-1})\rho_1^{t/h}\cos(\theta_1 \tfrac{t}{h}) \\
&\qquad + (b_{1,0} + b_{1,1}t + \cdots + b_{1,m(\rho_1)-1}t^{m(\rho_1)-1})\rho_1^{t/h}\sin(\theta_1 \tfrac{t}{h}) + \dots \\
&\qquad + (a_{s_2,0} + a_{s_2,1}t + \cdots + a_{s_2,m(\rho_{s_2})-1}t^{m(\rho_{s_2})-1})\rho_{s_2}^{t/h}\cos(\theta_{s_2} \tfrac{t}{h}) \\
&\qquad + (b_{s_2,0} + b_{s_2,1}t + \cdots + b_{s_2,m(\rho_{s_2})-1}t^{m(\rho_{s_2})-1})\rho_{s_2}^{t/h}\sin(\theta_{s_2} \tfrac{t}{h}) = 0, \qquad t \in \mathbb{T}.
\end{aligned}$$

By Lemma 2, the polynomial functions

$$\begin{aligned}
&c_{1,0} + c_{1,1}t + \cdots + c_{1,m(r_1)-1}t^{m(r_1)-1}, \dots, \\
&\qquad c_{s_1,0} + c_{s_1,1}t + \cdots + c_{s_1,m(r_{s_1})-1}t^{m(r_{s_1})-1}, \\
&\qquad a_{1,0} + a_{1,1}t + \cdots + a_{1,m(\rho_1)-1}t^{m(\rho_1)-1}, \\
&\qquad b_{1,0} + b_{1,1}t + \cdots + b_{1,m(\rho_1)-1}t^{m(\rho_1)-1}, \dots, \\
&\qquad a_{s_2,0} + a_{s_2,1}t + \cdots + a_{s_2,m(\rho_{s_2})-1}t^{m(\rho_{s_2})-1}, \\
&\qquad\qquad b_{s_2,0} + b_{s_2,1}t + \cdots + b_{s_2,m(\rho_{s_2})-1}t^{m(\rho_{s_2})-1}
\end{aligned}$$

must all vanish. But this implies that the scalars (4.39) and (4.40) must all vanish. This gives the desired linear independence. ■

4.6.2.4 Some examples The matter of carrying out Procedure 4.6.18 for difference equations is rather like carrying out the analogous Procedure 4.2.18 for differential equations. Thus what we give here are examples that illustrate some interesting behaviours for difference equations.

4.6.20 Example (First-order system behaviour) We consider the general 1st-order scalar linear homogeneous ordinary difference equation F defined on $\mathbb{T} \subseteq \mathbb{Z}(h)$ with right-hand side

$$\widehat{F}(t, x) = -\rho x$$

for $\rho \in \mathbb{R}$. Solutions $t \mapsto \xi(t)$ satisfy

$$\xi(t + h) + \rho\xi(t) = 0.$$

This is an easy equation to solve. It characteristic polynomial is $P_F = X + \rho$ which has the single real root $r_1 = -\rho$. Thus, by Procedure 4.6.18, *any* solution has the form $\xi(t) = c(-\rho)^{t/h}$. To determine c, we use initial conditions as in Proposition 4.6.1. We take a general initial time t_0 and prescribe $\xi(t_0) = x_0$. Thus

$$\xi(t_0) = c(-\rho)^{t_0/h} \implies c = x_0(-\rho)^{-t_0/h},$$

and so $\xi(t) = x_0(-\rho)^{(t-t_0)/h}$.

Let us think about this solution for a moment, and especially compare it to that obtained in Example 4.2.21 for the corresponding differential equation.

When $\rho \in \mathbb{R}_{<0}$, then the situation bears a strong resemblance to that observed for differential equations. Here, if $-\rho > 1$, then we have exponential growth; this is analogous to the case of $\tau > 0$ in Example 4.2.21. If $-\rho < 1$, then we have exponential decay; this is analogous to the case of $\tau < 0$ in Example 4.2.21. Finally, if $\rho = 1$, then solutions are constant, equal to the initial condition; this is analogous to the case of $\tau = 0$ in Example 4.2.21. The idea one should take away from this is that one should regard $-\rho$ in this case as being analogous to $e^{-1/\tau}$ in the case of Example 4.2.21.

Next we think about the case when $\rho \in \mathbb{R}_{>0}$. This case has no analogue in Example 4.2.21. Indeed, all solutions for the difference equation oscillate, and this oscillatory behaviour is not possible for first-order linear homogeneous scalar differential equations. In this case, if $|\rho| < 1$ then the oscillations decay in amplitude, if $|\rho| > 1$ then the oscillations grow in amplitude, and if $|\rho| = 1$ then the amplitude of the oscillations is constant. For the reason that there is, in general, no solution for the difference equation when $\rho \in \mathbb{R}_{>0}$, one often disregards this possibility.

Finally, we consider the case of $\rho = 0$. This corresponds, by Proposition 4.6.2, to the case when F is not invertible. In this case, the behaviour is simply that all initial conditions give rise to solutions that are identically zero after the initial time. This is behaviour that one does not see with the corresponding differential equation in Example 4.2.21.

In Figure 4.6 we graph $\xi(t)$ as a function of t for a few different ρ's. •

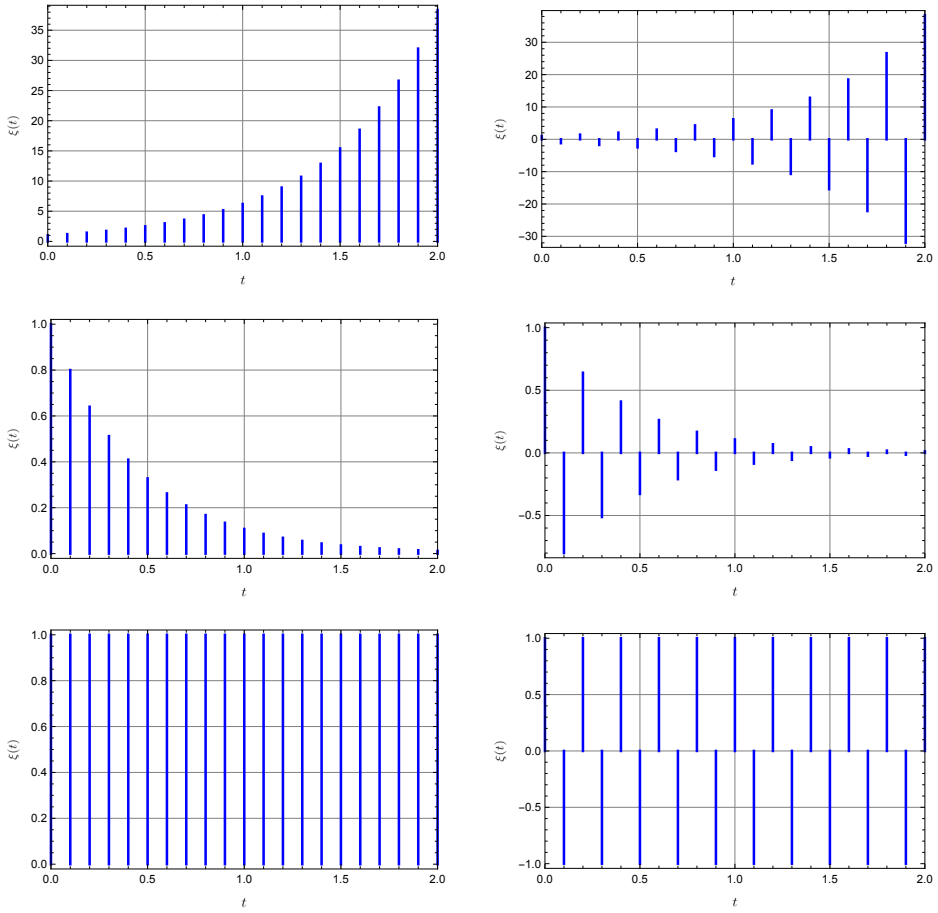


Figure 4.6 Solutions of a first-order scalar linear homogeneous ordinary differential equation with $\xi(0) = 1$ and $h = 0.1$. On the left we have $\rho = -1.2$ (top), $\rho = -0.8$ (middle) and $\rho = -1$ (bottom). We have $\rho = 1.2$ (top), $\rho = 0.8$ (middle) and $\rho = 1$ (bottom).

4.6.21 Example (Second-order system behaviour) Here we consider the second-order scalar linear homogeneous difference equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}) = -\rho^2 x + 2\rho f(\theta_0) x^{(1)}$$

for $\rho \in \mathbb{R}_{>0}$ and for $f \in \{\cos, \cosh, 1\}$, and where we take $\theta \in (0, \pi)$ if $f = \cos$ and $\theta \in \mathbb{R}_{>0}$ if $f = \cosh$. As we shall see, the strange definitions enable simple forms for solutions. Solutions satisfy

$$\xi(t + 2h) - 2\rho f(\theta_0)\xi(t + h) + \rho^2 \xi(t) = 0.$$

By assuming that ρ is nonzero, we ensure that the difference equation is invertible. We have the following cases characterising the forms of solutions:

1. $f = \cos$: $\xi(t) = c_1\rho^{t/h}\cos(\theta_0\frac{t}{h}) + c_2\rho^{t/h}\sin(\theta_0\frac{t}{h})$;
2. $f = \cosh$: $\xi(t) = c_1(\rho e^{\theta_0})^{t/h} + c_2(\rho e^{-\theta_0})^{t/h}$;
3. $f = 1$: $\xi(t) = c_1\rho^{t/h} + c_2 t\rho^{t/h}$.

The initial conditions are

$$\xi(0) = x_0, \quad \xi(h) = x_0^{(1)}.$$

If we solve for the constants using the initial conditions we get

1. $f = \cos\theta$:

$$\begin{aligned} c_1 &= x_0, \\ c_2 &= -x_0\cot(\theta_0) + \frac{x_0^{(1)}}{\rho}\csc(\theta_0); \end{aligned}$$

2. $f = \cosh$:

$$\begin{aligned} c_1 &= \frac{(\rho x_0 - e^{\theta_0}x_0^{(1)}}{\rho(1 - e^{2\theta_0})}, \\ c_2 &= \frac{e^{\theta_0}(e^{\theta_0}\rho x_0 - x_0^{(1)})}{\rho(e^{2\theta_0} - 1)}; \end{aligned}$$

3. $f = 1$:

$$\begin{aligned} c_1 &= x_0, \\ c_2 &= -\frac{-\rho x_0 + x_0^{(1)}}{\rho h}. \end{aligned}$$

Correspondingly to the differential equation in Example 4.2.22, we shall say the equation is ***positively damped*** if either (1) $f = \cos$ and $\rho < 1$ or (2) $f = \cosh$ and $\rho e^{\theta_0} < 1$. The equation is ***negatively damped*** if either (1) $f = \cos$ and $\rho > 1$ or (2) $f = \cosh$ and $\rho e^{\theta_0} > 1$. When either (1) $f = \cos$ and $\rho = 1$ or (2) $f = \cosh$ and $\rho e^{\theta_0} = 1$, the equation is ***undamped***. Let us concentrate on the positively damped case. Here we say that the equation is ***underdamped*** if $f = \cos$ and is ***overdamped*** if $f = \cosh$. The equation is ***critically damped*** when $f = 1$.

In Figures 4.7 and 4.8 we show the plots of the solutions in the various cases. •

Exercises

4.6.1 Let $\mathbb{T} \subseteq \mathbb{Z}(h)$ be a discrete time-domain and let $f_1, \dots, f_k \colon \mathbb{T} \to \mathbb{R}$.

(a) Show that $f_1, \dots, f_k$ are linearly independent in $\mathbb{R}^{\mathbb{T}}$ if and only if there exists $t_1, \dots, t_k \in \mathbb{T}$ such that

$$\det \begin{bmatrix} f_1(t_1) & \cdots & f_k(t_1) \\ \vdots & \ddots & \vdots \\ f_1(t_k) & \cdots & f_k(t_k) \end{bmatrix} \neq 0.$$

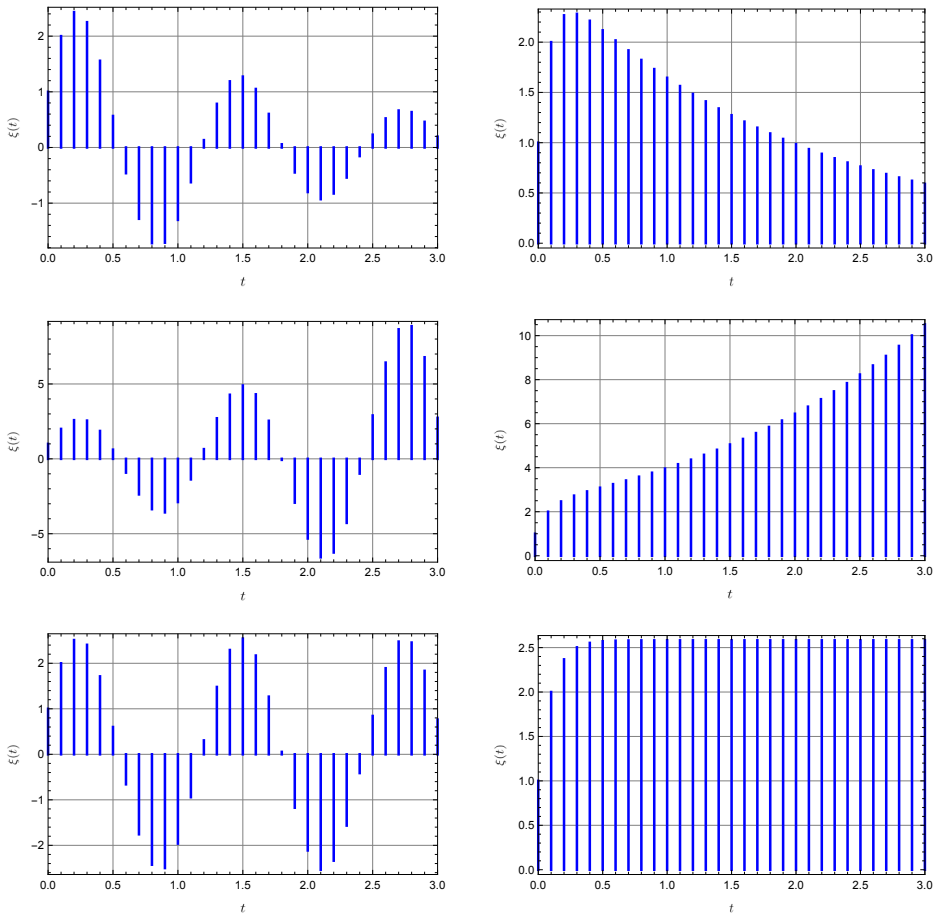


Figure 4.7 Solutions of a second-order scalarlinear homogeneous ordinary differential equation with $\xi(0) = 1$ and $\xi(h) = 2$. In all cases we have $h = 0.1$ and $\theta_0 = 0.5$. On the left we have $f = \cos$ and $\rho = 0.95$ (top), $\rho = 1.05$ (middle), and $\rho = 1$ (bottom). On the right we have $f = \cosh$ and $\rho = 0.95$ (top), $\rho = 1.05$ (middle), and $\rho = 1$ (bottom).

(b) Using your part (a), explain why the Casoratian is generally a poor device for showing the linear independence of an arbitrary collection of elements of $\mathbb{R}^{\mathbb{T}}$ for a discrete time-domain $\mathbb{T}$.

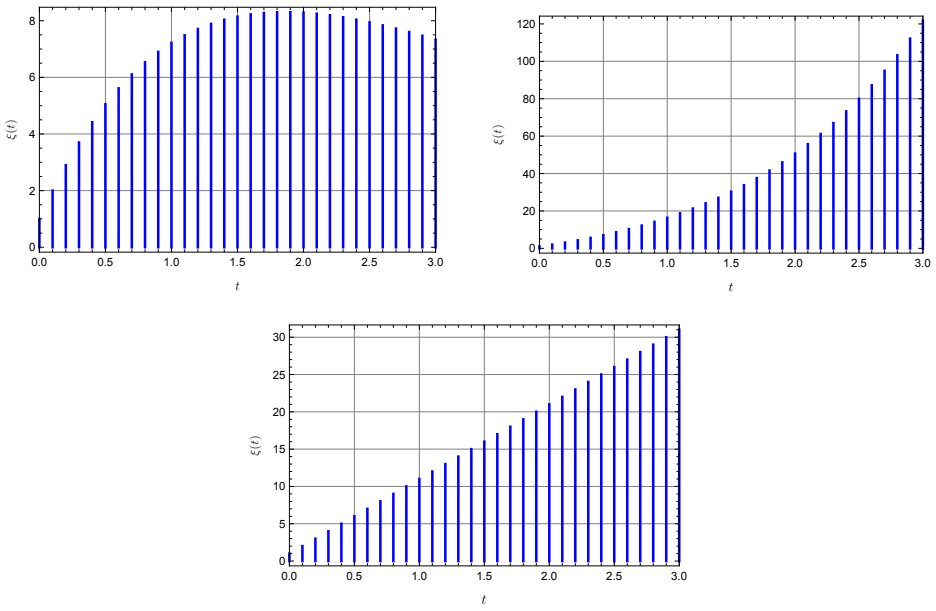


Figure 4.8 Solutions of a second-order scalarlinear homogeneous ordinary differential equation with $\xi(0) = 1$ and $\xi(h) = 2$. In all cases we have $h = 0.1$, $\theta_0 = 0.5$, and $f = 1$. In top we have $\rho = 0.95$ (left), $\rho = 1.05$ (right), and in bottom we have $\rho = 1$ (bottom).

Section 4.7

Scalar linear inhomogeneous ordinary difference equations

In this section we still consider scalar linear ordinary difference equations, but now we consider the inhomogeneous case. We still have the time-domain $\mathbb{T} \subseteq \mathbb{Z}(h)$ and the state space $U = \mathbb{R}$, but now we have a right-hand side of the form

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0 x + b(t) \tag{4.41}$$

for functions $a_0, a_1, a_{k-1}, b \colon \mathbb{T} \to \mathbb{R}$. Thus solutions $t \mapsto \xi(t)$ satisfy

$$\xi(t + kh) + a_{k-1}(t)\xi(t + (k-1)h) + \cdots + a_1(t)\xi(t + h) + a_0(t)\xi(t) = b(t).$$

We shall proceed in this section much as in the preceding section, first saying some things about the general case, and then focussing on the case where F has constant coefficients, as in this case there is more that can be said.

Do I need to read this section? This section contains tools that are standard for anyone claiming to know something about ordinary difference equations. •

4.7.1 Equations with time-varying coefficients

We begin by stating some general properties of general scalar linear inhomogeneous ordinary difference equations.

4.7.1.1 Solutions and their properties First we state the local existence and uniqueness result that one needs to get off the ground for any class of difference equations.

4.7.1 Proposition (Existence and uniqueness of solutions for scalar linear inhomogeneous ordinary difference equations) *Consider the linear inhomogeneous ordinary difference equation* F *with right-hand side equation (4.41). Let*

$$(t_0, x_0, x_0^{(1)}, \dots, x_{k-1}^{(0)}) \in \mathbb{T}_F \times \mathbb{R} \oplus L_{\mathrm{sym}}^{\leq k-1}(\mathbb{R}; \mathbb{R}).$$

Then there exists a unique $\xi \colon \mathbb{T}_{\geq t_0} \to \mathbb{R}$ *that is a solution for* F *and which satisfies*

$$\xi(t_0) = x_0, \xi(t_0 + h) = x_0^{(1)}, \dots, \xi(t_0 + (k-1)h) = x_0^{(k-1)}. \tag{4.42}$$

If F *is invertible, then there exists a unique* $\xi \colon \mathbb{T} \to \mathbb{R}$ *that is a solution for* F *and which satisfies (4.42).*

Proof This is Exercise 4.7.1. ■

The comments following Proposition 4.6.1 concerning the comparison in the homogeneous case between difference equations and differential equations are also valid here.

As in the homogeneous case, we can now talk sensibly about the set of *all* solutions for F. Thus we can define

$$\begin{aligned}\mathrm{Sol}_{t_0}(F) = \Big\{\xi \in \mathbb{R}^{\mathbb{T}_{\geq t_0}} \Big| \\ \xi(t+kh) + a_{k-1}(t)\xi(t+(k-1)h) + \cdots + a_1(t)\xi(t+h) + a_0(t)\xi(t) = b(t),\ t \in \mathbb{T}_{F,\geq t_0}\Big\}.\end{aligned}$$

which is exactly this set of all solutions for F from t_0. In case F is invertible, we can define

$$\begin{aligned}\mathrm{Sol}(F) = \Big\{\xi \in \mathbb{R}^{\mathbb{T}} \Big| \\ \xi(t+kh) + a_{k-1}(t)\xi(t+(k-1)h) + \cdots + a_1(t)\xi(t+h) + a_0(t)\xi(t) = 0,\ t \in \mathbb{T}_F\Big\}.\end{aligned}$$

While $\mathrm{Sol}_{t_0}(F)$ and $\mathrm{Sol}(F)$ were vectors space in the homogeneous case, in the inhomogeneous case this is no longer the case. However, the sets of solutions for the homogeneous case play an important rôle, even in the homogeneous case. To organise this discussion, we let F_{h} be the "homogeneous part" of F. Thus the right-hand side of F_{h} is

$$\widehat{F_{\mathrm{h}}}(t, x, x^{(1)}, \ldots, x^{(k-1)}) = -a_{k-1}(t)x^{(k-1)} - \cdots - a_1(t)x^{(1)} - a_0(t)x.$$

As in Theorem 4.6.3, $\mathrm{Sol}(F_{\mathrm{h}})$ is a $\mathbb{R}$-vector space of dimension k. We can now state the character of $\mathrm{Sol}(F)$.

4.7.2 Theorem (Affine space structure of sets of solutions) *Consider the linear inhomogeneous ordinary difference equation* F *with right-hand side equation (4.41). Let* $\xi_{\mathrm{p}} \in \mathrm{Sol}_{t_0}(\mathrm{F})$. *Then*

$$\mathrm{Sol}_{t_0}(\mathrm{F}) = \{\xi + \xi_{\mathrm{p}} \mid \xi \in \mathrm{Sol}_{t_0}(\mathrm{F_h})\}.$$

Moreover, if F *is additionally invertible and if* $\xi_{\mathrm{p}} \in \mathrm{Sol}(\mathrm{F})$. *Then*

$$\mathrm{Sol}(\mathrm{F}) = \{\xi + \xi_{\mathrm{p}} \mid \xi \in \mathrm{Sol}(\mathrm{F_h})\}.$$

Proof This can be proved, *mutatis mutandis*, as is Theorem 4.3.3. ∎

It is interesting to make some comments on the preceding theorem in the language of basic problems in linear algebra.

4.7.3 Remark (Comparison of Theorem 4.7.2 with systems of linear algebraic equations) The reader should compare here the result of Theorem 4.7.2 with the situation concerning linear algebraic equations of the form $L(u) = v_0$, for vector spaces U and V, a linear map $L \in \mathrm{L}(\mathsf{U}; \mathsf{V})$, and a fixed $v_0 \in \mathsf{V}$. In particular, we can make

reference to Section I-5.4.8. In the setting of scalar linear inhomogeneous ordinary difference equations, we have

$$\begin{aligned} \mathsf{U} &= \mathbb{R}^{\mathbb{T}_{\geq t_0}}, \\ \mathsf{V} &= \mathbb{R}^{\mathbb{T}_{\geq t_0}}, \\ L(f)(t) &= f(t+kh) + a_{k-1}(t)f(t+(k-1)h) + \cdots + a_1(t)f(t+h) + a_0(t)f(t), \\ v_0 &= b. \end{aligned}$$

Then Proposition 4.7.1 tells us that L is surjective, and so $v_0 \in \text{image}(L)$. Thus we are in case (ii) of Proposition I-5.4.48, which exactly the statement of Theorem 4.7.2. Note that L is not injective, since Theorem 4.6.3 tells us that $\dim_{\mathbb{R}}(\ker(L)) = k$. •

Note that Theorem 4.7.2 tells us that, to solve a scalar linear inhomogeneous ordinary difference equation, we must do two things: (1) find *some* solution for the equation; (2) find *all* solutions for the homogeneous part. Then we know our solution will be found amongst the set of sums of these. Generally, both of these things is impossible, in any general way. We do know, however, that Procedure 4.6.18 can be used, in principle, to find all solutions for the homogeneous part. Thus one need only find some solution of the equation in the constant coefficient case. Upon finding such a solution, one calls it a ***particular solution***. *Note that there are many particular solutions.* Indeed, Proposition 4.6.1 tells us that there is one solution for every set of initial conditions. So one should always speak of *a* particular solution, not *the* particular solution.

4.7.1.2 Finding a particular solution using the Casoratian So... how do we find a particular solution? In this section we outline a general (and not very efficient) way of arriving at some such solution, using the Casoratian of Definition 4.6.6. To state the result, suppose that we have a fundamental set of solutions $\{\xi_1, \dots, \xi_k\}$ for F_{h}, where F has right-hand side (4.41), and denote

$$\begin{aligned} &\mathrm{C}_{b,j}(\xi_1, \dots, \xi_k)(t) \\ &= \det \begin{bmatrix} \xi_1(t) & \cdots & \xi_{j-1}(t) & 0 & \xi_{j+1}(t) & \cdots & \xi_k(t) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \xi_1(t+(k-2)h) & \cdots & \xi_{j-1}(t+(k-2)h) & 0 & \xi_{j+1}(t+(k-2)h) & \cdots & \xi_k(t+(k-2)h) \\ \xi_1(t+(k-1)h) & \cdots & \xi_{j-1}(t+(k-1)h) & b(t-h) & \xi_{j+1}(t+(k-1)h) & \cdots & \xi_k(t+(k-1)h) \end{bmatrix}, \end{aligned}$$

for $j \in \{1, \dots, k\}$, i.e., $\mathrm{C}_{b,j}(\xi_1, \dots, \xi_k)(t)$ is the determinant of the matrix used to compute the Casoratian, but with the jth column replaced by $(0, \dots, 0, h^{-1}b(t-h))$.

We then have the following result.

4.7.4 Proposition (A particular solution using Casoratians) *Consider the linear inhomogeneous ordinary difference equation* F *with right-hand side equation (4.41). Let*

$\{\xi_1, \dots, \xi_k\}$ be a fundamental set of solutions for F_h and let $t_0 \in \mathbb{T}_F$. Suppose that F_h is invertible. Then the function $\xi_p \colon \mathbb{T} \to \mathbb{R}$ defined by

$$\xi_p(t) = \sum_{j=1}^{k} \xi_j(t) \sum_{l=1}^{(t-t_0)/h} \frac{C_{b,j}(\xi_1, \dots, \xi_k)(t_0 + lh)}{C(\xi_1, \dots, \xi_k)(t_0 + lh)}, \qquad t \in \mathbb{T}_{\geq t_0},$$

is a particular solution for F from t_0.

Proof Let us define

$$c_j(t) = \sum_{l=1}^{(t-t_0)/h} \frac{C_{b,j}(\xi_1, \dots, \xi_k)(t_0 + lh)}{C(\xi_1, \dots, \xi_k)(t_0 + lh)}, \qquad j \in \{1, \dots, k\},\ t \in \mathbb{T}_{\geq t_0},$$

so that

$$\begin{aligned}
\Delta^{1,+} c_j(t) &= \frac{1}{h}(c_j(t+h) - c_j(t)) \\
&= \frac{1}{h}\left(\sum_{l=1}^{(t+h-t_0)/h} \frac{C_{b,j}(\xi_1, \dots, \xi_k)(t_0 + lh)}{C(\xi_1, \dots, \xi_k)(t_0 + lh)} - \sum_{l=1}^{(t-t_0)/h} \frac{C_{b,j}(\xi_1, \dots, \xi_k)(t_0 + lh)}{C(\xi_1, \dots, \xi_k)(t_0 + lh)} \right) \\
&= \frac{1}{h} \frac{C_{b,j}(\xi_1, \dots, \xi_k)(t+h)}{C(\xi_1, \dots, \xi_k)(t+h)}.
\end{aligned}$$

Note that this is equivalent, by Cramer's Rule for linear systems of algebraic equations (Proposition I-5.3.12), to the set of equations

$$\begin{bmatrix} \xi_1(t+h) & \xi_2(t+h) & \cdots & \xi_k(t+h) \\ \xi_1(t+2h) & \xi_2(t+2h) & \cdots & \xi_k(t+2h) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1(t+kh) & \xi_2(t+kh) & \cdots & \xi_k(t+kh) \end{bmatrix} \begin{bmatrix} h\Delta^{1,+} c_1(t) \\ h\Delta^{1,+} c_2(t) \\ \vdots \\ h\Delta^{1,+} c_k(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ b(t) \end{bmatrix}, \qquad t \in \mathbb{T}_{\geq t_0}. \tag{4.43}$$

Note that the proposition is then that

$$\xi_p(t) = \sum_{j=1}^{k} c_j(t) \xi_j(t), \qquad t \in \mathbb{T}_{\geq t_0},$$

defines a particular solution for F. This we shall prove by direct computation.

We compute

$$\begin{aligned}
\xi_p(t+h) &= \sum_{j=1}^{k} c_j(t+h) \xi_j(t+h) \\
&= \sum_{j=1}^{k} (c_j(t+h) - c_j(t)) \xi_j(t+h) + \sum_{j=1}^{k} c_j(t) \xi_j(t+h) \\
&= h \sum_{j=1}^{k} \Delta^{1,+} c_j(t) \xi_j(t+h) + \sum_{j=1}^{k} c_j(t) \xi_j(t+h) = \sum_{j=1}^{k} c_j(t) \xi_j(t+h)
\end{aligned}$$

for $t \in \mathbb{T}_{\geq t_0}$, using the first of equations (4.43). Repeatedly shifting by h and using successive equations from (4.43), we deduce that

$$\xi_{\mathrm{p}}(t+lh) = \sum_{j=1}^{k} c_j(t)\xi_j(t+lh), \qquad l \in \{0,1,\dots,k-1\},\ t \in \mathbb{T}_{\geq t_0}.$$

We also have, using the last of equations (4.43),

$$\begin{aligned}
\xi_{\mathrm{p}}(t+kh) &= \sum_{j=1}^{k} c_j(t+h)\xi_j(t+kh) \\
&= \sum_{j=1}^{k} (c_j(t+h) - c_j(t))\xi_j(t+kh) + \sum_{j=1}^{k} c_j(t)\xi(t+kh) \\
&= h\sum_{j=1}^{k} \Delta^{1,+}c_j(t)\xi_j(t+kh) + \sum_{j=1}^{k} c_j(t)\xi(t+kh) \\
&= b(t) + \sum_{j=1}^{k} c_j(t)\xi(t+kh).
\end{aligned}$$

Therefore, combining these calculations,

$$\begin{aligned}
&\xi_{\mathrm{p}}(t+kh) + a_{k-1}(t)\xi_{\mathrm{p}}(t+(k-1)h) + \cdots + a_1(t)\xi_{\mathrm{p}}(t+h) + a_0(t)\xi(t) \\
&\quad = \sum_{j=1}^{k} c_j(t)\Big(\xi_j(t+kh) + a_{k-1}(t)\xi_j(t+(k-1)h) + \cdots + a_1(t)\xi_j(t+h) + a_0(t)\xi_j(t)\Big) + b(t) \\
&\quad = b(t),
\end{aligned}$$

using the fact that $\xi_1,\dots,\xi_k$ are solutions for F_{h}. Thus ξ_{p} is indeed a particular solution. ■

Let us illustrate the procedure of the preceding result with an example.

4.7.5 Example (First-order scalar linear inhomogeneous ordinary difference equations) We consider here the first-order equation F with right-hand side

$$\widehat{F}(t,x) = -a(t)x + b(t)$$

for $a, b \in \mathbb{R}^{\mathbb{T}}$. We have seen in Example 4.6.5 that a fundamental set of solutions for F from t_0 is given by $\{\xi_1(t)\}$, with

$$\xi_1(j_0h) = 1, \quad \xi_1(jh) = (-1)^{j-j_0}\prod_{l=j_0}^{j-1} a(lh), \qquad j \in \mathbb{T}_{>j_0h}.$$

Therefore,

$$\mathrm{C}(\xi_1)(t) = \det\Big[\xi_1(t)\Big] = \xi_1(t), \quad \mathrm{C}(\xi_1)_{b,1} = \det\Big[b(t-h)\Big] = b(t-h).$$

Thus $\xi_{\mathrm{p}}(j_0h) = 0$ (by convention) and, for $j > j_0$,

$$\begin{aligned}\xi_{\mathrm{p}}(jh) &= \xi_1(jh)\sum_{l=j_0}^{j-1}\frac{b(lh)}{\xi_1((l+1)h)}\\ &= \left((-1)^{j-j_0}\prod_{l=j_0}^{j-1}a(lh)\right)\sum_{l=j_0}^{j-1}\frac{b(lh)}{(-1)^{l+1-j_0}\prod_{r=j_0}^{l}a(rh)}\end{aligned}$$

defines a particular solution for F. Thus, as in Theorem 4.7.2, *any* solution for F has the form

$$\xi(t) = C(-1)^{j-j_0}\prod_{l=j_0}^{j-1}a(lh) + \left((-1)^{j-j_0}\prod_{l=j_0}^{j-1}a(lh)\right)\sum_{l=j_0}^{j-1}\frac{b(lh)}{(-1)^{l+1-j_0}\prod_{r=j_0}^{l}a(rh)}$$

for some $C \in \mathbb{R}$. In we apply an initial condition $\xi(t_0) = x_0$, then we see that $C = x_0$. Therefore, finally, we have the solution

$$\xi(t) = x_0(-1)^{j-j_0}\prod_{l=j_0}^{j-1}a(lh) + \left((-1)^{j-j_0}\prod_{l=j_0}^{j-1}a(lh)\right)\sum_{l=j_0}^{j-1}\frac{b(lh)}{(-1)^{l+1-j_0}\prod_{r=j_0}^{l}a(rh)}$$

to the initial value problem

$$\xi(t+h) + a(t)\xi(t) = b(t), \quad \xi(t_0) = x_0.$$

Because we have expressed the solution of a differential equation as a sum, we declare victory![6] •

4.7.1.3 The discrete-time Green's function

In this section we describe another means of determining a particular solution. In this case, what we arrive at is a description of a particular solution that allows for the inhomogeneous term "b" to be plugged into an integral. We shall see a close variant of this in Section 5.7 when we discuss linear inhomogeneous *systems* of equations.

The result is the following.

4.7.6 Theorem (Existence of, and properties of, the discrete-time Green's function) *Consider the linear homogeneous ordinary difference equation* F *with right-hand side equation (4.29). Then there exists*

$$\mathrm{G}_{\mathrm{F}} \colon \mathbb{T} \times \mathbb{T} \to \mathbb{R}$$

with the following properties:

[6] Because victories are few and far between in the business of solving difference equations.

(i) *for* $\tau \in \mathbb{T}$, *we have* $G_F(t, \tau) = 0$ *for* $t < \tau$ *and*

$$G_F(\tau + (k-1)h, \tau) = 1,\ G_F(\tau + lh, \tau) = 0, \qquad l \in \{0, 1, \dots, k-2\};$$

(ii) *for* $t \in \mathbb{T}_{\geq \tau}$ *we have*

$$G_F(t+kh, \tau) + a_{k-1}(t)G_F(t+(k-1)h, \tau) + \cdots + a_1(t)G_F(t+h, \tau) + a_0(t)G_F(t, \tau) = 0;$$

(iii) *if* $b \in \mathbb{R}^{\mathbb{T}}$, *if* $t_0 \in \mathbb{T}_F$, *and if* $\xi_{p,b} \colon \mathbb{T}_{\geq t_0} \to \mathbb{R}$ *is given by*

$$\xi_{p,b}(t) = \sum_{l=0}^{(t-t_0-h)/h} G_F(t-h, t_0+lh)b(t_0+lh),$$

then $\xi_{p,b}$ *solves the initial value problem*

$$\xi(t+kh) + a_{k-1}(t)\xi(t+(k-1)h) + \cdots + a_1(t)\xi(t+h) + a_0(t)\xi(t) = b(t),$$
$$\xi(t_0) = \cdots = \xi(t_0 + (k-1)h) = 0.$$

Moreover, there is only one such function satisfying all of the above properties.

Proof For $\tau \in \mathbb{T}$, let $\xi_\tau \colon \mathbb{T} \to \mathbb{R}$ be the solution to the initial value problem

$$\xi(t+kh) + a_{k-1}(t)\xi(t+(k-1)h) + \cdots + a_1(t)\xi(t+h) + a_0(t)\xi(t) = 0,$$
$$\xi(\tau) = \cdots = \xi(\tau + (k-2)h) = 0,\ \xi(\tau + (k-1)h) = 1.$$

Let $\{\xi_1, \dots, \xi_k\}$ be a fundamental set of solutions from τ and write

$$\xi_\tau(t) = \sum_{j=1}^{k} c_j(\tau)\xi_j(t), \qquad t \in \mathbb{T}_{\geq \tau}.$$

The specified initial conditions for ξ_τ can then be written in matrix form as

$$\begin{bmatrix} \xi_1(\tau) & \cdots & \xi_k(\tau) \\ \xi_1(\tau+h) & \cdots & \xi_k(\tau+h) \\ \vdots & \ddots & \vdots \\ \xi_1(\tau+(k-1)h) & \cdots & \xi_k(\tau+(k-1)h) \end{bmatrix} \begin{bmatrix} c_1(\tau) \\ c_2(\tau) \\ \vdots \\ c_k(\tau) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Following the construction preceding the statement of Proposition 4.7.4, denote

$$C_j(\xi_1, \dots, \xi_k)(t)$$
$$= \det \begin{bmatrix} \xi_1(t) & \cdots & \xi_{j-1}(t) & 0 & \xi_{j+1}(t) & \cdots & \xi_k(t) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \xi_1(t+(k-2)h) & \cdots & \xi_{j-1}(t+(k-2)h) & 0 & \xi_{j+1}(t+(k-2)h) & \cdots & \xi_k(t+(k-2)h) \\ \xi_1(t+(k-1)h) & \cdots & \xi_{j-1}(t+(k-1)h) & 1 & \xi_{j+1}(t+(k-1)h) & \cdots & \xi_k(t+(k-1)h) \end{bmatrix},$$

for $j \in \{1, \dots, k\}$. Then

$$c_j(\tau) = \frac{\mathrm{C}_j(\xi_1, \dots, \xi_k)(\tau)}{\mathrm{C}(\xi_1, \dots, \xi_k)(\tau)}.$$

Note that the denominator in the preceding expression is nonzero since $\xi_1, \dots, \xi_k$ are fundamental solutions from τ.

Now define

$$G_F(t, \tau) = \begin{cases} \xi_\tau(t), & t \geq \tau, \\ 0, & t < \tau. \end{cases}.$$

With this definition of G_F, let us check off the assertions of the theorem. The first two assertions follow immediately from the definition, so we only verify part (iii).

Let $b \in \mathbb{R}^{\mathbb{T}}$ and, for $t_0 \in \mathbb{T}_F$, define $\xi_{\mathrm{p},b} \colon \mathbb{T}_{\geq t_0} \to \mathbb{R}$ by

$$\xi_{\mathrm{p},b}(t) = \sum_{l=0}^{(t-t_0-h)/h} G_F(t-h, t_0+lh) b(t_0+lh).$$

For part (iii), we must show that $\xi_{\mathrm{p},b}$ solves the initial value problem in the theorem statement. We can take

$$\xi_{\mathrm{p},b}(t_0) = \cdots = \xi_{\mathrm{p},x_0}(t_0 + (k-1)h) = 0,$$

and so then we need only verify that the difference equation holds. We compute

$$\begin{aligned} \xi_{\mathrm{p},b}(t+h) &= \sum_{l=0}^{(t-t_0)/h} G_F(t, t_0+lh) b(t_0+lh) \\ &= \sum_{l=0}^{(t-t_0-h)/h} G_F(t, t_0+lh) b(t_0+lh) + G_F(t,t) b(t) \\ &= \sum_{l=0}^{(t-t_0-h)/h} G_F(t, t_0+lh) b(t_0+lh). \end{aligned}$$

We can then recursively compute

$$\xi_{\mathrm{p},b}(t+mh) = \sum_{l=0}^{(t-t_0-h)/h} G_F(t+(m-1)h, t_0+lh) b(t_0+lh), \qquad m \in \{0, 1, \dots, k-1\},$$

and

$$\xi_{\mathrm{p},b}(t+kh) = \sum_{l=0}^{(t-t_0-h)/h} G_F(t+(k-1)h, t_0+lh) b(t_0+lh) + G_F(t+(k-1)h, t) b(t).$$

Combining the preceding two formulae and using part (ii), we have, for $t \in \mathbb{T}_{\geq t_0}$,

$$\xi_{\mathrm{p},b}(t+kh) + a_{k-1}(t)\xi_{\mathrm{p},b}(t+(k-1)h) + \cdots + a_1(t)\xi_{\mathrm{p},b}(t+h) + a_0(t)\xi_{\mathrm{p},b}(t) = b(t),$$

giving (iii).

The final uniqueness assertion of the theorem is obtained from the following observations:

1. for $t < \tau$, $t \mapsto G_F(t, \tau)$ is the zero element of $\mathrm{Sol}(F)$;
2. for $t \geq \tau$, $t \mapsto G_F(t, \tau)$ is the unique element of $\mathrm{Sol}_\tau(F)$ with initial conditions

$$G_F(\tau + lh, \tau) = 0, \qquad l \in \{0, 1, \dots, k-2\},$$
$$G_F(\tau + (k-1)h, \tau) = 1.$$

These, combined with Proposition 4.7.1, give the theorem. ■

Of course, we can give a name to the function G_F from the preceding theorem.

4.7.7 Definition (Discrete-time Green's function) Consider the linear homogeneous ordinary difference equation F with right-hand side equation (4.41). The function G_F of Theorem 4.7.6 is the ***discrete-time Green's function*** for F. •

There are a few observations one can make about the discrete-time Green's function.

4.7.8 Remarks (Attributes of the discrete-time Green's function)

1. As we observed in Remark 4.7.3, the mapping

$$L_F \colon \mathbb{R}^{\mathbb{T}_{\geq t_0}} \to \mathbb{R}^{\mathbb{T}_{\geq t_0}}$$
$$\xi \mapsto F_{\mathrm{h}}(t, \xi(t), \xi(t+h), \dots, \xi(t+kh))$$

is surjective, and so, for any $b \in \mathbb{R}^{\mathbb{T}_{\geq t_0}}$, there exists one (indeed, many by Theorem 4.7.2), solution of the difference equation with solutions

$$F_{\mathrm{h}}(t, \xi(t), \xi(t+h), \dots, \xi(t+kh)) = b(t).$$

One can think of the mapping

$$b \mapsto \left(t \mapsto \sum_{l=0}^{(t-t_0-h)/h} G_F(t-h, t_0+lh) b(t_0+lh) \right) \tag{4.44}$$

as prescribing a right-inverse of L_F. Of course, the prescription of a particular right-inverse amounts to a prescription for choosing initial conditions, since initial conditions are what distinguish elements of $\mathrm{Sol}_{t_0}(F)$. We refer the reader to Exercise 4.7.2 for just what initial condition choice is being made by the assignment (4.44).

2. There is also a physical interpretation of the mapping $t \mapsto G_F(t, \tau)$. For $t < \tau$, the solution is zero, until something happens at $t = \tau$. At $t = \tau$, we imagine the system being given an "impulse" input. Unlike in the continuous-time case where one needs distribution theory to make precise the notion of an impulse, in the discrete time case this is elementary as one can give an input that is zero, except at the time of the impulse where it has value 1. The reader can make this precise in Exercise 4.7.3. •

Let us give the simplest possible example to illustrate the use of the discrete-time Green's function.

4.7.9 Example (Discrete-time Green's function for first-order scalar linear ordinary differential equation) We consider the first order equation F with right-hand side

$$\widehat{F}(t,x) = -a(t)x.$$

Let us take $\mathbb{T} \subseteq \mathbb{Z}(h)$ to be the time-domain for the equation. The way one determines the continuous-time Green's function is by first taking $\tau \in \mathbb{T}$ and then solving the initial value problem

$$\xi(t+h) + a(t)\xi(t) = 0, \quad \xi(\tau) = 1,$$

just as prescribed in part (iii) of Theorem 4.7.6. However, in Example 4.6.5 we obtained the solution to this initial value problem as

$$\xi(t) = (-1)^{(t-\tau)/h} \prod_{l=\tau/h}^{(t-\tau-h)/h} a(lh)$$

Then the continuous-time Green's function is given by

$$G_F(t,\tau) = \begin{cases} 0, & t < \tau, \\ (-1)^{(t-\tau)/h} \displaystyle\prod_{l=\tau/h}^{(t-\tau-h)/h} a(lh), & t \geq \tau. \end{cases}$$

Therefore, given $b \in \mathbb{R}^{\mathbb{T}}$, the solution to the initial value problem

$$\xi(t+h) + a(t)\xi(t) = b(t), \quad \xi(t_0) = 0,$$

is given by

$$\xi_{\mathrm{p},b}(t) = \sum_{l=0}^{(t-t_0-h)/h} (-1)^{(t-t_0+lh)/h} \prod_{l=(t_0+lh)/h}^{(t-t_0+lh-h)/h} a((l-1)h)b(t_0+lh).$$

Note that this is, in this first-order case, the same particular solution as in Example 4.7.5 using the Casoratian method of Proposition 4.7.4. This is simply because both solutions satisfy the same initial value problem. To rectify that the solutions are, in fact the same, can be done by a change of summation variable. •

4.7.10 Remark (Discrete-time Green's function for constant coefficient equations and convolution) Suppose that F is a kth-order scalar linear inhomogeneous ordinary differential equation with constant coefficients, and take $\mathbb{T} = [0,\infty)$. As in the statement of Theorem 4.7.6, for each $\tau \in \mathbb{T}$, $t \mapsto G_F(t,\tau)$ is a solution for F satisfying the initial conditions

$$\begin{aligned} G_F(\tau + jh, \tau) &= 0, \qquad j \in \{0,1,\dots,k-2\}, \\ G_F(\tau + (k-1)h, \tau) &= 1. \end{aligned}$$

Since F has constant coefficients, it is autonomous, and so by Exercise 3.3.5 there exists $H_F\colon \mathbb{T} \to \mathbb{R}$ such that $G_F(t,\tau) = H_F(t-\tau)$. Then, if we add an inhomogeneous term b to F, the particular solution of Theorem 4.7.6(iii) is

$$\xi_{\mathrm{p},b}(t) = \sum_{l=0}^{(t-h)/h} H_F(t-h-lh)b(lh).$$

Sums of the type

$$\sum f(t-\tau)g(\tau)$$

are known as ***convolution sums***. These arise in system theory and Fourier theory, for example. We shall consider convolution in the context of transform theory in . • what

4.7.2 Equations with constant coefficients

We now specialise the general discussion from the preceding section to equations with constant coefficients. Thus we are looking at scalar linear inhomogeneous ordinary difference equations with right-hand sides given by

$$\widehat{F}(t,x,x^{(1)},\dots,x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1x^{(1)} - a_0x + b(t) \tag{4.45}$$

for $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$ and $b\colon \mathbb{T} \to \mathbb{R}$. Thus a solution $t \mapsto \xi(t)$ satisfies the equation

$$\xi(t+kh) + a_{k-1}\xi(t+(k-1)h) + \cdots + a_1\xi(t+h) + a_0\xi(t) = b(t). \tag{4.46}$$

These equations are, of course, a special case of the equations considered in Section 4.7.1, and so all statements made about the general case of time-varying coefficients hold in the special case of constant coefficients. In particular, Proposition 4.7.1 and Theorem 4.7.2 hold for equations of the form (4.46). However, for these constant coefficient equations, it is possible to say some things a little more explicitly, and this is what we undertake to do.

4.7.2.1 The "method of undetermined coefficients" We present in this section a so-called method for solving scalar linear inhomogeneous ordinary difference equations with constant coefficients. With this method, one guesses a form of particular solution based on the form of the function b, and then does algebra to determine the precise solution. The "pros" and cons of the method for difference equations are the same as those for differential equations, so we refer the reader back to the beginning of Section 4.3.2.1 for this discussion.

First let us indicate the sorts of "b's" we allow.

4.7.11 Definition (Pretty uninteresting function) Let $\mathbb{T} \subseteq \mathbb{Z}(h)$ be a discrete time-domain. A function $f\colon \mathbb{T} \to \mathbb{R}$ is ***pretty uninteresting*** if it has one of the following three forms:

(i) $f(t) = \begin{cases} 1, & t = \hat{t}, \\ 0, & t \neq \hat{t}, \end{cases}$ for $\hat{t} \in \mathbb{T}$;

(ii) $f(t) = t^m r^{t/h}$ for $m \in \mathbb{Z}_{\geq 0}$ and $r \in \mathbb{R} \setminus \{0\}$;

(iii) $f(t) = t^m \rho^{t/h} \cos(\frac{\theta}{h} t)$ for $m \in \mathbb{Z}_{\geq 0}$, $\rho \in \mathbb{R}_{>0}$, and $\theta \in (0, \pi)$;

(iv) $f(t) = t^m \rho^{t/h} \sin(\frac{\theta}{h} t)$ for $m \in \mathbb{Z}_{\geq 0}$, $\rho \in \mathbb{R}_{>0}$, and $\theta \in (0, \pi)$.

For $t_0 \in \mathbb{T}$ with $\hat{t} \geq t_0$, the $\mathbf{t_0}$***-order*** in the form (i) is $h^{-1}(\hat{t} - t_0)$ and is denoted by $o(f)$. The nonnegative integer m in the forms (ii)–(iv) is the ***order*** of f and is denoted by $o(f)$. If $f\colon \mathbb{T} \to \mathbb{R}$ has the form

$$f(t) = c_1 f_1(t) + \cdots + c_r f_r(t)$$

where $c_1, \ldots, c_r \in \mathbb{R}$ and each of $f_1, \ldots, f_r$ is pretty uninteresting, then f is ***also pretty uninteresting***. •

Note that the signal from part (i) of the definition is $\tau^*_{-\hat{t}}\mathsf{P}$, the shifted pulse function, cf. Example IV-1.1.9–5.

Here are some examples of useful pretty uninteresting functions.

4.7.12 Examples (Examples of interesting pretty uninteresting functions)

1. Consider the function $\mathsf{1}_{\geq 0}\colon \mathbb{Z}_{\geq 0} \to \mathbb{R}$ defined by $\mathsf{1}_{\geq 0}(t) = 1$ for all $t \in \mathbb{Z}_{\geq 0}$. This is a "step function" and is pretty uninteresting. Often it is taken to be defined on all of $\mathbb{Z}$, and to be zero for negative times. The idea is that it gives an input to a differential equation that "switches on" at $t = 0$. Among the many particular solutions for a difference equation with $b = \mathsf{1}_{\geq 0}$, there is one that is known as the "step response," and it is determined by a specific choice of initial condition. We shall consider this in . what
2. Next consider the function $H_\theta\colon \mathbb{Z}_{\geq 0} \to \mathbb{R}$ defined by $H_\theta(t) = \sin(\frac{\theta}{h} t)$ for $\theta \in \mathbb{Z}(0, \pi)$. This is an example of an "harmonic" function, and specifically is a "sinusoid." In this case, one can think of prescribing a "b" of this form as "shaking" a difference equation. It can be interesting to know how the behaviour of the system will vary as we change θ. This gives rise in system theory to something called the "frequency response." •

We now state a few elementary properties of pretty uninteresting functions.

4.7.13 Lemma (Properties of pretty uninteresting functions) *Let* $\mathbb{T} \subseteq \mathbb{Z}(\mathrm{h})$ *be a discrete time-domain, let* $\mathrm{t}_0 \in \mathbb{T}$, *let* $\mathrm{f}, \mathrm{f}_1, \ldots, \mathrm{f}_\mathrm{r}\colon \mathbb{T} \to \mathbb{R}$ *be pretty uninteresting functions, and consider a scalar linear homogeneous ordinary difference equation* F *with constant coefficients with right-hand side of the form (4.45). Define normalised scalar linear inhomogeneous ordinary difference equations* F_j, $\mathrm{j} \in \{1, \ldots, \mathrm{r}\}$, *by*

$$\mathrm{F}_\mathrm{j}(\mathrm{t}, \mathrm{x}, \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(\mathrm{k})}) = \mathrm{x}^{(\mathrm{k})} - \widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}, \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(\mathrm{k}-1)}) - \mathrm{f}_\mathrm{j}(\mathrm{t}).$$

Then the following statements hold:

(i) there exists a unique normalised scalar linear homogeneous ordinary difference equation $\mathrm{F_f}$ *of order* $\mathrm{o(f)}$ *such that*

$$\mathrm{F_f}\,(\mathrm{t}, \mathrm{f(t)}, \mathrm{f(t+h)}, \dots, \mathrm{f(t+o(f))}) = 0, \qquad \mathrm{t} \in \mathbb{T}_{\geq \mathrm{t_0}};$$

(ii) if $\xi_\mathrm{j} \in \mathrm{Sol(F_j)}$, $\mathrm{j} \in \{1, \dots, \mathrm{r}\}$, *and if*

$$\mathrm{g} = \mathrm{c_1 f_1} + \cdots + \mathrm{c_r f_r}$$

is also pretty uninteresting, then, if $\xi = \mathrm{c_1}\xi_1 + \cdots + \mathrm{c_r}\xi_\mathrm{r}$, *then* $\xi \in \mathrm{Sol_{t_0}(F_g)}$, *where*

$$\mathrm{F_j(t, x, x^{(1)}, \dots, x^{(k)}) = x^{(k)} - \widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) - g(t).}$$

Proof (i) An examination of Procedure 4.6.18 and the attendant Theorem 4.6.19 shows that F_f can be defined by defining their characteristic polynomials as follows:

1. $f = \tau^*_{-\hat{t}}\mathsf{P}$: take
$$P_{F_f} = X^{o(f)+1}$$
2. $f(t) = t^m r^{t/h}$: take
$$P_{F_f} = (X - r)^{m+1};$$
3. $f(t) = t^m \rho^{t/h} \cos(\frac{\theta}{h}t)$ or $f(t) = t^m \rho^{t/h} \sin(\frac{\theta}{h}t)$: take
$$P_{F_f} = (X^2 - 2\rho\cos(\theta)X + \rho^2)^{m+1}.$$

(ii) This is a mere verification, once one understands the symbols involved. ∎

The differential equation F_f in the first part of the lemma we call the ***annihilator*** of the pretty uninteresting function f. The following examples illustrate how one finds the annihilator in practice, based on the proof of the first part of the lemma.

4.7.14 Examples (Annihilator) We shall work with the time-domain $\mathbb{T} = \mathbb{Z}$ and take $t_0 = 0$.

1. Consider the function $f(t) = \mathsf{P}$. As in Procedure 4.6.18, this corresponds to a root of 0 with multiplicity 1 of the characteristic polynomial. Thus we have $P_{F_f} = X$, and so
$$F_f(t, x, x^{(1)}) = x^{(1)}.$$
2. Consider the function $f(t) = 1$. This is the pretty uninteresting function $t \mapsto t^k r^t$ with $k = 0$ and $r = 1$. This corresponds, from Procedure 4.6.18, to a root $r = 1$ of a polynomial with multiplicity 1. Thus $P_{F_f} = X - 1$, and so
$$F_f(t, x, x^{(1)}) = x^{(1)} - x.$$

3. Now consider $f(t) = (-2)^t$. This is the pretty uninteresting function $t \mapsto t^k r^t$ with $k = 0$ and $r = -2$. This corresponds to a root $r = -2$ of a polynomial with multiplicity 1. Thus $P_{F_f} = X + 2$ and so
$$F_f(t, x, x^{(1)}) = x^{(1)} + 2x.$$
4. Next we take $f(t) = 2\,3^t \sin(2t) + t^2$. This is an also pretty uninteresting function, being a linear combination of $f_1(t) = 3^t \sin(2t)$ and $f_2(t) = t^2$. Note that f_1 is the pretty uninteresting function $t \mapsto t^k \rho^t \sin(\theta t)$ with $k = 0$, $\rho = 3$, and $\theta = 2$. This function is associated, via Procedure 4.6.18, with a root $\rho = 3\mathrm{e}^{2\mathrm{i}}$ of a polynomial with multiplicity 1. Of course, we must also have the root $\bar{\rho} = 3\mathrm{e}^{-2\mathrm{i}}$.
Note that f_2 is the pretty uninteresting function $t \mapsto t^k \rho^t \cos(\theta t)$ with $k = 2$, $\sigma = 0$, and $\theta = 0$. This is associated with a root $r = 0$ with multiplicity 3.
Putting this all together,
$$P_{F_f} = (X - 3\mathrm{e}^{2\mathrm{i}})(X - 3\mathrm{e}^{-2\mathrm{i}})X^3 = X^5 - 6\cos(2)X^4 + 9X^3. \qquad \bullet$$

The second part of the lemma points out, in short, the obvious fact that if "b" is also pretty uninteresting, then one can obtain a particular solution by obtaining a particular solution for each of its pretty uninteresting components, and then summing these with the same coefficients as in the also pretty uninteresting function. The point of this is that, to obtain a particular solution for an also pretty uninteresting "b," it suffices to know how to do this for a pretty uninteresting b. Thus we deliver the following construction.

4.7.15 Procedure (Method of undetermined coefficients) We let F be a normalised scalar linear inhomogeneous ordinary difference equation with constant coefficients with right-hand side
$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x + f(t),$$
where f is pretty uninteresting. Do the following.

1. Let F_f be the annihilator of f.
2. Let G_f be the normalised scalar linear homogeneous ordinary differential equation whose characteristic polynomial is $P_{G_f} = P_{F_f} P_{F_\mathrm{h}}$.
3. Using Procedure 4.6.18, find
   (a) pretty uninteresting functions $\xi_1, \dots, \xi_k$ for which $\{\xi_1, \dots, \xi_k\}$ is a fundamental set of solutions for F_h and
   (b) pretty uninteresting functions $\eta_1, \dots, \eta_{o(f)+1}$ for which $\{\xi_1, \dots, \xi_k, \eta_1, \dots, \eta_{o(f)+1}\}$ is a fundamental set of solutions for G_f.
4. For (as yet) undetermined coefficients $c_1, \dots, c_{o(f)+1} \in \mathbb{R}$, denote
$$\xi_\mathrm{p} = c_1\eta_1 + \cdots + c_{o(f)+1}\eta_{o(f)+1}.$$

5. Determine $c_1, \dots, c_{o(f)+1}$ by demanding that ξ_{p} be a particular solution for F.

We shall show that this procedure makes sense and defines a particular solution for F. •

Let us verify that the preceding procedure gives what we want.

4.7.16 Proposition (Validity of the method of undetermined coefficients) *Let* F *be a normalised scalar linear inhomogeneous ordinary difference equation with constant coefficients with right-hand side*

$$\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}, \mathrm{x}^{(1)}, \dots, \mathrm{x}^{(\mathrm{k}-1)}) = -\mathrm{a}_{\mathrm{k}-1}\mathrm{x}^{(\mathrm{k}-1)} - \cdots - \mathrm{a}_1\mathrm{x}^{(1)} - \mathrm{a}_0\mathrm{x} + \mathrm{f}(\mathrm{t}),$$

where f *is pretty uninteresting. Then all steps in Procedure 4.7.15 are unambiguously defined, and the result is a particular solution for* F *from* t_0.

Proof We first consider the case where $f = \tau^*_{-\hat{t}}\mathsf{P}$, for $\hat{t} \geq t_0$. Then $o(f) = h^{-1}(\hat{t} - t_0)$. From Procedure 4.6.18 we have $P_{F_f} = X^{o(f)+1}$. Let us write

$$P_{F_{\mathrm{h}}} = X^{m(0)}P,$$

where P does not have a zero root. Therefore,

$$P_{G_f} = X^{m(0)+o(f)+1}P.$$

Then, according to Procedure 4.6.18, among the pretty uninteresting solutions for F_{h} are

$$\tau^*_{-(t_0+jh)}\mathsf{P}, \qquad j \in \{0, 1, \dots, m(0) - 1\}.$$

The rest of the pretty uninteresting solutions for F_{h} have nothing to do with the root "r" of the characteristic polynomial, and are not interesting to us here. Now the $o(f)+1$ pretty uninteresting solutions for G_f that are added to those for F_{h} are

$$\tau^*_{-(t_0+jh)}\mathsf{P}, \qquad j \in \{m(0), \dots, m(0) + o(f)\},$$

again according to Procedure 4.6.18. This demonstrates the viability of the first three steps of Procedure 4.6.18. We now need to show that one can solve for the coefficients $c_1, \dots, c_{o(f)+1}$ to obtain a particular solution for F. If

$$\xi_{\mathrm{p}}(t) = c_1\tau^*_{-(t_0+m(0)h)}\mathsf{P} + \cdots + c_{o(f)+1}\tau^*_{-(t_0+(m(0)+o(f))h)}\mathsf{P},$$

then (4.48) gives

$$\begin{aligned} F_{\mathrm{h}}\left(t, \xi_{\mathrm{p}}(t), \xi_{\mathrm{p}}(t+h), \dots, \xi_{\mathrm{p}}(t+kh)\right) &= \sum_{j=m(0)}^{k} a_j\xi_{\mathrm{p}}(t+jh) \\ &= \sum_{j=m(0)}^{k}\sum_{a=1}^{o(f)+1} a_j c_a \tau_{-(t_0+(m(0)+a-1)h)}\mathsf{P}(t+jh) \\ &= \sum_{j=m(0)}^{k} a_j c_{h^{-1}(t-t_0)+j-(m(0)-1)}. \end{aligned} \tag{4.47}$$

Note that the sum is meaningful for

$$0 \leq h^{-1}(t - t_0) + j - m(0) \leq o(f).$$

For $t = o(f)h$, we thus must have $j \in \{m(0)\}$. For $t = (o(f) - 1)h$, we must have $j \in \{m(0), m(0) + 1\}$. If we write the corresponding equations obtained by setting the expression (4.48) equal to f for times $t = o(f)h, t = (o(f) - 1)h, \ldots$, we get

$$\begin{aligned} a_{m(0)}c_1 &= 1, \\ a_{m(0)+1}c_1 + a_{m(0)}c_2 &= 0, \\ &\vdots \end{aligned}$$

Thus we can recursively solve for c_1, then c_2, and so on.

Next we shall assume that $f(t) = t^{o(f)}r^{t/h}$ for $r \in \mathbb{R} \setminus \{0\}$. Entirely similar reasoning works for the remaining two sorts of pretty uninteresting functions.

From Procedure 4.6.18 we know that $P_{F_f} = (X - r)^{o(f)+1}$. Let us suppose that

$$P_{F_{\mathrm{h}}} = (X - r)^{m(r)}P, \tag{4.48}$$

where P does not have r as a root. Therefore,

$$P_{G_f} = (X - r)^{m(r)+o(f)+1}P.$$

Then, according to Procedure 4.6.18, among the pretty uninteresting solutions for F_{h} are

$$t \mapsto t^j r^{t/h}, \qquad j \in \{0, 1, \ldots, m(r) - 1\}.$$

The rest of the pretty uninteresting solutions for F_{h} have nothing to do with the root "r" of the characteristic polynomial, and are not interesting to us here. Now the $o(f)+1$ pretty uninteresting solutions for G_f that are added to those for F_{h} are

$$t \mapsto t^j r^{t/h}, \qquad j \in \{m(r), \ldots, m(r) + o(f)\},$$

again according to Procedure 4.6.18. This demonstrates the viability of the first three steps of Procedure 4.6.18. We now need to show that one can solve for the coefficients $c_1, \ldots, c_{o(f)+1}$ to obtain a particular solution for F. If

$$\xi_{\mathrm{p}}(t) = c_1 t^{m(r)} r^{t/h} + \cdots + c_{o(f)+1} t^{m(r)+o(f)} r^{t/h},$$

then Lemma 1 from the proof of Theorem 4.6.19 shows that $D_r^{m(r)}\xi_{\mathrm{p}}$ is an also pretty uninteresting function whose highest order (as a pretty uninteresting function) term is of order $o(f)$. By Corollary 4.6.17, and since the forward differences of a pretty uninteresting function of order m are also pretty uninteresting function of order m (as can be verified by a direct computation), we have that

$$F_{\mathrm{h}}\big(t, \xi_{\mathrm{p}}(t), \xi_{\mathrm{p}}(t + h), \ldots, \xi_{\mathrm{p}}(t + kh)\big)$$

is an also pretty uninteresting function of order $o(f)$ associated with the root r. Therefore, we can use the equality

$$F_{\mathrm{h}}\big(t, \xi_{\mathrm{p}}(t), \xi_{\mathrm{p}}(t + h), \ldots, \xi_{\mathrm{p}}(t + kh)\big) = f(t)$$

to solve for the coefficients $c_1, \ldots, c_{o(f)+1}$, as asserted in Procedure 4.6.18. ■

While the preceding discussion does indeed provide a means of solving, in principle, scalar linear inhomogeneous ordinary difference equations with also pretty uninteresting "b's," it does tend to be a lot of work, and there are precisely zero equations that can be solved by this procedure that cannot far more easily be solved with a computer.

4.7.2.2 Some examples Here we consider two interesting examples, one with a first-order difference equation with a step function as right-hand side and the other with a second order difference equation with harmonic right-hand side. Since much of the behaviour exhibited by these systems resembles that for differential equations exhibited in Examples 4.3.19 and 4.3.20, we shall not go into as much detail here as we did with the differential equation examples.

4.7.17 Example (First-order system with step input) The difference equation we consider here is an inhomogeneous version of the equation considered in Example 4.6.20. We take the first-order scalar linear inhomogeneous ordinary difference equation F with constant coefficients and with right-hand side

$$\widehat{F}(t, x) = -\rho x + 1.$$

Thus solutions $t \mapsto \xi(t)$ to this difference equation satisfy

$$\xi(t + h) + \rho\xi(t) = 1.$$

We have already determined that a solution to the homogeneous equation will have the form $\xi(t) = c(-\rho)^{t/h}$, when $\rho \neq 0$. When $\rho = 0$, then the homogeneous equation is not invertible, and so the homogeneous solutions from t_0 will be

$$\xi(t) = \begin{cases} c, & t = t_0, \\ 0, & t > t_0. \end{cases}$$

So next we find a particular solution. The annihilator F_f of the pretty uninteresting function $f(t) = 1$ has characteristic polynomial $P_{F_f} = X - 1$. The characteristic polynomial for F_{h} is $P_{F_{\mathrm{h}}} = X + \rho$. Thus we must list the fundamental solutions for G_f, where

$$P_{G_f} = (X - 1)(X + \rho).$$

There are two cases.

First, when $\rho \neq -1$, the fundamental solutions are $t \mapsto (-\rho)^{t/h}$ (when $\rho \neq 0$) or the impulse at t_0 (when $\rho = 0$), and $t \mapsto 1$, using Procedure 4.6.18. The first of these is a solution for the homogeneous equation, so we take a particular solution to be a multiple of the second: $\xi_{\mathrm{p}}(t) = c$. To find c we substitute into the differential equation:

$$\xi(t + h) + \rho\xi(t) = (1 + \rho)c.$$

To be a particular solution, we must have $(1+\rho)c = 1$ and so $c = 1/(1+\rho)$. Thus $\xi_{\mathrm{p}}(t) = 1/(1+\rho)$.

Next we consider the case when $\rho = -1$, and in this case the fundamental solutions for G_f are $t \mapsto 1$ and $t \mapsto t$, again using Procedure 4.6.18. In this case, the first of these functions is a solution for the homogeneous system, and so a multiple of the second will be a particular solution, i.e., $\xi_{\mathrm{p}}(t) = ct$. To determine c we require that ξ_{p} be a particular solution:

$$\xi(t+h) - \xi(t) = c(t+h) - ct = ch,$$

from which we deduce that $ch = 1$. Thus $\xi_{\mathrm{p}}(t) = \frac{t}{h}$.

In summary, a particular solution is

$$\xi_{\mathrm{p}}(t) = \begin{cases} \frac{1}{1+\rho}, & \rho \neq -1, \\ \frac{t}{h}, & \rho = -1. \end{cases}$$

Let us now determine the full solution, including initial conditions at $t_0 = 0$. We have three cases.

1. $\rho \notin \{0, -1\}$: Any solution has the form

$$\xi(t) = c(-\rho)^{t/h} + \frac{1}{1+\rho}.$$

Applying the initial condition $\xi(0) = 0$ gives $c = -\frac{1}{1+\rho}$ and so

$$\xi(t) = \frac{1}{1+\rho}(1 - (-\rho^{t/h})).$$

To allow a fruitful comparison of the effects of changing ρ, let us normalise this solution by multiplying by $1+\rho$ to get the ***step response***

$$1_F(t) = 1 - (-\rho)^{t/h}.$$

2. $\rho = -1$: Here any solution has the form

$$\xi(t) = c + \frac{t}{h}.$$

Applying the initial condition $\xi(0) = 0$ gives $c = 0$ and so the solution is

$$1_F(t) = \frac{t}{h}.$$

3. $\rho = 0$: Here we have any solution from 0 given by

$$\xi(t) = \begin{cases} c+1, & t = 0, \\ 1, & t > 0. \end{cases}$$

The initial condition $\xi(0) = 0$ gives $c = -1$ and so

$$\xi(t) = \begin{cases} 0, & t = 0, \\ 1, & t > 0. \end{cases}$$

In Figure 4.9 we graph this step response for a couple of values of $\rho \in \mathbb{R}_{<0}$.

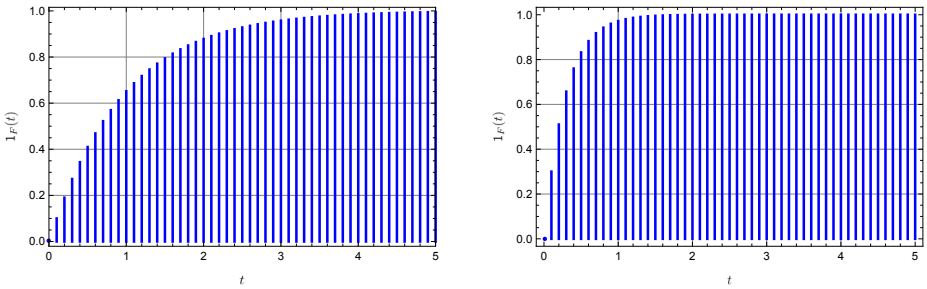


Figure 4.9 The step response of a first-order system for $\rho = -0.9$ (left) and $\rho = -0.7$ (right)

We note that as $-\rho$ gets smaller in magnitude, the step response rises more quickly, i.e., responds faster. •

4.7.18 Example (Second-order system with sinusoidal input) Next we consider a special case of the second-order differential equation of Example 4.6.21, but with a sinusoidal inhomogeneous term. Thus we take the second-order scalar linear inhomogeneous ordinary differential equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}) = -\rho^2 x + 2\rho \cos(\theta_0) x^{(1)} + A \sin(\tfrac{\theta}{h} t)$$

for $A, \theta \in \mathbb{R}_{>0}$. Solutions $t \mapsto \xi(t)$ then satisfy

$$\xi(t + 2h) - 2\rho \cos(\theta_0)\xi(t + h) + \rho^2 \xi(t) = A \sin(\tfrac{\theta}{h} t).$$

In Example 4.6.21 we carefully and thoroughly investigated the nature of the solutions for the homogeneous system. There we saw, for example, that as long as $\rho \in (0, 1)$, solutions to the homogeneous equation decay to zero as $t \to \infty$. For $\rho = 1$, solutions were periodic. Here we will thus focus on $\rho \in (0, 1]$ and on the nature of the particular solution. When $\rho \in (0, 1)$, this means that we are looking at the "steady-state" behaviour of the system, i.e., what we see after a long time. When $\zeta = 0$, we do not have this steady-state interpretation, but nonetheless we will interpret these solutions in light of our understanding of what happens when $\zeta \in \mathbb{R}_{>0}$.

The annihilator F_f for the pretty uninteresting function $f(t) = A\sin(\frac{\theta}{h}t)$ has characteristic polynomial $P_{F_f} = X^2+\theta^2$. We have two cases to consider for particular solutions.

The first case is when $\zeta \neq 1$ or when $\zeta = 1$ and $\theta \neq \theta_0$. Here the characteristic polynomial for G_f in Procedure 4.7.15 is

$$P_{G_f} = (X^2 + \theta^2)(X^2 - 2\rho\cos(\theta_0)X + \rho^2).$$

The fundamental solutions for G_f associated to this polynomial, according to Procedure 4.6.18, are

$$\xi_1(t),\ \xi_2(t),\ \cos(\tfrac{\theta}{h}t),\ \sin(\tfrac{\theta}{h}t),$$

where ξ_1 and ξ_2 are homogeneous solutions as determined in Example 4.6.21. Thus a particular solution will be of the form

$$\xi_{\mathrm{p}}(t) = c_1\cos(\tfrac{\theta}{h}t) + c_2\sin(\tfrac{\theta}{h}t).$$

To determine c_1 asnd c_2 we require that ξ_{p} be a particular solution. Thus we compute

$$\begin{aligned}
&\xi_{\mathrm{p}}(t+2h) - 2\rho\cos(\theta_0)\xi_{\mathrm{p}}(t+h) + \rho^2\xi_{\mathrm{p}}(t)\\
&\quad(c_1\rho^2 - 2c_1\rho\cos(\theta)\cos(\theta_0) + c_1\cos(2\theta) - 2c_2\rho\sin(\theta)\cos(\theta_0) + c_2\sin(2\theta))\cos(\tfrac{\theta}{h}t)\\
&+ (2c_1\rho\sin(\theta)\cos(\theta_0) - c_1\sin(2\theta) + c_2\rho^2 - 2c_2\rho\cos(\theta)\cos(\theta_0) + c_2\cos(2\theta))\sin(\tfrac{\theta}{h}t),
\end{aligned}$$

upon employing some standard trigonometric identities. We can solve these equations to give

$$\begin{aligned}
c_1 &= -\frac{2A\sin(\theta)(\cos(\theta) - \rho\cos(\theta_0))}{\rho^4 - 4(\rho^2+1)\rho\cos(\theta)\cos(\theta_0) + 2\rho^2\cos^2(\theta) + \rho^2\cos(2\theta) + 2\rho^2\cos(2\theta_0) + \rho^2 + 1},\\
c_2 &= \frac{A(\rho(\rho - 2\cos(\theta)\cos(\theta_0)) + \cos(2\theta))}{\rho^4 - 4(\rho^2+1)\rho\cos(\theta)\cos(\theta_0) + 2\rho^2\cos^2(\theta) + \rho^2\cos(2\theta) + 2\rho^2\cos(2\theta_0) + \rho^2 + 1}.
\end{aligned}$$

Thus a particular solution is

$$\xi_{\mathrm{p}}(t) = c_1\cos(\tfrac{\theta}{h}t) + c_2\sin(\tfrac{\theta}{h}t),$$

with c_1 and c_2 as just defined.

The other case is when $\rho = 1$ and $\theta = \theta_0$. Here the characteristic polynomial for G_f in Procedure 4.7.15 is

$$P_{G_f} = (X^2 + \theta^2)^2$$

The fundamental solutions for G_f associated to this polynomial, according to Procedure 4.6.18, are

$$\xi_1(t),\ \xi_2(t),\ t\cos(\tfrac{\theta}{h}t),\ t\sin(\tfrac{\theta}{h}t),$$

where ξ_1 and ξ_2 are homogeneous solutions as determined in Example 4.6.21. Therefore, a particular solution will have the form

$$\xi_{\mathrm{p}}(t) = c_1 t \cos(\tfrac{\theta}{h}t) + c_2 t \sin(\tfrac{\theta}{h}t).$$

To determine c_1 and c_2 we ask that this be a particular solution. Thus we compute, after a tedious computation,

$$\begin{aligned}&\xi_{\mathrm{p}}(t+2h) - 2\cos(\theta)\xi_{\mathrm{p}}(t+h) + \xi_{\mathrm{p}}(t)\\ &= 2h\sin(\theta)(-c_1\sin(\theta) + c_2\cos(\theta))\cos(\tfrac{\theta}{h}t) - 2h\sin(\theta)(c_1\cos(\theta) + c_2\sin(\theta))\sin(\tfrac{\theta}{h}t).\end{aligned}$$

We can now solve for c_1 and c_2 to get

$$\begin{aligned}c_1 &= -\frac{A\cot(\theta)}{2h},\\ c_2 &= -\frac{A}{2h},\end{aligned}$$

and so the particular solution we obtain is

$$\xi_{\mathrm{p}}(t) = -\frac{A\cot(\theta)}{2h}\cos(\tfrac{\theta}{h}t) - \frac{A}{2h}\sin(\tfrac{\theta}{h}t).$$

In both of the above cases, any solution will be a sum of the obtained particular solution, plus some solution to the homogeneous equation as determined in Example 4.6.21.

In Figure 4.10 we graph particular solutions for various θ's, keeping other parameters fixed. The observations one makes here echo those made in Example 4.3.20. •

Exercises

4.7.1 Consider the ordinary difference equation F with right-hand side given by (4.41).

(a) Convert this to a first-order equation with k states, following Exercise 3.3.7.

(b) Show that the resulting first-order equation satisfies the conditions of Theorem 3.4.2 for existence of a unique solution $t \mapsto \xi(t)$ satisfying the initial conditions

$$\xi(t_0) = x_0, \xi(t_0 + h) = x_0^{(1)}, \dots, \xi(t_0 + (k-1)h) = x_0^{(k-1)}$$

at time $t_0 \in \mathbb{T}$.

4.7.2 Consider the ordinary difference equation F with right-hand side given by (4.41). Answer the following questions.

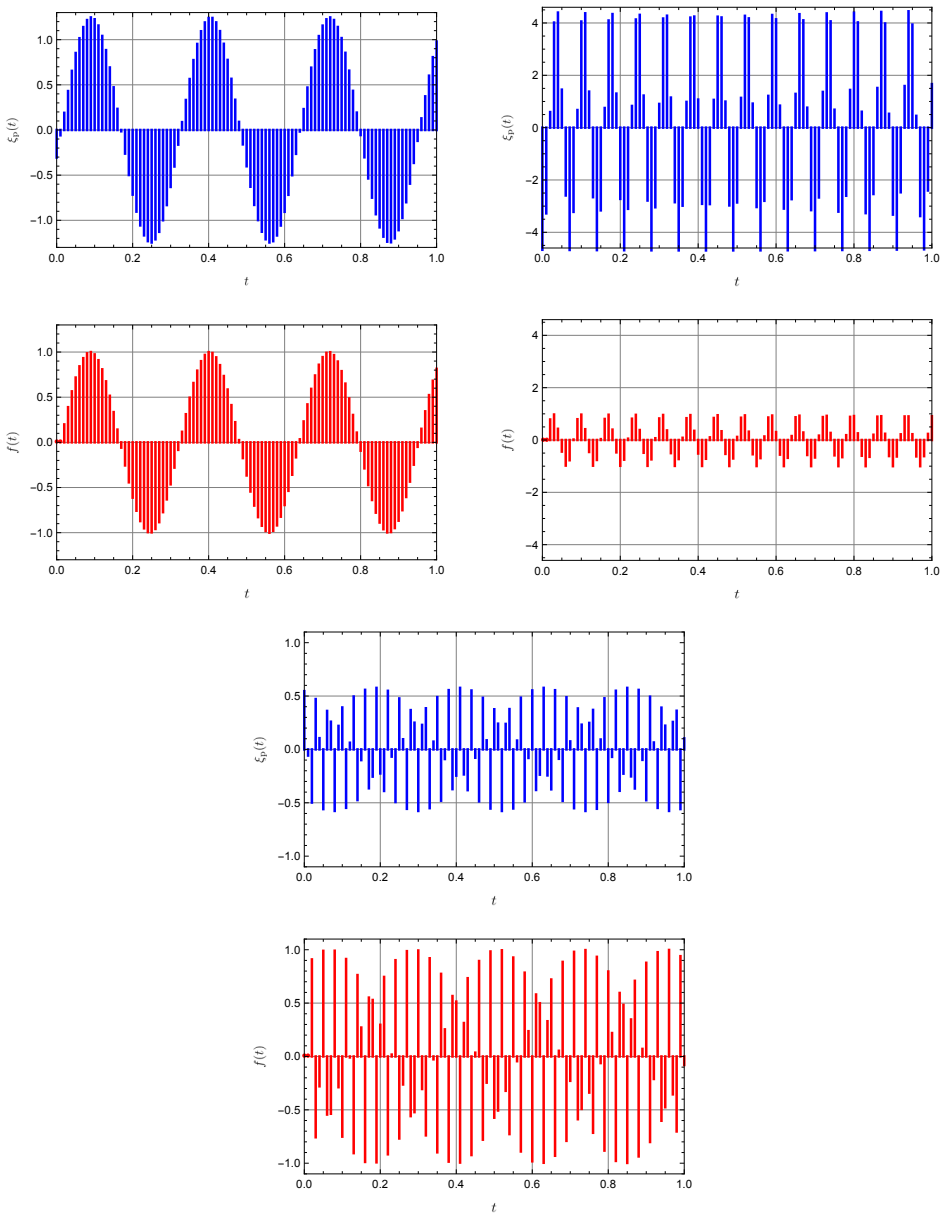


Figure 4.10 Response (in blue) of a second-order system with $h = 0.01$, $\rho = 0.9$, and $\theta_0 = 1$ to a sinusoidal input with $A = 1$ (in red) for varying θ (top left: $\theta = 0.2$; top right: $\theta = 0.9$; bottom: $\theta = 2$)

(a) Show that the particular particular solution

$$\xi_{\mathrm{p},b}(t) = \sum_{l=0}^{(t-t_0-h)/h} G_F(t, t_0 + lh) b(t_0 + lh),$$

satisfies the initial value problem

$$\begin{aligned}\xi(t+kh)+a_{k-1}(t)\xi(t+(k-1)h)+\cdots+a_1(t)\xi(t+h)+a_0(t)\xi(t)=b(t),\\ \xi(t_0)=0,\ \xi(t_0+h)=0,\dots,\ \xi(t_0+(k-1)h)=0.\end{aligned}$$

(b) Show that the solution to the initial value problem

$$\begin{aligned}\xi(t+kh)+a_{k-1}(t)\xi(t+(k-1)h)+\cdots+a_1(t)\xi(t+h)+a_0(t)\xi(t)=b(t),\\ \xi(t_0)=x_0,\ \xi(t_0+h)=x_0^{(1)},\dots,\ \xi(t_0+(k-1)h)=x_0^{(k-1)}\end{aligned}$$

is given by $\xi(t)=\xi_{\mathrm{h}}+\xi_{\mathrm{p},b}$, where ξ_{h} is the solution to the homogeneous initial value problem

$$\begin{aligned}\xi_{\mathrm{h}}(t+kh)+a_{k-1}(t)\xi_{\mathrm{h}}(t+(k-1)h)+\cdots+a_1(t)\xi_{\mathrm{h}}(t+h)+a_0(t)\xi_{\mathrm{h}}(t)=0,\\ \xi_{\mathrm{h}}(t_0)=x_0,\ \xi_{\mathrm{h}}(t_0+h)=x_0^{(1)},\dots,\ \xi_{\mathrm{h}}(t_0+(k-1)h)=x_0^{(k-1)}.\end{aligned}$$

4.7.3 Let $\mathbb{T}\subseteq\mathbb{Z}(h)$ and consider a scalar linear homogeneous ordinary difference equation F with right-hand side given by (4.29). Show that the solutions to the following initial value problems are the same:

$$\begin{aligned}\xi(t+kh)+a_{k-1}(t)\xi(t+(k-1)h)+\cdots+a_1(t)\xi_1(t+h)+a_0(t)\xi(t)=0,\\ \xi(t_0)=0,\ \xi(t+h)=0,\dots,\xi(t+(k-1)h)=1,\end{aligned}$$

and

$$\begin{aligned}\xi(t+kh)+a_{k-1}(t)\xi(t+(k-1)h)+\cdots+a_1(t)\xi_1(t+h)+a_0(t)\xi(t)=1,\\ \xi(t_0)=0,\ \xi(t+h)=0,\dots,\xi(t+(k-1)h)=0.\end{aligned}$$

(The point is that the solution to an impulse at t_0 with zero initial condition can be determined by the homogeneous equation with the initial conditions that show up for the Green's function.)

Section 4.8

Laplace transform methods for scalar ordinary difference equations

Jut as the causal CLT can be used to study various sorts of differential equations, the causal DLT can be used to study difference equations. In this section, we will stick to considering the application of Laplace transform techniques to the study of scalar linear ordinary difference equations with constant coefficients. We shall consider systems of equations in Section 5.8.

Do I need to read this section? This is a section that can, maybe, be skipped. It will have its best context in the setting of transfer functions in Chapter 7. •

4.8.1 Scalar homogeneous equations

We begin our discussion with scalar linear homogeneous ordinary difference equations with constant coefficients, first considered in detail in Section 4.6.2. Thus, as in that section we are working with difference equations

$$F\colon \mathbb{Z}_{\geq 0}(h) \times \mathbb{R} \oplus \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) \to \mathbb{R}$$

with right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x \tag{4.49}$$

for $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$. Given Proposition IV-9.2.16, the causal DLT is particularly well suited for working with ordinary difference equations with initial conditions. Thus we shall consider the initial value problem

$$\begin{aligned} \xi(t+kh) + a_{k-1}\xi(t+(k-1)h) + \cdots + a_1\xi(t+h) + a_0\xi(t) = 0, \\ \xi(0) = x_0,\ \xi(h) = x_0^{(1)},\ \cdots,\ \xi(k-1)h) = x_0^{(k-1)}. \end{aligned} \tag{4.50}$$

We shall now take the causal DLT of this initial value problem. To do so, it is tacitly assumed that all members of $\mathrm{Sol}(F)$ and their derivatives are in $\mathsf{LT}^{1,+}(\mathbb{Z}_{\geq 0}(h);\mathbb{C})$ so that we may use the difference rule of Proposition IV-9.2.16. This is true, however, since all members of $\mathrm{Sol}(F)$ are also pretty uninteresting functions, and so are in what? $\mathsf{LT}^{1,+}(\mathbb{Z}_{\geq 0}(h);\mathbb{C})$, when restricted to the domain $\mathbb{Z}_{\geq 0}(h)$, as we saw in . Another way to think of taking the causal DLT of the equation, were one to not know *a priori* that solutions were Laplace transformable, would be to go ahead and take the transform assuming this is so, and then see if the assumption is valid by seeing if the equation can be solved (or by some other means). In any case, the following result records what happens when we take the causal DLT of the initial value problem.

4.8.1 Proposition (Causal DLT of scalar homogeneous equation) *The causal DLT of the initial value problem (4.50) has the solution*

$$\mathscr{L}^1_{\mathrm{D}}(\xi)(\mathrm{z}) = \frac{\sum_{\mathrm{j}=0}^{\mathrm{k}} \sum_{\mathrm{l}=0}^{\mathrm{j}-1} \mathrm{a_j} (\mathrm{hz})^{\mathrm{l}+1} \xi((\mathrm{j}-\mathrm{l}-1)\mathrm{h})}{\mathrm{z}^{\mathrm{k}} + \mathrm{a}_{\mathrm{k}-1}\mathrm{z}^{\mathrm{k}-1} + \cdots + \mathrm{a}_1 \mathrm{z} + \mathrm{a}_0},$$

with the convention that $\mathrm{a_k} = 1$.

Proof By Corollary IV-9.2.17 we have

$$\mathscr{L}^1_{\mathrm{D}}(\tau^*_{-jh}\xi)(z) = z^j \mathscr{L}^1_{\mathrm{D}}(\xi)(z) - \sum_{l=0}^{j-1} (hz)^{l+1} \xi((j-l-1)h), \qquad j \in \{0, 1, \ldots, k\}.$$

Therefore, with the stated convention that $a_k = 1$,

$$\mathscr{L}^1_{\mathrm{D}}\left(\sum_{j=0}^{k} a_j \tau^*_{-jh}\xi\right) = \sum_{j=0}^{k} a_j \left(z^j \mathscr{L}^1_{\mathrm{D}}(\xi)(z) - \sum_{l=0}^{j-1} (hz)^{l+1} \xi((j-l-1)h)\right),$$

and solving this equation for $\mathscr{L}^1_{\mathrm{D}}(\xi)(z)$ gives the asserted conclusion. ■

To obtain the solution to the initial value problem in the time-domain, we should apply the inverse transform to the expression from the proposition. To do this, one could, in principle, apply the definition of the inverse causal DLT of Theorem IV-9.2.13. However, in cases where one can actually compute the inverse transform, it is not typically done in this way. Indeed, typically one "looks up" the answer, and one can do this using partial fraction expansion as in Procedure 4.5.2. Let us see how one uses the partial fraction expansion to compute the inverse causal DLT of the expression from Proposition 4.8.1. This is most easily done via examples.

4.8.2 Examples (Solving scalar homogeneous equations using the causal DLT)
1.

Our comments about using partial fraction expansions to solve homogeneous scalar linear ordinary differential equations applies here as well.

4.8.2 Scalar inhomogeneous equations

We next consider scalar linear inhomogeneous ordinary difference equations, first considered in Section 4.7.2. Thus we are working with scalar ordinary difference equations with right-hand sides given by

$$\widehat{F}(t, x, x^{(1)}, \ldots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x + b(t) \tag{4.51}$$

for $a_0, a_1, \ldots, a_{k-1} \in \mathbb{R}$ and $b \colon \mathbb{Z}_{\geq 0}(h) \to \mathbb{R}$. The initial value problem we consider is then

$$\begin{aligned} \xi(t+kh) + a_{k-1}\xi(t+(k-1)h) + \cdots + a_1\xi(t+h) + a_0\xi(t) = b(t), \\ \xi(0) = x_0,\ \xi(h) = x_0^{(1)},\ \cdots,\ \xi((k-1)h) = x_0^{(k-1)}. \end{aligned} \tag{4.52}$$

If $b \in \ell_{\mathrm{loc}}(\mathbb{Z}_{\geq 0}(h); \mathbb{R})$, then the solution to the initial value problem (4.52) is given by $\xi(t) = \xi_{\mathrm{h}}(t) + H_F * b(t)$, where ξ_{h} satisfies the homogeneous initial value problem

$$\xi_{\mathrm{h}}(t+kh) + a_{k-1}\xi_{\mathrm{h}}(t+(k-1)h) + \cdots + a_1\xi_{\mathrm{h}}(t+h) + a_0\xi_{\mathrm{h}}(t) = 0,$$
$$\xi_{\mathrm{h}}(0) = x_0,\ \xi_{\mathrm{h}}(h) = x_0^{(1)},\ \cdots,\ \xi_{\mathrm{h}}((k-1)h) = x_0^{(k-1)},$$

where $H_F(t-\tau) = G_F(t,\tau)$ and where G_F is the Green's function from Section 4.7.1.3. This follows from Remark 4.7.10 and Exercise 4.7.2.

We wish to provide an interpretation of this strategy using the causal DLT and the connection of the transform and convolution from Propositions IV-9.2.9 and IV-9.2.10. As with inhomogeneous equations above, we take the causal DLT of the equation (4.52). However, unlike in the homogeneous case, here taking the transform is not generally valid; indeed, it is valid if and only if $b \in \mathsf{LT}^{1,+}(\mathbb{Z}_{\geq 0}(h); \mathbb{R})$. To apply the convolution result from Propositions IV-9.2.9 and IV-9.2.10, we further assume that $b \in \mathsf{LT}^{1,+}(\mathbb{Z}_{\geq 0}(h); \mathbb{R})$.

First let us give determine the causal DLT of the Green's function in this case.

4.8.3 Proposition (Causal DLT and the Green's function) *Consider the scalar linear homogeneous ordinary difference equation* F *with right-hand side (4.49). Let* $\mathrm{G_F}$ *be the Green's function and denote* $\mathrm{H_F(t-\tau) = G_F(t,\tau)}$. *Then the causal DLT of* $\mathrm{H_F}$ *is given by*

$$\mathscr{L}^1_{\mathrm{D}}(\mathrm{H_F})(\mathrm{z}) = \frac{\mathrm{hz}}{\mathrm{z^k + a_{k-1}z^{k-1} + \cdots + a_z z + a_0}}.$$

Proof According to Remark 4.7.10, $G_F(t,\tau) = H_F(t-\tau)$, where H_F satisfies the initial value problem

$$H_F(t+kh) + a_{k-1}H_F((k-1)h) + \cdots + a_1H_F(t+h) + a_0H_F(t) = 0,$$
$$H_F(0) = 0,\ H_K(h) = 0, \ldots,\ H_F((k-2)h)(0) = 0,\ H_F((k-1)h) = 1.$$

Therefore, according to Proposition 4.8.1,

$$\mathscr{L}_{\mathrm{D}}(H_F)(z) = \frac{hz}{z^k + a_{k-1}z^{k-1} + \cdots + a_z z + a_0},$$

as claimed. ∎

By combining Proposition 4.8.1 with the preceding result and the convolution solution $\xi(t) = \xi_{\mathrm{h}}(t) + H_F * b(t)$ of the initial value problem (4.52), we obtain the following result.

4.8.4 Proposition (Causal DLT of scalar inhomogeneous equation) *Consider the scalar ordinary difference equation with right-hand side (4.51), and suppose that* $\mathrm{b} \in \mathsf{LT}^{1,+}(\mathbb{Z}_{\geq 0}(\mathrm{h}); \mathbb{R})$. *The causal DLT of the solution of the initial value problem (4.52) is given by*

$$\mathscr{L}^1_{\mathrm{D}}(\xi)(\mathrm{z}) = \frac{\sum_{\mathrm{j}=0}^{\mathrm{k}} \sum_{\mathrm{l}=0}^{\mathrm{j}-1} \mathrm{a_j(hz)^{l+1}}\xi(\mathrm{(j-l-1)h}) + \mathscr{L}^1_{\mathrm{D}}(\mathrm{b})(\mathrm{z})}{\mathrm{z^k + a_{k-1}z^{k-1} + \cdots + a_1 z + a_0}},$$

with the convention that $\mathrm{a_k} = 1$.

There are two ways in which the proposition has value. One is theoretical and one is that it provides another tedious algorithmic procedure—augmenting the "method of undetermined coefficients"—for computing solutions when the inhomogeneous term is an also pretty uninteresting function. Let us consider these in turn.

Next let us turn to a less interesting but somehow more concrete application of the causal DLT in the study of scalar linear inhomogeneous ordinary difference equations. Specifically, we consider such an equation F with right-hand side (4.51), and where b is an also pretty uninteresting function. In this case, as we see from , the causal DLT $\mathscr{L}^1_{\mathrm{D}}(b)$ of b will be a rational function of the complex variable z what? whose numerator polynomial has degree strictly less than that of the denominator polynomial. Therefore, as per Proposition 4.8.4, the causal DLT $\mathscr{L}^1_{\mathrm{D}}(\xi)$ of the solution ξ of the initial value problem (4.52) will itself be such a rational function of z. Thus we can perform a partial fraction expansion of $\mathscr{L}^1_{\mathrm{D}}(\xi)$ as per Procedure 4.5.2, and then perform the inversion of the causal DLT as per Example 4.8.2 to obtain the solution. This is not something to be belaboured—not least because we already have the often easier "method of undetermined coefficients" for such situations–and we content ourselves with an illustration via a example.

4.8.5 Example (Solving scalar inhomogeneous equations using the causal DLT)

Exercises

4.8.1

Section 4.9

Using a computer to work with scalar ordinary differential equations

We thank Jack Horn for putting together the MATHEMATICA® *and* MATLAB® *results in this section.*

In Sections 4.2 and 4.3 we have discussed the character of, and solved very specific examples of, scalar linear ordinary differential equations. This, however, represents a tiny subset (but, arguably, an important tiny subset) of the differential equations one might encounter in practice. Moreover, even in the simple examples where the analytical methods we have learnt *are* applicable, to apply them is often extremely tedious and error-prone. Therefore, in this section we illustrate how computers can make working with differential equations, specifically scalar ordinary differential equations, a bearable undertaking.

list packages

In the we listed a couple of computer packages—some symbolic, some numerical, some both—available for working with differential equations. We shall not attempt to illustrate how all of these work, but choose two as illustrative. We choose MATHEMATICA® to illustrate a computer algebra package[7] and MATLAB® to illustrate a numerical package. There is no reason for this choice, other than personal familiarity (in the case of MATHEMATICA®) and ease of access (in the case of MATLAB®).

4.9.1 Using MATHEMATICA® to obtain analytical and/or numerical solutions

For some ordinary differential equations, one can simply plug them into a computer algebra package, and out will pop the answer. So, this is always worth a shot.

Our first example illustrates this in MATHEMATICA®.

4.9.1 Example (Solving simple scalar ordinary differential equation) The first ordinary differential equation we will solve is the simple first order equation:

$$\frac{\mathrm{d}y}{\mathrm{d}t}(t) = \frac{-ty(t)}{2 - y(t)}.$$

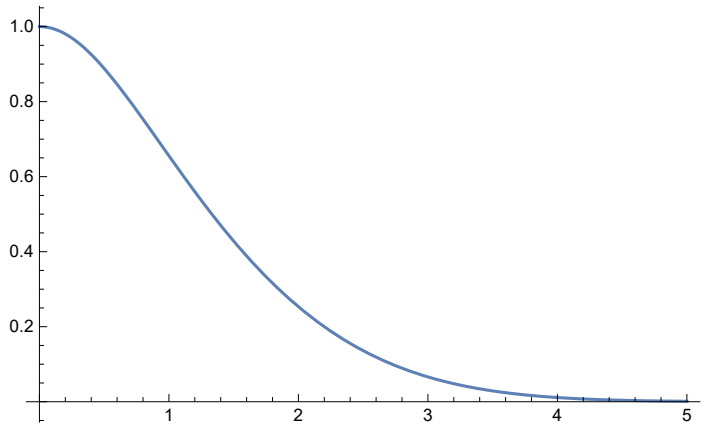
The following MATHEMATICA® script will use the `DSolve` command to solve this ordinary differential equation, then plot the solution.

```
soln = DSolve[{y'[t] == (-t * y[t])/(2 - y[t]), y[0] == 1}, y[t], t]

Plot[y[t]/.soln, {t, 0, 5}]
```

[7] MATHEMATICA® also does numerical computations, and indeed was used to produce the numerical results used in the book.

This gives the output

$$\left\{\left\{y[t] \to -2\text{ProductLog}\left[-\tfrac{1}{2}\sqrt{e^{-1-\frac{t^2}{2}}}\right]\right\}\right\}$$



Note that, as arguments to `DSolve` we give the conditions for a solution to the differential, as well as initial conditions. The syntax `y[t]/.soln` simply means that one should replace `y[t]` with its value as determined by the assignment `soln`. Also, the "`;`" at the end of a MATHEMATICA® command line means that the output will be suppressed. •

While `DSolve` is a useful command, it is also possibe to solve ordinary differential equations using MATHEMATICA® as an assistive tool, rather than just having it belt out solutions.

4.9.2 Example (Solving ordinary differential equations without using `DSolve`) We illustrate Procedure 4.2.18 for the fourth-order equation

$$\frac{\mathrm{d}^4 s}{\mathrm{d}x}(x) - \frac{\mathrm{d}^2 s}{\mathrm{d}x}(x) + 9s(x) = 0.$$

First we must find the roots of the characteristic polynomial.

```
CharPoly = a^4 - 10a^2 + 9 == 0;

roots = Solve[CharPoly, a];
```

Next, we will find the general solution.

```
S1 = C1 * Exp[a * x]/.roots[[1]];

S2 = C2 * Exp[a * x]/.roots[[2]];

S3 = C3 * Exp[a * x]/.roots[[3]];

S4 = C4 * Exp[a * x]/.roots[[4]];

GenSol = S1 + S2 + S3 + S4;
```

Once we have the general solution, we will create a system of equations using the given initial conditions to find the values for $C1$, $C2$, $C3$, and $C4$.

```
A1 = GenSol == 5/.x → 0;

A2 = D[GenSol, x] == −1/.x → 0;

A3 = D[GenSol, {x, 2}] == 21/.x → 0;

A4 = D[GenSol, {x, 3}] == −49/.x → 0;

Const = Solve[{A1, A2, A3, A4}, {C1, C2, C3, C4}];

Sol = GenSol/.Const
```

This gives the solution

$$\{2e^{-3x} - e^{-x} + 4e^{x}\}$$

We can verify this by using `DSolve`:

```
expr = s''''[x] − 10s''[x] + 9s[x] == 0;

DSolve[{expr, s[0] == 5, s'[0] == −1, s''[0] == 21, s'''[0] == −49}, s[x], x]
```

$$\left\{\left\{s[x] \to e^{-3x}\left(2 - e^{2x} + 4e^{4x}\right)\right\}\right\}$$

As you can see, both methods give the same result. •

Let us now work with a particular example with some physical motivation.

4.9.3 Example (Skydiver) Next we will look at another example, this time a second-order equation. Consider a skydiver jumping from a plane. Using Newton's laws of force balance, the governing equation is found to be:

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t}(t) = -a_g + \frac{\rho}{m}\left(\frac{\mathrm{d}y}{\mathrm{d}t}(t)\right)^2.$$

The following script will solve the ordinary differential equation, and plot the jumpers position and velocity for the first twenty seconds.

```
m = 80;

g = 9.81;

p = 1.225;

sol =  DSolve[{y''[t] == −g + (p * y'[t]^2) * (1/m), y[0] == 500, y'[0] == 0}, y[t], t];

a[t] = y[t]/.sol;
```

```
b[t] = D[a[t], t];

position = Plot[a[t], {t, 0, 20}]

velocity = Plot[Evaluate[b[t]], {t, 0, 20}]
```

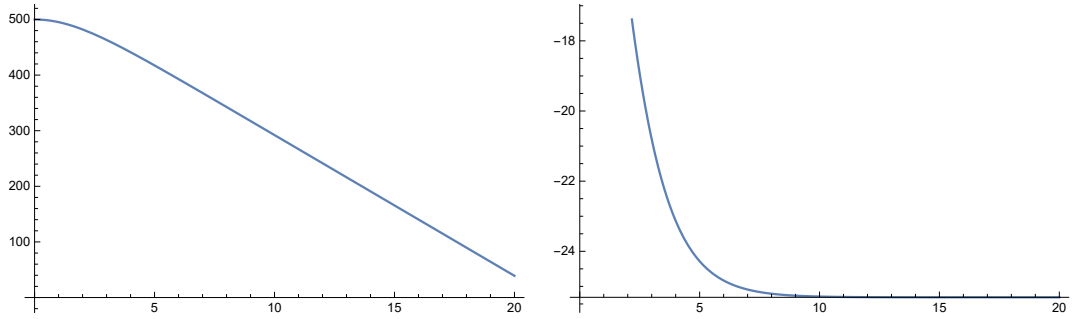


Figure 4.11 Parachuter's position (left) and velocity (right)

As can be seen from the plots, the parachuter's velocity asymptotically reaches a value determined as the inertial forces balance the aerodynamic drag forces. •

In the above examples, we obtained analytical solutions for the differential equations. Typically this is not possible, and one must obtain numerical solutions.

4.9.4 Example (Solving ordinary differential equations numerically) In this example we will show that mathematica also has the ability to solve differential equations numerically as well, again modelling a parachuter jumping from a plane. The `NDSolve` command works very similarly to the `DSolve` command, however it solves the ordinary differential equation, returning a numerical solution. We work again with the parachuter equation

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t}(t) = -a_g + \frac{\rho}{m}\left(\frac{\mathrm{d}y}{\mathrm{d}t}(t)\right)^2.$$

The MATHEMATICA® code is as follows.

```
NumericalSol = NDSolve[{y''[t] == -g + (p * y'[t]^2) * (1/m),
        y[0] == 500, y'[0] == 0}, y, {t, 0, 20}];

Plot[Evaluate[y[t]/.NumericalSol], {t, 0, 20}]

Plot[Evaluate[y'[t]/.NumericalSol], {t, 0, 20}]
```

As you can see, the results are nearly identical when compared to the analytically obtained solutions. •

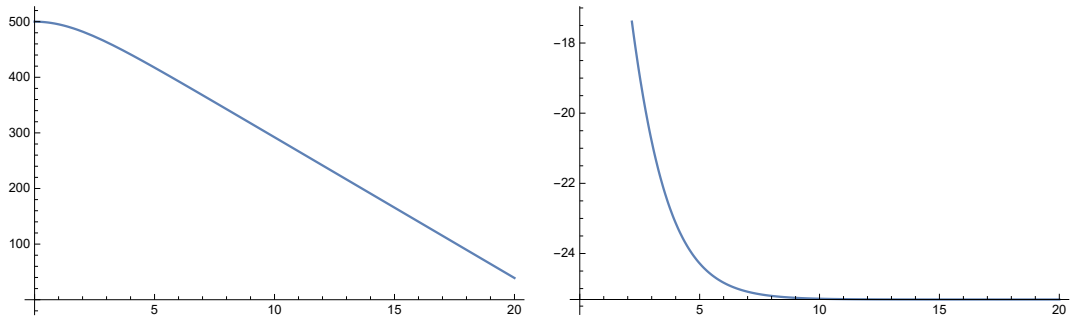


Figure 4.12 Parachuter's position (left) and velocity (right)

4.9.2 Using MATLAB® to obtain numerical solutions

MATLAB® is a very powerful tool for solving complicated differential equations. However, the process is not quite as simple as MATHEMATICA®. To use the `ode45` solver, you must first create a function that is your ordinary differential equation in the form $\frac{dy}{dt}(t) = F(t, y(t))$. This function must then be passed into another script that will solve it. If one types

```
odeexamples
```

at the MATLAB® prompt, you will be given you a list of examples and from these you can easily figure out how to do commonplace things using MATLAB®. To edit an example file named `foo.m`, type

```
edit foo.m
```

To run this file type

```
foo
```

in MATLAB®.

We will now consider the same two examples we covered in the section on MATHEMATICA®.

4.9.5 Example (Solving simple scalar ordinary differential equation) Below is the function that contains the same ordinary differential equation from Exercise 4.9.1. We will pass this into the following main script that will find the solution.

```
1 function [ dydt ] = Example1( t,y )
2
3 dydt = (-t*y)/(2-y);
4
5 end
```

Next we have the main script that will solve this ordinary differential equation. Note that `ode45` has three input arguments: the ordinary differential equation iteslf, time, and initial conditions. The plot that is produced by this script can be found in Figure 4.13.

```
clc
clear all
close all
%% Solving Numerically

t = linspace(0,5);
y0 = 1;

solution = ode45(@(t,y)Example1(t,y),t,y0);

%% Plotting

figure(1)
plot(solution.x,solution.y,'b')
xlabel('Time [s]');
ylabel('y(t)');

print -deps Example1Plot
```

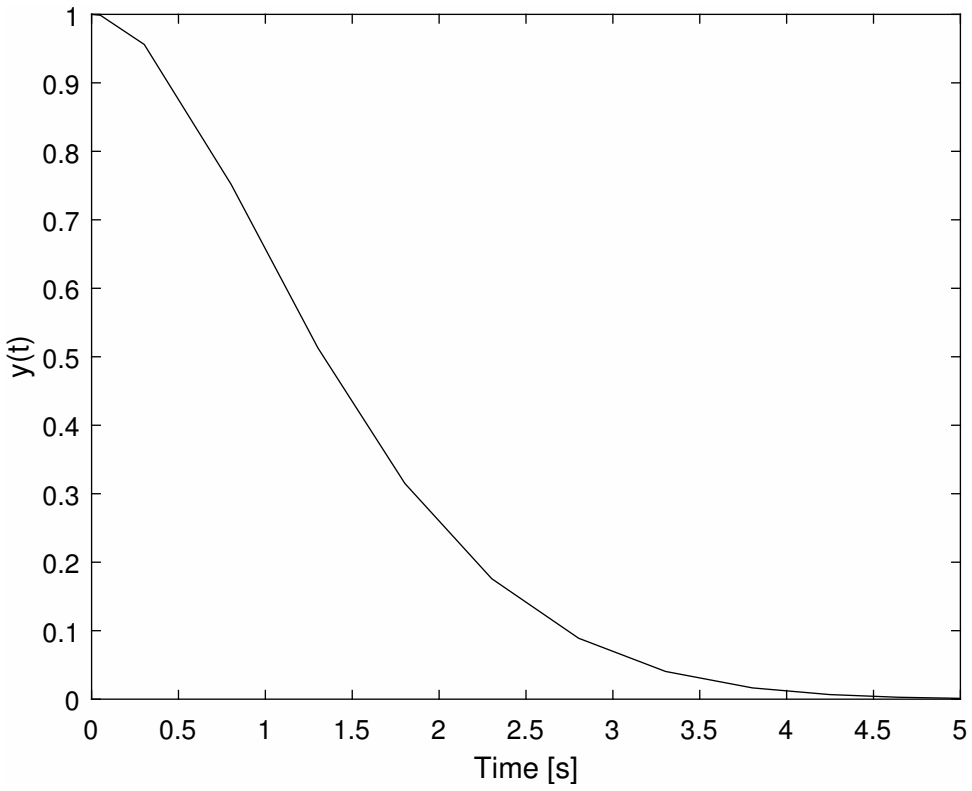


Figure 4.13 Plot generated by MATLAB® for Exercise 4.9.5

Of course, the numerical result here agrees closely with the plot of the analytical result produced in Exercise 4.9.1. •

Next we consider the parachuter example initiated in Exercise 4.9.3.

4.9.6 Example (Skydiver) Next we will consider the same skydiver example as in Exercise 4.9.3. Again we must create a function containing the ordinary differential equation that will then be passed into the main script.

```
function [ dydt ] = Parachute(t,y)

    m = 80; %Mass, in kg, of the parachuter and their gear
    g = 9.81; %Gravitational constant
    p = 1.225; %Density  of  air in kg/m^3

    dydt = [y(2); -g+p*y(2).^2*(1/m)];
end
```

Here is the main script. The plots generated by this script can be found in Figure 4.14.

```
clc
close all
clear all

t = linspace(0,20);

y0 = [500 0];

y = ode45(@(t,y)Parachute(t,y),t,y0);

figure(1)

subplot(2,1,1)
plot(y.x,y.y(1,:))
ylabel('Height [m]');
xlabel('Time [s]');

subplot(2,1,2)
plot(y.x,y.y(2,:))
ylabel('Velocity [m/s]');
xlabel('Time [s]');
```

Again, of course, the numerical results agree with those produced by MATHEMATICA®, both analytically and numerically. •

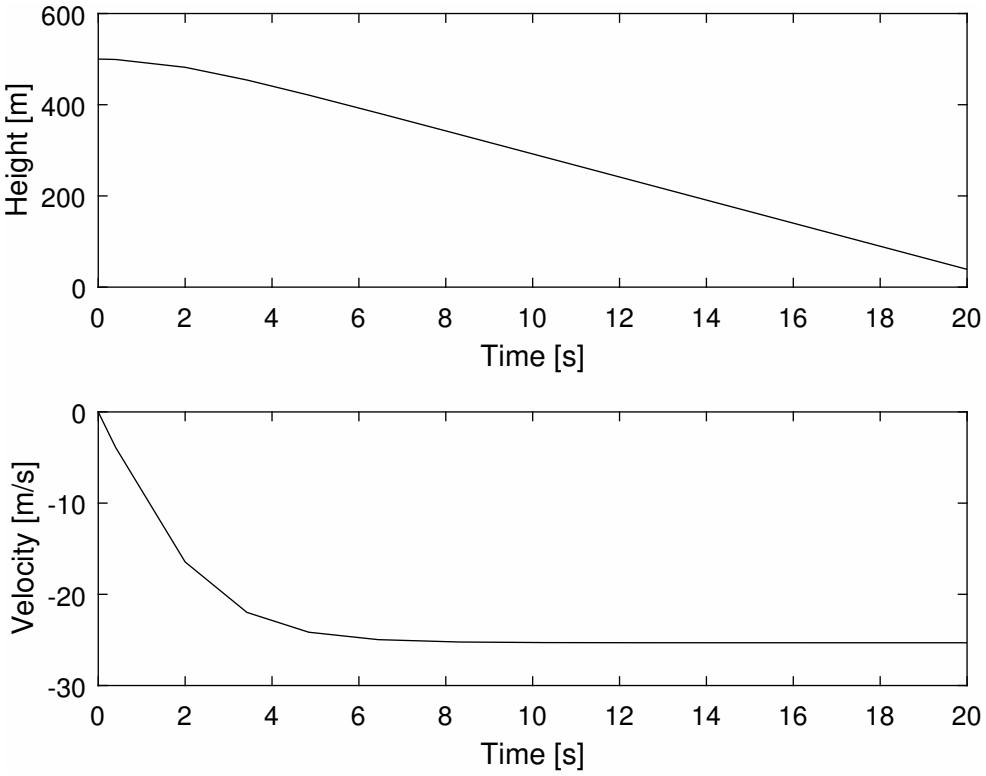


Figure 4.14 Position and velocity graphs of the parachuter Exercise 4.9.6

Chapter 5

Systems of ordinary differential and ordinary difference equations

In this chapter we extend our discussion of scalar differential and difference equations in Chapter 4 to systems of equations. Thus, in the notation of Section 3.1.3, we consider an ordinary differential equation with time-domain $\mathbb{T} \subseteq \mathbb{R}$, state space $U \subseteq \mathbb{R}^m$, and with right-hand side

$$\widehat{F} \colon \mathbb{T} \times U \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

giving the equation

$$\frac{\mathrm{d}^k \xi}{\mathrm{d}t^k}(t) = \widehat{F}\left(t, \xi(t), \frac{\mathrm{d}\xi}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^{k-1}\xi}{\mathrm{d}t^{k-1}}(t)\right)$$

for solutions $t \mapsto \xi(t)$. In the notation of Section 3.3.3, we consider an ordinary difference equation with time-domain $\mathbb{T} \subseteq\subseteq \mathbb{Z}(h)$, state space $U \subseteq \mathbb{R}^m$, and with right-hand side

$$\widehat{F} \colon \mathbb{T} \times U \times \mathrm{L}^{\leq k}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}^m) \to \mathbb{R}^m$$

giving the equation

$$\xi(t + kh) = \widehat{F}\,(t, \xi(t), \xi(t + h), \dots, \xi(t + (k-1)h))$$

for solutions $t \mapsto \xi(t)$. When we studied scalar equations in Chapter 4, we retained this higher-order form of the equations, because doing so allowed us to continue working with scalar equations. However, every scalar equation of order k can be represented as a first-order equation with k unknowns, cf. Exercises 3.1.23 and 3.3.7. In fact, in that exercise we see how to convert a kth-order differential or difference equation in m unknowns into a first-order equation in km unknowns. The point is that, since in this chapter we are working already with vector equations, we will always suppose that our equations are first-order. Also, we will swap around our lettering from Sections 3.1.3 and 3.3.3 and suppose that U is an open subset of $\mathbb{R}^n$. Thus we have a right-hand side

$$\widehat{F} \colon \mathbb{T} \times U \to \mathbb{R}^n$$

and solutions satisfy

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) = \widehat{F}(t, \xi(t))$$

(for differential equations) or

$$\xi(t + h) = \widehat{F}(t, \xi(t))$$

(for difference equations). Note, however, that physically it may still be interesting to retain the higher-order form, even for vector equations, cf. the equation (1.2) modelling a coupled mass-spring system.

As with scalar ordinary differential and difference equations, there is little that one can say in much generality about general systems of ordinary differential or difference equations. Therefore, we focus almost entirely on linear equations in this chapter. One of the reasons that linear systems are so important is that, even for systems that are not linear, a first step towards understanding them is often to linearise them. Thus we shall begin in Section 5.1 with a discussion of linearisation. The next two sections, 5.2 and 5.3, deal with linear systems of ordinary differential equations in some detail. In Section 5.5 we study, essentially, graphical representations for two-dimensional systems of ordinary differential equations, not necessarily linear. While the planar nature of the systems we consider limits the generality of the ideas we discuss, it is nonetheless the case that the ideas seen here form the basis for any serious further study of ordinary differential equations in more advanced treatments of the subject. In Sections 5.6–5.9 we mirror the results for described above, now for difference equations. In Section 5.11 we introduce numerical consideration of systems of ordinary differential equations.

Do I need to read this chapter? Many of the results and techniques in this chapter are prerequisite for our treatments of system theory, particularly in Sections 6.6, 6.8, 6.7, and 6.9. •

Contents

Section 5.1

Linearisation

As we have said, if one is given a completely general system of ordinary differential or difference equations, there is little that one can do. However, sometimes one might be able to find an isolated solution to the differential or difference equation, and then it becomes interesting to know what one can say given this information. The first thing one typically tries is linearisation, i.e., look at the "first-order" variation of solutions from the given solution. In this section we present this method in some detail. We shall not at this point say much about what one can do after linearisation; our main objective is to understand why it might be interesting to focus our attention on linear systems, which is exactly what we do in the subsequent two sections. We shall see in Section 10.5 that linearisation is the foundation of a key set of techniques and results concerning stability.

Do I need to read this section? Some of the results in this section are quite technical, and the technical details are not an essential feature of much of what follows. But, in order to understand the importance of linear systems of differential and difference equations, linearisation is very important, and the reader ought to embrace this. •

5.1.1 Linearisation of ordinary differential equations

We begin by considering linearisation for systems of ordinary differential equations. We first consider linearisation along arbitrary solutions, and then specialise to the case of equilibrium solutions. In practice, it is linearisation about equilibria that is most commonly used in practice, but linearisation along general solutions is something that arises in crucial ways once one starts engaging in more sophisticated undertaking in control theory, none of which shall be done here.

5.1.1.1 Linearisation along solutions Suppose that we have a system of ordinary differential equations $\boldsymbol{F}$ with right-hand side $\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n$ and that we have a solution $\xi_0\colon \mathbb{T}' \to X$ for $\boldsymbol{F}$. We wish to understand what happens to solutions "nearby" this fixed solution ξ_0.

To do this, we suppose that the map

$$\begin{aligned} \widehat{\boldsymbol{F}}_t\colon X &\to \mathbb{R}^n \\ \boldsymbol{x} &\mapsto \widehat{\boldsymbol{F}}(t, \boldsymbol{x}) \end{aligned}$$

is of class C^1. We denote

$$\boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) = \boldsymbol{D}\widehat{\boldsymbol{F}}_t(\boldsymbol{x}), \qquad t \in \mathbb{T}.$$

We then suppose that we have a solution $\xi\colon \mathbb{T}' \to X$ for $\boldsymbol{F}$ for which the deviation $\nu \triangleq \xi - \xi_0$ is small. Let us try to understand the behaviour of ν. Naïvely, we can do this as follows:

$$\dot{\xi}(t) = \frac{\mathrm{d}(\xi_0 + \nu)}{\mathrm{d}t}(t) = \widehat{\boldsymbol{F}}(t, \xi_0(t) + \nu(t)) = \widehat{\boldsymbol{F}}(t, \xi_0(t)) + \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \xi_0(t)) \cdot \nu(t) + \cdots .$$

We will not here try to be precise about what "$\cdots$" might mean, but merely say that the idea of the preceding equation is that we approximate using the constant and first-order terms in the Taylor expansion, and then pray that this gives us something meaningful. Note that, since ξ_0 is a solution for $\boldsymbol{F}$, the approximation we arrive at is

$$\dot{\nu}(t) \approx \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \xi_0(t)) \cdot \nu(t).$$

Meaningful or not, the preceding naïve calculations give rise to the following definition.

5.1.1 Definition (Linearisation of an ordinary differential equation along a solution) Let $\boldsymbol{F}$ be an ordinary differential equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

supposing that $\widehat{\boldsymbol{F}}_t$ is of class C^1 for every $t \in \mathbb{T}$. For a solution $\xi_0\colon \mathbb{T}' \to X$ for $\boldsymbol{F}$, the ***linearisation of*** $\mathbf{F}$ ***along*** $\boldsymbol{\xi_0}$ is the linear ordinary differential equation $\boldsymbol{F}_{L,\xi_0}$ with right-hand side

$$\begin{aligned} \widehat{\boldsymbol{F}}_{L,\xi_0}\colon \mathbb{T}' \times \mathbb{R}^n &\to \mathbb{R}^n \\ (t, v) &\mapsto \boldsymbol{D}\widehat{\boldsymbol{F}}(\xi_0(t)) \cdot v. \end{aligned}$$

•

Note that a solution $t \mapsto \nu(t)$ for the linearisation of $\boldsymbol{F}$ along ξ_0 satisfies

$$\dot{\nu}(t) = \boldsymbol{A}(t)(\nu(t)),$$

where

$$\boldsymbol{A}(t) = \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \xi_0(t)).$$

This is indeed a linear ordinary differential equation. We note that, even when $\boldsymbol{F}$ is autonomous, the linearisation will generally be nonautonomous, due to the dependence of the reference solution ξ_0 on time.

Note that there is an alternative view of linearisation that can be easily developed, one where linearisation is of the *equation*, not just along a solution. The construction we make is the following.

5.1.2 Definition (Linearisation of an ordinary differential equation) Let $\boldsymbol{F}$ be an ordinary differential equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

supposing that $\widehat{\boldsymbol{F}}_t$ is of class C^1 for every $t \in \mathbb{T}$. The ***linearisation*** of $\boldsymbol{F}$ is the ordinary differential equation $\boldsymbol{F}_L$ with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}_L\colon \mathbb{T} \times (X \times \mathbb{R}^n) &\to \mathbb{R}^n \oplus \mathbb{R}^n\\ (t, (\boldsymbol{x}, \boldsymbol{v})) &\mapsto (\widehat{\boldsymbol{F}}(t, \boldsymbol{x}), \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x})(\boldsymbol{v})).\end{aligned}$$

•

Solutions of the linearisation of $\boldsymbol{F}$ are then curves $t \mapsto (\boldsymbol{\xi}(t), \boldsymbol{\nu}(t))$ satisfying

$$\begin{aligned}\dot{\boldsymbol{\xi}}(t) &= \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)),\\ \dot{\boldsymbol{\nu}}(t) &= \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)) \cdot \boldsymbol{\nu}(t).\end{aligned}$$

We see, then, that in this version of linearisation we carry along the original differential equation $\boldsymbol{F}$ as part of the linearisation. This is, in no way, incompatible with the definition of linearisation along a solution $\boldsymbol{\xi}_0$, since one needs $\boldsymbol{F}$ to provide the solution.

Let us illustrate how this works in an example. Finding nonlinear ordinary differential equations whose nontrivial solutions we can explicitly compute is not easy,[1] so we are sort of stuck with systems with one state. However, this will suffice for the illustrative purposes here.

5.1.3 Example (Linearisation of an ordinary differential equation along a solution) We work here with the logistical population model of (1.8). This is the scalar first-order ordinary differential equation with right-hand side

$$\widehat{F}(t, x) = kx(1 - x).$$

Solutions $t \mapsto \xi(t)$, therefore, satisfy

$$\dot{\xi}(t) = k\xi(t)(1 - \xi(t)).$$

This equation is separable and so can be solved using the method from Section 4.1.1. We skip the details, and instead just say that

$$\xi_0(t) = \frac{x_0 \mathrm{e}^{kt}}{1 + x_0(\mathrm{e}^{kt} - 1)}$$

is the solution for F satisfying $\xi_0(0) = x_0$, as long as $x_0 \notin \{0, 1\}$ (we shall consider the cases $x_0 \in \{0, 1\}$ in Example 5.1.7–1). We have

$$\boldsymbol{D}\widehat{F}(t, x) \cdot v = k(1 - 2x)v,$$

[1] We shall see in the next section that working with trivial solutions is easier.

and so the linearisation F_{L,ξ_0} about the solution ξ_0 has the right-hand side

$$\widehat{F}_{L,\xi_0}(t,v) = \frac{k(1 - x_0(e^{kt} + 1))}{1 + x_0(e^{kt} - 1)}v.$$

Thus a solution $t \mapsto v(t)$ for the linearisation satisfies

$$\dot{v}(t) = \underbrace{\frac{k(1 - x_0(e^{kt} + 1))}{1 + x_0(e^{kt} - 1)}}_{a(t)} v(t).$$

This equation can actually be solved, as we saw in Example 4.2.5:

$$v(t) = v_0 e^{-\int_0^t a(\tau)\,d\tau} = v_0 e^{k(t-t_0)} \frac{(1 + x_0(e^{kt_0} - 1))^2}{(1 + x_0(e^{-kt} - 1))^2},$$ [2]

where $v(t_0) = v_0$. Just what conclusions we can draw from this are not clear... nor should they be... The connection between a differential equation and its linearisation is not so clear at the moment. In Section 5.1.1.3 we shall describe the flow of the linearisation in some detail, and in doing so will arrive at a precise interpretation of linearisation. •

5.1.1.2 Linearisation about equilibria In this section we consider what amounts to a special case of linearisation about a solution. The solution we consider is a very particular sort of solution, as given by the following definition.

5.1.4 Definition (Equilibrium state for an ordinary differential equation) Let $\boldsymbol{F}$ be an ordinary differential equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n.$$

A state $\boldsymbol{x}_0 \in X$ is an ***equilibrium state*** if $\widehat{\boldsymbol{F}}(t, \boldsymbol{x}_0) = \boldsymbol{0}$ for every $t \in \mathbb{T}$. •

The following result gives the relationship between equilibrium states and solutions.

5.1.5 Proposition (Equilibrium states and constant solutions) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{X} \to \mathbb{R}^{\mathrm{n}}.$$

A state $\mathbf{x}_0 \in \mathrm{X}$ *is an equilibrium state if and only if the constant function* $\mathrm{t} \mapsto \mathbf{x}_0$ *is a solution for* $\mathbf{F}$.

[2] Integration courtesy of MATHEMATICA®.

Proof Let us denote by ξ_0 the constant function $t \mapsto x_0$.

First suppose that x_0 is an equilibrium state. Then $\dot{\xi}_0(t) = \mathbf{0}$ for every $t \in \mathbb{T}$ and $\widehat{F}(t, \xi_0(t)) = \mathbf{0}$ and so

$$\dot{\xi}_0(t) = \widehat{F}(t, \xi_0(t)), \qquad t \in \mathbb{T},$$

and thus ξ_{x_0} is a solution.

Next suppose that ξ_0 is a solution. Then

$$\mathbf{0} = \dot{\xi}_0(t) = \widehat{F}(t, \xi_0(t)) = \widehat{F}(t, x_0), \qquad t \in \mathbb{T},$$

so giving that x_0 is an equilibrium state. ■

Note that, as a consequence of the preceding simple result, we can linearise about the constant solution $t \mapsto x_0$ in the event that x_0 is an equilibrium state. Let us, however, use some particular language in this case.

5.1.6 Definition (Linearisation of an ordinary differential equation about an equilibrium state) Let F be an ordinary differential equation with right-hand side

$$\widehat{F} \colon \mathbb{T} \times X \to \mathbb{R}^n,$$

supposing that $\widehat{F}_t$ is of class C^1 for every $t \in \mathbb{T}$, and let x_0 be an equilibrium state. The ***linearisation of*** $\mathbf{F}$ ***about*** $\mathbf{x_0}$ is the linear ordinary differential equation F_{L,x_0} with right-hand side

$$\begin{aligned} \widehat{F}_{L,x_0} \colon \mathbb{T} \times \mathbb{R}^n &\to \mathbb{R}^n \\ (t, v) &\mapsto D\widehat{F}(t, x_0) \cdot v. \end{aligned}$$

•

A solution $t \mapsto \nu(t)$ for F_{L,x_0} satisfies

$$\dot{\nu}(t) = A(t)(\nu(t)),$$

where

$$A(t) = D\widehat{F}(t, x_0).$$

Thus we see that the linearisation about an equilibrium point is indeed a linear ordinary differential equation, just as it should be since the same is true of the linearisation about an arbitrary solution. What is special here, however, is that the linearisation is autonomous if F is autonomous. Thus the linearisation when F is autonomous is a linear ordinary differential equation with constant coefficients.

5.1.7 Examples (Linearisation of an ordinary differential equation about an equilibrium state)

1. Let us first return to the linearisation of the logistical population model of Example 5.1.3. We have

$$\widehat{F}(t, x) = kx(1 - x),$$

and so there are two equilibrium states, $x_0 = 0$ and $x_0 = 1$. In Example 5.1.3 we computed the derivative of $\widehat{F}$ to be $\boldsymbol{D}\widehat{F}(t,x) \cdot v = k(1-2x)v$. We thus have the linearisations about $x_0 = 0$ and $x_0 = 1$ given by

$$\widehat{F}_{L,0}(t,v) = kv, \quad \widehat{F}_{L,1}(t,v) = -kv.$$

The solutions then satisfy the equations

$$\dot{v}_0(t) = kv_0(t), \quad \dot{v}_1(t) = -kv_1(t),$$

respectively. These are easily solved using Procedure 4.2.18 to give

$$v_0(t) = v_0(0)\mathrm{e}^{kt}, \quad v_1(t) = v_1(0)\mathrm{e}^{-kt}.$$

We see that we have exponential growth for the solutions of the linearisation about $x_0 = 0$ and exponential decay for the solutions about $x_0 = 1$.
It turns out that this behaviour of the linearisations about the equilibrium state *is* an accurate approximation of the behaviour of the actual system near these states. We do not develop this here, but will address matters such as this in . what?

2. Let us consider the simple pendulum model of (1.3). This is a scalar second-order equation F whose right-hand side is

$$\widehat{F}(t,x,x^{(1)}) = -\frac{a_g}{\ell}\sin(x).$$

In order to fit this differential equation into our linearisation framework, we must convert it into a first-order equation, as in Exercise 3.1.23. Doing this gives the first-order ordinary differential equation $\boldsymbol{F}$ with right-hand side

$$\boldsymbol{F}(t,(x_1,x_2)) = \left(x_2, -\frac{a_g}{\ell}\sin(x)\right).$$

This differential equation has equilibria $\boldsymbol{x}_n = (n\pi, 0)$, $n \in \mathbb{Z}$, corresponding to periodically repeated copies of the "down" and "up" rest positions of the pendulum. We shall work with two of these, $\boldsymbol{x}_0 = (0,0)$ and $\boldsymbol{x}_1 = (\pi,0)$, as they are representative. We compute

$$\boldsymbol{D}\widehat{\boldsymbol{F}}(t,(x_1,x_2)) \cdot (v_1,v_2) = \begin{bmatrix} 0 & 1 \\ -\frac{a_g}{\ell}\cos(x) & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

Now, if we compute this at the two equilibria, we have

$$\boldsymbol{DF}(t,\boldsymbol{x}_0) \cdot \boldsymbol{v} = \begin{bmatrix} 0 & 1 \\ -\frac{a_g}{\ell} & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \quad \boldsymbol{DF}(t,\boldsymbol{x}_1) \cdot \boldsymbol{v} = \begin{bmatrix} 0 & 1 \\ \frac{a_g}{\ell} & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

A solution $t \mapsto \boldsymbol{v}(t)$ of the linearisations satisfies

$$\begin{bmatrix} \dot{v}_1(t) \\ \dot{v}_2(t) \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -\frac{a_g}{\ell} & 0 \end{bmatrix} \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix}, \quad \begin{bmatrix} \dot{v}_1(t) \\ \dot{v}_2(t) \end{bmatrix} \begin{bmatrix} 0 & 1 \\ \frac{a_g}{\ell} & 0 \end{bmatrix} \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix}.$$

It is possible to solve these equation using Procedure 5.2.23 below, and it turns out that the solutions are

$$\begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix} = \begin{bmatrix} \cos(\sqrt{a_g/\ell}t) & \sqrt{\ell/a_g}\sin(\sqrt{a_g/\ell}t) \\ -\sqrt{a_g/\ell}\sin(\sqrt{a_g/\ell}t) & \cos(\sqrt{a_g/\ell}t) \end{bmatrix} \begin{bmatrix} v_1(0) \\ v_2(0) \end{bmatrix},$$
$$\begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix} = \begin{bmatrix} \cosh(\sqrt{a_g/\ell}t) & \sqrt{\ell/a_g}\sinh(\sqrt{a_g/\ell}t) \\ \sqrt{a_g/\ell}\sinh(\sqrt{a_g/\ell}t) & \cosh(\sqrt{a_g/\ell}t) \end{bmatrix} \begin{bmatrix} v_1(0) \\ v_2(0) \end{bmatrix}.$$

In Section 1.1.2 we said a few quite informal things about how this process of linearisation is reflected in the behaviour of the pendulum near the "down" and "up" equilibria. This is reflected in the behaviour of the linearisations, in that, about the "down" equilibrium, the motion for the linearisation is periodic, and, about the "up" equilibrium, the motion diverges from $(0, 0)$ most of the time. We shall be more rigorous about this in . • what?

Summary of linearisation constructions In this section we have illustrated the idea of linearisation in a few different contexts. The take away from these constructions is as follows.

1. The linearisation of an ordinary differential equation $\boldsymbol{F}$ about a solution ξ_0 gives rise to a linear ordinary differential equation that will generally be time-varying, even when $\boldsymbol{F}$ is autonomous.
2. It is possible to linearise an equation with n states in its entirety, to give an ordinary differential equation with $2n$ states.
3. The linearisation of an ordinary differential equation about an equilibrium state gives rise to a linear ordinary differential equation, and this linear equation is autonomous if $\boldsymbol{F}$ is autonomous.
4. At this point, we know nothing about what the linearisation of $\boldsymbol{F}$ says about $\boldsymbol{F}$. However, what is true is that linear ordinary differential equations, even with constant coefficients, arise naturally in the context of linearisation, and so are worthy of some study.

5.1.1.3 The flow of the linearisation

In this section, in contrast with the preceding sections, we give a very precise characterisation of linearisation. It has the benefit of being precise, but the drawback of being complicated. However, the constructions we give in this section are of some importance in subjects like optimal control theory. We shall do three things: (1) provide conditions under which the flow of an ordinary differential equation is differentiable in state and initial time, as well as final time with respect to which it is always differentiable almost everywhere; (2) give explicit formulae for the derivatives; (3) give an interpretation of these derivatives in terms of "wiggling" of initial conditions in state and time.

We shall first investigate thoroughly the properties of the flow of an ordinary differential equation that has more regularity properties than are required for the

basic existence and uniqueness theorem, Theorem 3.2.8. In order to state the result we want, we will make use of some ideas that we will not develop fully until Section 5.2. Let us suppose that we have a system of ordinary differential equations $\boldsymbol{F}$ with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

and let $(t_0, \boldsymbol{x}_0) \in \mathbb{T} \times X$. We then have the solution

$$t \mapsto \xi_0(t) \triangleq \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}_0)$$

defined for $t \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0)$. We then define

$$\begin{aligned} \boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}\colon J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) &\to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n) \\ t &\mapsto \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}_0)). \end{aligned}$$

Now consider the linear time-varying differential equation $\boldsymbol{F}^T_{(t_0,\boldsymbol{x}_0)}$ with right-hand side

$$\begin{aligned} \widehat{\boldsymbol{F}}^T_{(t_0,\boldsymbol{x}_0)}\colon J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) \times \mathbb{R}^n &\to \mathbb{R}^n \\ (t, \boldsymbol{v}) &\mapsto \boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}(t) \cdot \boldsymbol{v}. \end{aligned}$$

To describe solutions of this linear ordinary differential equation, we consider first the following ordinary differential equation. For $t \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) \times \mathbb{R}^n$, we consider the following initial value problem:

$$\frac{\mathrm{d}\boldsymbol{\Psi}}{\mathrm{d}s}(s) = \boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}(s) \circ \boldsymbol{\Psi}(s), \quad \boldsymbol{\Psi}(t) = \boldsymbol{I}_n.$$

As we shall see in the proof of the theorem immediately following, this initial value problem has solutions defined for all $s \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0)$. Moreover, we denote the solution at time s by $\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}}(s, t)$; the associated map

$$\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}}\colon J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) \times J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$$

is what we shall call the "state transition map" in Section 5.6.1.2, and we shall use some of the results from this section in the proof below. In particular, we shall use the fact that the solution to the initial value problem

$$\frac{\mathrm{d}\boldsymbol{\nu}}{\mathrm{d}s}(s) = \boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}(s) \cdot \boldsymbol{\nu}(s), \quad \boldsymbol{\nu}(t) = \boldsymbol{v}_0$$

is

$$\boldsymbol{\nu}(s) = \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}}(s, t) \cdot \boldsymbol{v}_0, \qquad s \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0).$$

With the preceding background, we can now state the theorem.

5.1.8 Theorem (Differentiability of flows for ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times X\to\mathbb{R}^n,$$

and make the following assumptions:

(i) *the map* $t\mapsto\widehat{\mathbf{F}}(t,\mathbf{x})$ *is measurable for each* $\mathbf{x}\in X$*;*

(ii) *the map* $\mathbf{x}\mapsto\widehat{\mathbf{F}}(t,\mathbf{x})$ *is continuously differentiable for each* $t\in\mathbb{T}$*;*

(iii) *for each* $(t,\mathbf{x})\in\mathbb{T}\times X$*, there exist* $r,\rho\in\mathbb{R}_{>0}$ *and*

$$g_0,g_1\in L^1([t_0-\rho,t_0+\rho];\mathbb{R}_{\geq0})$$

such that

(a) $\|\widehat{\mathbf{F}}(s,\mathbf{y})\|\leq g_0(s)$ *for* $(s,\mathbf{y})\in([t_0-\rho,t_0+\rho]\cap\mathbb{T})\times B(r,\mathbf{x})$ *and*

(b) $\left|\dfrac{\partial\widehat{F}_j}{\partial x_k}(s,\mathbf{y})\right|\leq g_1(t)$ *for* $(s,\mathbf{y})\in([t_0-\rho,t_0+\rho]\cap\mathbb{T})\times B(r,\mathbf{x})$ *and* $j,k\in\{1,\dots,n\}$.

Then the following statements hold:

(iv) *for* $t,t_0\in\mathbb{T}$, $\Phi^{\mathbf{F}}_{t,t_0}\colon D_{\mathbf{F}}(t,t_0)\to X$ *is a* C^1*-diffeomorphism onto its image and its derivative is given by* $\mathbf{D}\Phi^{\mathbf{F}}_{t,t_0}(\mathbf{x}_0)=\mathbf{\Phi}_{\mathbf{A}_{(t_0,\mathbf{x}_0)}}(t,t_0)$*;*

(v) *the map*

$$\begin{aligned}\mathbf{D}\Phi^{\mathbf{F}}\colon D_{\mathbf{F}}&\to L(\mathbb{R}^n;\mathbb{R}^n)\\ (t,t_0,\mathbf{x})&\mapsto\mathbf{D}\Phi^{\mathbf{F}}_{t,t_0}(\mathbf{x})\end{aligned}$$

is continuous;

(vi) *for* $(t_0,\mathbf{x}_0)\in\mathbb{T}\times X$*, the set*

$$I_{\mathbf{F}}(t_0,\mathbf{x}_0)=\{t\in\mathbb{T}\mid\ t_0\in J_{\mathbf{F}}(t,\mathbf{x}_0)\}$$

is an open interval, the map

$$\begin{aligned}\iota_{\mathbf{F},t_0,\mathbf{x}_0}\colon I_{\mathbf{F}}(t_0,\mathbf{x}_0)&\to X\\ t&\mapsto\Phi^{\mathbf{F}}(t_0,t,\mathbf{x}_0)\end{aligned}$$

is locally absolutely continuous, and its derivative at a time t *where it is differentiable is given by*

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{\mathbf{F}}(t_0,t,\mathbf{x}_0)=-\mathbf{\Phi}_{\mathbf{A}_{(t_0,\mathbf{x}_0)}}(t_0,t)\cdot\widehat{\mathbf{F}}(t,\mathbf{x}_0).$$

Proof Let us first show that the hypotheses of the theorem imply those of Theorem 3.2.8(ii). Let $\boldsymbol{x}\in X$ and let $r\in\mathbb{R}_{>0}$ and $g_0,g_1\colon\mathbb{T}\to\mathbb{R}_{\geq0}$ be as in the statement of the theorem. To do so, we need to fuss with the manner in which various norms for matrices are related. For $\boldsymbol{A}\in L(\mathbb{R}^n;\mathbb{R}^m)$, we denote by $|||\boldsymbol{A}|||$ the norm induced by the Euclidean norms for $\mathbb{R}^n$ and $\mathbb{R}^m$, as in Section II-1.1.4. Let us also denote

$$\|\boldsymbol{A}\|_\infty=\max\{|A_{jk}|\mid\ j,k\in\{1,\dots,n\}\}.$$

Finally, let us denote by $\|\boldsymbol{A}\|_{\mathrm{Fr}}$ the Frobenius norm of $\boldsymbol{A}$ as in Section II-1.1.5. Let us now make a couple of observations.

1. The Frobenius norm of $\boldsymbol{A}$ is the Euclidean norm of $\boldsymbol{A}$ thought of as a vector in $\mathbb{R}^{nm}$ by listing all of its components.
2. By Proposition II-1.1.11(iv), it follows that
$$\|\boldsymbol{A}\|_{\mathrm{Fr}} \leq \sqrt{nm}\|\boldsymbol{A}\|_{\infty}.$$
3. By Theorem II-1.1.14(v), $|||\boldsymbol{A}|||$ is the largest eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$.
4. By a choice of basis for $\mathbb{R}^n$ in which $\boldsymbol{A}^T\boldsymbol{A}$ is diagonal, we have
$$\|\boldsymbol{A}\|_{\mathrm{Fr}} = \left(\sum_{j=1}^{n} |\lambda_j|^2\right)^2,$$
where $\lambda_1, \dots, \lambda_n$ are the real eigenvalues of $\boldsymbol{A}$.
5. Combining the preceding two observations with Proposition II-1.1.11(vi), we have
$$|||\boldsymbol{A}||| \leq \|\boldsymbol{A}\|_{\mathrm{Fr}}.$$
6. Thus $|||\boldsymbol{A}||| \leq \sqrt{nm}\|\boldsymbol{A}\|_{\infty}$.

Now, for $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathsf{B}(r, \boldsymbol{x})$, the Mean Value Theorem (Theorem II-1.4.38) gives

$$\begin{aligned}\left\|\widehat{\boldsymbol{F}}(t, \boldsymbol{y}_1) - \widehat{\boldsymbol{F}}(t, \boldsymbol{y}_2)\right\| &\leq \sup\{|||\boldsymbol{D}\widehat{\boldsymbol{F}}(\boldsymbol{y})||| \mid \boldsymbol{y} \in \mathsf{B}(r, \boldsymbol{x})\}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| \\ &\leq n g_1(t)\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|,\end{aligned}$$

giving the desired conclusion.

(iv) By virtue of the proof of Theorem 3.2.13 there exists $r, r', \alpha \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{x} \in \mathsf{B}(r, \boldsymbol{x}_0)$ and $t \in [t_0 - \alpha, t_0 + \alpha]$, then $\Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x})$ is defined and takes values in $\mathsf{B}(r', \boldsymbol{x}_0)$. Moreover, we have

$$\Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}) = \boldsymbol{x} + \int_{t_0}^{t} \widehat{\boldsymbol{F}}(s, \Phi^{\boldsymbol{F}}(s, t_0, \boldsymbol{x}))\,\mathrm{d}s$$

in this case. We note that r', r, and α depend on g_0 and L_0 according to the required inequalities

$$\left|\int_{t_0}^{t} g_0(s)\,\mathrm{d}s\right| < \frac{r'}{2}, \quad \left|\int_{t_0}^{t} L_0(s)\,\mathrm{d}s\right| < \lambda$$

for some $\lambda \in (0, 1)$.

By choosing r' and α small enough, there exists $g_1 \in \mathsf{L}^1([t_0 - \alpha, t_0 + \alpha]; \mathbb{R}_{\geq 0})$ such that

$$\left|\frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{x})\right| \leq g_1(t), \qquad (t, \boldsymbol{x}) \in ([t_0 - \alpha, t_0 + \alpha] \cap \mathbb{T}) \times \mathsf{B}(r', \boldsymbol{x}_0).$$

We claim that, if $\boldsymbol{x} \in \mathsf{B}(r, \boldsymbol{x}_0)$, then the ordinary differential equation $\boldsymbol{F}^T_{(t_0, \boldsymbol{x}_0)}$ with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}^T_{(t_0, \boldsymbol{x}_0)} \colon (t_0 - \alpha, t_0 + \alpha) \times \mathbb{R}^n &\to \mathbb{R}^n \\ (t, \boldsymbol{v}) &\mapsto \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x})) \cdot \boldsymbol{v}\end{aligned}$$

possesses unique solutions on $(t_0 - \alpha, t_0 + \alpha)$. To show this, we note by Lemma 1 from the proof of Theorem 3.2.8 that

$$t \mapsto \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}))$$

is locally integrable. Our assertion then follows from Proposition 5.2.2 below.

Now we show that, for each $t \in (t_0 - \alpha, t_0 + \alpha)$, $\Phi^{\boldsymbol{F}}_{t,t_0}$ is differentiable at each $\boldsymbol{x}$ sufficiently close to $\boldsymbol{x}_0$. Let $\rho \in (0, r)$ be small enough that $\mathsf{B}(\rho, \boldsymbol{x}) \subseteq \mathsf{B}(r, \boldsymbol{x}_0)$ for every $\boldsymbol{x} \in \mathsf{B}(r, \boldsymbol{x}_0)$. Let $\boldsymbol{h} \in \mathsf{B}(\rho, \boldsymbol{0})$. By the Fundamental Theorem of Calculus, for $\boldsymbol{x} \in \mathsf{B}(r - \rho, \boldsymbol{x}_0)$, we have

$$\int_0^1 \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x} + s\boldsymbol{h}) \cdot \boldsymbol{h} \, \mathrm{d}s = \widehat{\boldsymbol{F}}(t, \boldsymbol{x} + \boldsymbol{h}) - \widehat{\boldsymbol{F}}(t, \boldsymbol{x}).$$

Therefore,

$$\widehat{\boldsymbol{F}}(t, \boldsymbol{x} + \boldsymbol{h}) - \widehat{\boldsymbol{F}}(t, \boldsymbol{x}) - \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) \cdot \boldsymbol{h} = \int_0^1 (\boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x} + s\boldsymbol{h}) - \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x})) \cdot \boldsymbol{h} \, \mathrm{d}s \tag{5.1}$$

Define

$$M_t(\boldsymbol{h}) = \sup \left\{ \int_0^1 |||\boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x} + s\boldsymbol{h}) - \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x})||| \, \mathrm{d}s \;\middle|\; \boldsymbol{x} \in \mathsf{B}(r - \rho, \boldsymbol{x}_0) \right\},$$

and note that M_t is continuous for $\boldsymbol{h}$ small[3] and that $M_t(\boldsymbol{0}) = 0$. For $\boldsymbol{x} \in \mathsf{B}(r - \rho, \boldsymbol{x}_0)$ and $\boldsymbol{h} \in \mathsf{B}(\rho, \boldsymbol{0})$, consider the initial value problems

$$\dot{\xi}_0(t) = \widehat{\boldsymbol{F}}(t, \xi_0(t)), \quad \xi_0(t_0) = \boldsymbol{x},$$

and

$$\dot{\xi}_1(t) = \widehat{\boldsymbol{F}}(t, \xi_1(t)), \quad \xi_1(t_0) = \boldsymbol{x} + \boldsymbol{h}.$$

Denote $\boldsymbol{\delta}(t) = \xi_1(t) - \xi_0(t)$. We then have

$$\begin{aligned}
\dot{\boldsymbol{\delta}}(t) &= \widehat{\boldsymbol{F}}(t, \xi_0(t) + \boldsymbol{\delta}(t)) - \widehat{\boldsymbol{F}}(t, \xi_0(t)) \\
&= \underbrace{\boldsymbol{D}\widehat{\boldsymbol{F}}(t, \xi_0(t))}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}(t)} \cdot \boldsymbol{\delta}(t) + \underbrace{\int_0^1 (\boldsymbol{D}\widehat{\boldsymbol{F}}(t, \xi_0(t) + s\boldsymbol{\delta}(t)) - \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \xi_0(t))) \cdot \boldsymbol{\delta}(t) \, \mathrm{d}s}_{\boldsymbol{e}(t)},
\end{aligned}$$

[3]The argument here is as follows. We can suppose that we are working in a compact subset of X if $\boldsymbol{h}$ is small, and so the function

$$(\boldsymbol{h}, \boldsymbol{x}) \mapsto \int_0^1 |||\boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x} + s\boldsymbol{h}) - \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x})||| \, \mathrm{d}s$$

is uniformly continuous by the Heine–Cantor Theorem (Theorem II-1.3.33). The continuity of the function obtained by taking the supremum then follows from Exercise II-1.3.4.

using (5.1). Note that

$$\begin{aligned}\|\boldsymbol{e}(t)\| &\leq \int_0^1 \|\widehat{\boldsymbol{DF}}(t, \xi_0(t) + s\boldsymbol{\delta}(t)) - \widehat{\boldsymbol{DF}}(t, \xi_0(t)) \cdot \boldsymbol{\delta}(t)\| \, \mathrm{d}s \\ &\leq \int_0^1 |||\widehat{\boldsymbol{DF}}(t, \xi_0(t) + s\boldsymbol{\delta}(t)) - \widehat{\boldsymbol{DF}}(t, \xi_0(t))||| \, \|\boldsymbol{\delta}(t)\| \, \mathrm{d}s \\ &\leq \|\boldsymbol{\delta}(t)\| M_t(\boldsymbol{\delta}(t)).\end{aligned}$$

Let $\boldsymbol{\nu}$ be the solution to the initial value problem

$$\dot{\boldsymbol{\nu}}(t) = \boldsymbol{A}_{(t_0,\boldsymbol{x})}(t) \cdot \boldsymbol{\nu}(t), \quad \boldsymbol{\nu}(t_0) = \boldsymbol{h}.$$

Now, for fixed $t \in (t_0 - \alpha, t_0 + \alpha)$, we have

$$\boldsymbol{\delta}(t) = \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, t_0) \cdot \boldsymbol{h} + \int_{t_0}^t \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, \tau)\boldsymbol{e}(\tau) \, \mathrm{d}\tau,$$

by Corollary 5.3.3, noting that $\boldsymbol{\delta}(t_0) = \boldsymbol{h}$. Here $\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}$ is the state transition map from Section 5.2.1.2. Thus

$$\boldsymbol{\delta}(t) = \boldsymbol{\nu}(t) + \int_{t_0}^t \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, \tau)\boldsymbol{e}(\tau) \, \mathrm{d}\tau.$$

Thus

$$\begin{aligned}\|\boldsymbol{\delta}(t) - \boldsymbol{\nu}(t)\| &\leq \int_{t_0}^t |||\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, \tau)||| \, \|\boldsymbol{e}(\tau)\| \, \mathrm{d}\tau \leq (t - t_0)|||\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, \cdot)|||_\infty \|\boldsymbol{e}\|_\infty \\ &\leq (t - t_0)|||\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, \cdot)|||_\infty \|\boldsymbol{\delta}(t)\| M_t(\boldsymbol{\delta}(t)),\end{aligned}$$

where the ∞-norm is over the interval $[t_0, t]$. As in the proof of Lemma 2(i) from the proof of Theorem 3.2.13, we have

$$\|\boldsymbol{\delta}(t)\| \leq C\|\boldsymbol{h}\|$$

for some $C \in \mathbb{R}_{>0}$. Therefore,

$$\|\boldsymbol{\delta}(t) - \boldsymbol{\nu}(t)\| \leq C'\|\boldsymbol{h}\| M_t(\boldsymbol{\delta}(t)),$$

where $C' = C\alpha|||\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, \cdot)|||_\infty$. Restoring the pre-abbreviation notation, this reads

$$\frac{\Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x} + \boldsymbol{h}) - \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}) - \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, t_0) \cdot \boldsymbol{h}}{\|\boldsymbol{h}\|} \leq C' M_t(\boldsymbol{\delta}(t)).$$

Since $\lim_{\boldsymbol{h}\to\boldsymbol{0}} \boldsymbol{\delta}(t) = \boldsymbol{0}$ by continuity of solutions with respect to initial conditions and by definition of M_t, we have

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{\Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x} + \boldsymbol{h}) - \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}) - \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x})}}(t, t_0) \cdot \boldsymbol{h}}{\|\boldsymbol{h}\|} = 0,$$

which shows that Φ^F_{t,t_0} is differentiable on $\mathsf{B}(r, x_0)$ and for every $t \in (t_0 - \alpha, t_0 + \alpha)$, and that, moreover, the derivative satisfies the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t} D\Phi^F_{t,t_0}(x) = D\widehat{F}(t, \Phi^F(t, t_0, x)) \circ D\Phi^F_{t,t_0}(x), \quad D\Phi^F_{t_0,t_0}(x) = I_n.$$

Next we show that Φ^F_{t,t_0} is *continuously* differentiable. To show this, let $x \in \mathsf{B}(r, x_0)$ and let ρ be such that $x + h \in \mathsf{B}(\rho, x_0)$. As we showed in the preceding part of the proof,

$$D\Phi^F_{t,t_0}(x + h) = \mathbf{\Phi}_{A_{(t_0,x+h)}}(t, t_0) = I_n + \int_{t_0}^t A_{(t_0,x+h)}(\tau) \circ \mathbf{\Phi}_{A_{(t_0,x+h)}}(\tau, t_0)\,\mathrm{d}\tau.$$

We have

$$\begin{aligned}
|||\mathbf{\Phi}_{A_{(t_0,x+h)}}(t, t_0) - \mathbf{\Phi}_{A_{(t_0,x)}}(t, t_0)|||& \\
&\leq \int_{t_0}^t |||A_{(t_0,x+h)}(\tau) \circ \mathbf{\Phi}_{A_{(t_0,x+h)}}(\tau, t_0) - A_{(t_0,x)}(\tau) \circ \mathbf{\Phi}_{A_{(t_0,x)}}(\tau, t_0)|||\,\mathrm{d}\tau \\
&\leq \int_{t_0}^t |||A_{(t_0,x+h)}(\tau) \circ \mathbf{\Phi}_{A_{(t_0,x+h)}}(\tau, t_0) - A_{(t_0,x+h)}(\tau) \circ \mathbf{\Phi}_{A_{(t_0,x)}}(\tau, t_0)|||\,\mathrm{d}\tau \\
&\quad + \int_{t_0}^t |||A_{(t_0,x+h)}(\tau) \circ \mathbf{\Phi}_{A_{(t_0,x)}}(\tau, t_0) - A_{(t_0,x)}(\tau) \circ \mathbf{\Phi}_{A_{(t_0,x)}}(\tau, t_0)|||\,\mathrm{d}\tau \\
&\leq \int_{t_0}^t g_1(\tau)|||\mathbf{\Phi}_{A_{(t_0,x+h)}}(t, t_0) - \mathbf{\Phi}_{A_{(t_0,x)}}(t, t_0)|||\,\mathrm{d}\tau \\
&\quad + |||\mathbf{\Phi}_{A_{(t_0,x)}}|||_\infty \int_{t_0}^t |||A_{(t_0,x+h)}(\tau) - A_{(t_0,x)}(\tau)|||\,\mathrm{d}\tau.
\end{aligned}$$

By Lemma 1 from the proof of Theorem 3.2.13, we have

$$|||\mathbf{\Phi}_{A_{(t_0,x+h)}}(t, t_0) - \mathbf{\Phi}_{A_{(t_0,x)}}(t, t_0)||| \leq |||\mathbf{\Phi}_{A_{(t_0,x)}}|||_\infty \mathrm{e}^{\int_{t_0}^t g_1(\tau)\,\mathrm{d}\tau} \int_{t_0}^t |||A_{(t_0,x+h)}(\tau) - A_{(t_0,x)}(\tau)|||\,\mathrm{d}\tau.$$

By the Dominated Convergence Theorem,

$$\lim_{h \to 0} \int_{t_0}^t |||A_{(t_0,x+h)}(\tau) - A_{(t_0,x)}(\tau)|||\,\mathrm{d}\tau = 0,$$

which gives

$$\lim_{h \to 0} |||D\Phi^F_{t,t_0}(x + h) - D\Phi^F_{t,t_0}(x)||| = 0,$$

which, for $t \in (t_0 - \alpha, t_0 + \alpha)$, gives the continuity of the derivative of Φ^F_{t,t_0} on $\mathsf{B}(r, x_0)$.

The final part of the proof of the local part of the proof is to show that Φ^F_{t,t_0} is invertible with a continuously differentiable inverse. Let $r', \alpha' \in \mathbb{R}_{>0}$ and let $r \in (0, r']$ and $\alpha \in (0, \alpha']$ as above, and so such that

$$\Phi^F_{t,t_0}(x) \in \mathsf{B}(r', x_0), \qquad x \in \mathsf{B}(r, x_0),\ t \in [t_0 - \alpha, t_0 + \alpha].$$

Let $t \in (t_0 - \alpha, t_0 + \alpha) \cap \mathbb{T}$ and denote

$$V = \Phi^{\boldsymbol{F}}_{t,t_0}(\mathsf{B}(r, \boldsymbol{x}_0)) \subseteq \mathsf{B}(r', \boldsymbol{x}_0).$$

Let $\boldsymbol{x} \in \mathsf{B}(r, \boldsymbol{x}_0)$. Since $\boldsymbol{y} \triangleq \Phi^{\boldsymbol{F}}_{t,t_0}(\boldsymbol{x}) \in \mathsf{B}(r', \boldsymbol{x}_0)$ and $t \in [t_0 - \alpha', t_0 + \alpha'] \cap \mathbb{T}$, there exists $\rho \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{y}' \in \mathsf{B}(\rho, \boldsymbol{y})$, then $(t_0, t, \boldsymbol{y}') \in D_{\boldsymbol{F}}$. Moreover, since $\Phi^{\boldsymbol{F}}_{t_0,t}$ is continuous (indeed, continuously differentiable) and $\Phi^{\boldsymbol{F}}_{t_0,t}(\boldsymbol{y}) = \boldsymbol{x}$, we may choose ρ sufficiently small that $\Phi^{\boldsymbol{F}}_{t_0,t}(\boldsymbol{y}') \in \mathsf{B}(r, \boldsymbol{x}_0)$ if $\boldsymbol{y}' \in \mathsf{B}(\rho, \boldsymbol{y})$. By the preceding part of the proof, $\Phi^{\boldsymbol{F}}_{t_0,t}|\mathsf{B}(\rho, \boldsymbol{y})$ is continuously differentiable. Thus there is a neighbourhood of $\boldsymbol{x}$ on which the restriction of $\Phi^{\boldsymbol{F}}_t t_0$ is invertible, continuously differentiable, and with a continuously differentiable inverse.

To complete this part of the proof, we need to prove the statement globally. To this end, let $(t_0, \boldsymbol{x}_0) \in \mathbb{T} \times X$ and denote by $J_+(t_0, \boldsymbol{x}_0) \subseteq \mathbb{T}$ the set of $b > t_0$ such that, for each $b' \in [t_0, b)$, there exists a relatively open interval $J \subseteq \mathbb{T}$ and a $r \in \mathbb{R}_{>0}$ such that

1. $b' \in J$,
2. $J \times \{t_0\} \times \mathsf{B}(r, \boldsymbol{x}_0) \subseteq D_{\boldsymbol{F}}$, and
3. for each $t \in J$, $\mathsf{B}(r, \boldsymbol{x}_0) \ni \boldsymbol{x} \mapsto \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x})$ is a C^1-diffeomorphism onto its image.

By the local part of the proof above, $J_+(t_0, \boldsymbol{x}_0) \neq \varnothing$. We then consider two cases.

The first case is $J_+(t_0, \boldsymbol{x}_0) \cap [t_0, \infty) = \mathbb{T} \cap [t_0, \infty)$. In this case, for each $t \in \mathbb{T}$ with $t \geq t_0$, there exists a relatively open interval $J \subseteq \mathbb{T}$ and $r \in \mathbb{R}_{>0}$ such that

1. $t \in J$,
2. $J \times \{t_0\} \times \mathsf{B}(r, \boldsymbol{x}_0) \subseteq D_{\boldsymbol{F}}$, and
3. for each $\tau \in J$, $\mathsf{B}(r, \boldsymbol{x}_0) \ni \boldsymbol{x} \mapsto \Phi^{\boldsymbol{F}}(\tau, t_0, \boldsymbol{x})$ is a C^1-diffeomorphism onto its image.

The second case is $J_+(t_0, \boldsymbol{x}_0) \cap [t_0, \infty) \subset \mathbb{T} \cap [t_0, \infty)$. In this case we let $t_1 = \sup J_+(t_0, \boldsymbol{x}_0)$ and note that $t_1 \neq \sup \mathbb{T}$. We claim that $t_1 \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0)$. Were this not the case, then we must have $b \triangleq \sup J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) < t_1$. Since $b \in J_+(t_0, \boldsymbol{x}_0)$, there must be a relatively open interval $J \subseteq \mathbb{T}$ containing b such that $t \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0)$ for all $t \in J$. But, since there are t's in J larger than b, this contradicts the definition of $J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0)$, and so we conclude that $t_1 \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0)$. Let us denote $\boldsymbol{x}_1 = \Phi^{\boldsymbol{F}}(t_1, t_0, \boldsymbol{x}_0)$. By our local conclusions from the first part of the proof, there exists $\alpha_1, r_1 \in \mathbb{R}_{>0}$ such that $(t, t_1, \boldsymbol{x}) \in D_{\boldsymbol{F}}$ for every $t \in (t_1 - \alpha_1, t_1 + \alpha_1)$ and $\boldsymbol{x} \in \mathsf{B}(r_1, \boldsymbol{x}_1)$, and such that the map

$$\mathsf{B}(r_1, \boldsymbol{x}_1) \ni \boldsymbol{x} \mapsto \Phi^{\boldsymbol{F}}(t, t_1, \boldsymbol{x})$$

is a C^1-diffeomorphism onto its image for every $t \in (t_1 - \alpha_1, t_1 + \alpha_1)$. Since $t \mapsto \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}_0)$ is continuous and $\Phi^{\boldsymbol{F}}(t_1, t_0, \boldsymbol{x}_0) = \boldsymbol{x}_1$, let $\delta \in \mathbb{R}_{>0}$ be such that $\delta < \frac{\alpha_1}{2}$ and $\Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}_0) \in \mathsf{B}(r_1/4, \boldsymbol{x}_1)$ for $t \in (t_1 - \delta, t_1)$. Now let $\tau_1 \in (t_1 - \delta, t_1)$ and, by our hypotheses on t_1, there exists an open interval J and $r_1' \in \mathbb{R}_{>0}$ such that

1. $\tau_1 \in J$,
2. $J \times \{t_0\} \times \mathsf{B}(r_1', \boldsymbol{x}_0) \subseteq D_{\boldsymbol{F}}$, and
3. for each $\tau \in J$, $\mathsf{B}(r_1', \boldsymbol{x}_0) \ni \boldsymbol{x} \mapsto \Phi^{\boldsymbol{F}}(\tau, t_0, \boldsymbol{x})$ is a C^1-diffeomorphism onto its image.

We also choose J and r_1' sufficiently small that

$$\{\Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}) \mid \ t \in J, \ \boldsymbol{x} \in \mathsf{B}(r_1', \boldsymbol{x}_0)\} \subseteq \mathsf{B}(r_1/2, \boldsymbol{x}_1).$$

Now we claim that

$$(\tau_1 - \alpha_1, \tau_1 + \alpha_1) \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F.$$

We first show that

$$[\tau_1, \tau_1 + \alpha_1) \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F. \tag{5.2}$$

Indeed, we have $(\tau_1, t_0, x) \in D_F$ for every $x \in \mathsf{B}(r_1', x_0)$ since $\tau_1 \in J$. By definition of J, $\Phi^F(\tau_1, t_0, x) \in \mathsf{B}(r_1/2, x_1)$. By definition of τ_1, $t_1 - \tau_1 < \delta < \frac{\alpha_1}{2}$. Then, by definition of α_1 and r_1,

$$(t_1, \tau_1, \Phi^F(\tau_1, t_0, x)) \in D_F$$

for every $x \in \mathsf{B}(r_1', x_0)$. From this we conclude that $(t_1, t_0, x) \in D_F$ for every $x \in \mathsf{B}(r_1', x_0)$. Now, since

$$t \in [\tau_1, \tau_1 + \alpha_1) \quad \Longrightarrow \quad t \in (t_1 - \alpha_1, t_1 + \alpha_1),$$

we have $(t, t_1, \Phi^F(t, t_1, x)) \in D_F$ for every $t \in (\tau_1 - \alpha_1, \tau_1 + \alpha_1)$ and $x \in \mathsf{B}(r_1', x_0)$. Since

$$\Phi^F(t, t_1, \Phi^F(t_1, t_0, x)) = \Phi^F(t, t_0, x),$$

we conclude (5.2). A similar but less complicated argument gives

$$(\tau_1 - \alpha_1, \tau_1) \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F.$$

Next we claim that the map

$$\mathsf{B}(r_1', x_0) \ni x \mapsto \Phi^F(t, t_0, x)$$

is a C^1-diffeomorphism onto its image for every $t \in (\tau_1 - \alpha_1, \tau_1 + \alpha_1)$. By definition of τ_1, the map

$$\Phi^F_{t,t_0} \colon \mathsf{B}(r_1', x_0) \to \mathsf{B}(r_1/2, x_1)$$

is a C^1-diffeomorphism onto its image for $t \in (\tau_1 - \alpha_1, \tau_1]$. We also have that

$$\Phi^F_{t,\tau_1} \colon \mathsf{B}(r_1, x_1) \to X$$

is a C^1-diffeomorphism onto its image for $t \in (\tau_1, \tau_1 + \alpha_1)$. Since the composition of C^1-diffeomorphisms onto their image is a C^1-diffeomorphism onto its image, our assertion follows.

By our above arguments, we have an open interval J' and $r_1' \in \mathbb{R}_{>0}$ such that

1. $t_1 \in J'$,
2. $J' \times \{t_0\} \times \mathsf{B}(r_1', x_0) \subseteq D_F$, and
3. for each $t \in J'$, $\mathsf{B}(r_1', x_0) \ni x \mapsto \Phi^F(t, t_0, x)$ is a C^1-diffeomorphism onto its image.

This contradicts the fact that $t_1 = \sup J_+(t_0, x_0)$ and so the condition

$$J_+(t_0, x_0) \cap [t_0, \infty) \subset \mathbb{T} \cap [t_0, \infty)$$

cannot obtain.

One similarly shows that it must be the case that $J_-(t_0, x_0) \cap (-\infty, t_0] = \mathbb{T} \cap (-\infty, t_0]$, where $J_-(t_0, x_0)$ has the obvious definition.

Now we note that Φ^F_{t,t_0} injective by uniqueness of solutions for $\boldsymbol{F}$. Now, assertion (iv) of the theorem now follows since the notion of "C^1-diffeomorphism" can be tested locally, i.e., in a neighbourhood of any point.

To conclude, we must show that the derivative satisfies the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{D}\Phi^F(t,t_0,\boldsymbol{x}_0) = \boldsymbol{D}\widehat{\boldsymbol{F}}(t,\Phi^F(t,t_0,\boldsymbol{x}_0))\circ\boldsymbol{D}\Phi^F(t,t_0,\boldsymbol{x}_0),\quad \boldsymbol{D}\Phi^F(t_0,t_0,\boldsymbol{x}_0)=\boldsymbol{I}_n,$$

on $J_F(t_0,\boldsymbol{x}_0)$. Let $J_+(t_0,\boldsymbol{x}_0)$ (reusing the notation from the preceding part of the proof) be the set of $t\geq t_0$ such that $\tau\mapsto \boldsymbol{D}\Phi^F(\tau,t_0,\boldsymbol{x}_0)$ satisfies the preceding initial value problem on $[t_0,t_1]$. Note that $J_+(t_0,\boldsymbol{x}_0)\neq\emptyset$ by our arguments in the first part of the proof. Let $t_1=\sup J_+(t_0,\boldsymbol{x}_0)$. We claim that $t_1=\sup J_F(t_0,\boldsymbol{x}_0)$. If $t_1=t_0$ there is nothing to prove. So suppose that $t_1>t_0$ and suppose that $t_1\neq\sup J_F(t_0,\boldsymbol{x}_0)$. Therefore, $t_1\in J_F(t_0,\boldsymbol{x}_0)$ and so there exists $\alpha_1\in\mathbb{R}_{>0}$ such that $(t_1-\alpha_1,t_1+\alpha_2)\subseteq J_F(t_0,\boldsymbol{x}_0)$. Let $\boldsymbol{x}_1=\Phi^F(t_1,t_0,\boldsymbol{x}_0)$. Note that our arguments from the first part of the proof show that, on $(t_1-\alpha_1,t_1+\alpha_1)$, $t\mapsto\boldsymbol{D}\Phi^F(t,t_1,\boldsymbol{x}_1)$ satisfies the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{D}\boldsymbol{F}(t,t_1,\boldsymbol{x}_1) = \boldsymbol{D}\widehat{\boldsymbol{F}}(t,\Phi^F(t,t_1,\boldsymbol{x}_1))\circ\boldsymbol{D}\Phi^F(t,t_1,\boldsymbol{x}_1),\quad \boldsymbol{D}\Phi^F(t_1,t_1,\boldsymbol{x}_1)=\boldsymbol{I}_n.$$

Now define $\boldsymbol{\Xi}\colon[t_0,t_1+\alpha_1)\to\mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$ by

$$\boldsymbol{\Xi}(t)=\begin{cases}\boldsymbol{D}\Phi^F(t,t_0,\boldsymbol{x}_0), & t\in[t_0,t_1],\\ \boldsymbol{D}\Phi^F(t,t_1,\boldsymbol{x}_1), & t\in(t_1,t_1+\alpha_1).\end{cases}$$

As we showed in the first part of the proof, if we denote $\boldsymbol{A}(t)=\boldsymbol{D}\widehat{\boldsymbol{F}}(t,\Phi^F(t,t_0,\boldsymbol{x}_0))$, then, since

$$\Phi^F(t,t_0,\boldsymbol{x}_0)=\Phi^F(t,t_1,\Phi^F(t_1,t_0,\boldsymbol{x}_0))=\Phi^F(t,t_1,\boldsymbol{x}_1)$$

for $t\in[t_1,t_1+\alpha_1)$, we have $\boldsymbol{\Xi}(t)=\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}}(t,t_0)$ for $t\in[t_0,t_1+\alpha_1)$. Thus we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{D}\boldsymbol{F}(t,t_1,\boldsymbol{x}_1) = \boldsymbol{D}\widehat{\boldsymbol{F}}(t,\Phi^F(t,t_1,\boldsymbol{x}_1))\circ\boldsymbol{D}\Phi^F(t,t_1,\boldsymbol{x}_1),\quad \boldsymbol{D}\Phi^F(t_1,t_1,\boldsymbol{x}_1)=\boldsymbol{I}_n,$$

on $[t_0,t_1+\alpha_1)$, which contradicts the definition of $J_+(t_0,\boldsymbol{x}_0)$. Thus we must have $t_1=\sup J_F(t_0,\boldsymbol{x}_0)$. A similar argument can be made for $t<t_0$, and we have thus completed this part of the proof.

(v) Let us consider the ordinary differential equation $\boldsymbol{F}_1$ with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}_1\colon\mathbb{T}\times X\times\mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)&\to\mathbb{R}^n\times\mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)\\ (t,\boldsymbol{x},\boldsymbol{X})&\mapsto(\widehat{\boldsymbol{F}}(t,\boldsymbol{x}),\boldsymbol{D}\widehat{\boldsymbol{F}}(t,\boldsymbol{x})\circ\boldsymbol{X}).\end{aligned}$$

This ordinary differential equation satisfies the conditions of Theorem 3.2.8(ii). Moreover, as we saw from the previous part of the proof, $J_{F_1}(t_0,(\boldsymbol{x}_0,\boldsymbol{X}_0))=J_F(t_0,\boldsymbol{x}_0)$ for every $\boldsymbol{X}_0\in\mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$. Thus

$$D_{F_1}=\{(t,t_0,(\boldsymbol{x}_0,\boldsymbol{X}_0))\mid(t,t_0,\boldsymbol{x}_0)\in D_F\}.$$

From Theorem 3.2.13 we know that Φ^{F_1} is continuous. Moreover, from the first part of the proof,

$$\Phi^{F_1}(t,t_0,(\boldsymbol{x}_0,\boldsymbol{X}_0))=(\Phi^F(t,t_0,\boldsymbol{x}_0),\boldsymbol{D}\Phi^F_{t,t_0}(\boldsymbol{x}_0)\circ\boldsymbol{X}).$$

From this, the desired conclusion follows.

(vi) We will show something more than is stated in this part of the theorem. The setup we will make is the following. We suppose that we have $a, b \in \mathbb{T}$ with $a < b$ and $x_0 \in X$, and we suppose that, for some $\rho \in \mathbb{R}_{>0}$, we have a solution

$$[a - \rho, b + \rho] \ni t \mapsto \Phi^F(t, a, x_0).$$

Let us abbreviate $\xi_0(t) = \Phi^F(t, a, x_0)$. Then, according to Theorem 3.2.13, there exists $r \in \mathbb{R}_{>0}$ such that, if $\tau \in [a, b]$ and if $(t, x) \in (\tau - r, \tau + r) \times \mathsf{B}(r, \xi_0(\tau))$, then the solution

$$s \mapsto \Phi^F(s, t, x)$$

is defined for $s \in [a - \rho, b + \rho]$.[4] We denote

$$W_r = \cup_{\tau \in [a,b]} (\tau - r, \tau + r) \times \mathsf{B}(r, \xi_0(\tau)).$$

We shall show that, for $t_0, t_1 \in [a, b]$, if $x_0 = \xi_0(t_0)$ and if ξ_0 is differentiable at t_0, then the function

$$W_r \ni (t, x) \mapsto \Phi^F(t_1, t, x)$$

is differentiable at (t_0, x_0), and that its derivative is the linear map

$$(\sigma, v) \mapsto \mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1, t_0) \cdot (v - \sigma \dot{\xi}_0(t_0)).$$

This implies the conclusions of the theorem, since the conclusions of the theorem are only about the function of t, not of t and x.

We make some preliminary constructions. Let $B \in \mathbb{R}_{>0}$ be such that

$$\|\mathbf{\Phi}_{A_{(t_0,\xi_0(t_0))}}(t_1, t_0) \cdot v\| \leq B\|v\|, \qquad t_1, t_0 \in [a - \rho, b + \rho],$$

this being possible by part (v). Now define

$$\sigma(\tau) = \sup\{\|\mathbf{\Phi}_{A_{(t_0+\tau,x_0)}}(t_1, t_0 + \tau) - \mathbf{\Phi}_{A_{(t_0,x_0)}(t_1,t_0)}(t_1, t_0)\| \mid \ t_0, t_1 \in [a, b]\}.$$

By uniform continuity, σ is continuous and $\lim_{\tau \to 0} \sigma(\tau) = 0$. Now let $t_0, t_1 \in [a, b]$, let $x_0 = \xi_0(t_0)$, and suppose that ξ_0 is differentiable at t_0. Denote

$$v_0(\tau) = \frac{\|\xi_0(t_0 + \tau) - \xi_0(t_0) - \tau \dot{\xi}_0(t_0)\|}{|\tau|},$$

and note that v_0 is continuous for small τ and that $\lim_{\tau \to 0} v_0(\tau) = 0$. Next denote

$$D(\tau, h) = \sup \left\{ \frac{\|\Phi^F(t_1, t_0 + \tau, x_0 + h) - \Phi^F(t_1, t_0, x_0) - \mathbf{\Phi}_{A_{(t_0+\tau,x_0)}}(t_1, t_0 + \tau) \cdot h\|}{\|h\|} \;\middle|\; t_1 \in [a, b] \right\}.$$

Note that D is continuous and that $\lim_{(\tau,h) \to (0,0)} D(\tau, h) = 0$.

[4] The existence of such $r \in \mathbb{R}_{>0}$ follows from a compactness argument, using compactness of $\{(\tau, \xi_0(\tau)) \mid \ \tau \in [a, b]\}$.

Now we estimate

$$\begin{aligned}
&\|\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0+\tau,\boldsymbol{x}_0)}}(t_1,t_0+\tau)\cdot(\boldsymbol{x}_0+\boldsymbol{h}-\boldsymbol{\xi}_0(t_0+\tau))-\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}}(t_1,t_0)\cdot(\boldsymbol{h}-\tau\dot{\boldsymbol{\xi}}_0(t_0))\| \\
&\qquad\le \|\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0+\tau,\boldsymbol{x}_0)}}(t_1,t_0+\tau)\cdot(\boldsymbol{\xi}_0(t_0+\tau)-\boldsymbol{x}_0-\tau\dot{\boldsymbol{\xi}}_0(t_0))\| \\
&\qquad\quad + \|\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0+\tau,\boldsymbol{x}_0)}}(t_1,t_0+\tau)\cdot(\boldsymbol{h}-\tau\dot{\boldsymbol{\xi}}_0(t_0))-\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}}(t_1,t_0)\cdot(\boldsymbol{h}-\tau\dot{\boldsymbol{\xi}}_0(t_0))\| \\
&\qquad\le f_1(\tau)(|\tau|+\|\boldsymbol{h}\|),
\end{aligned}$$

where

$$f_1(\tau)=Bv_0(\tau)+(1+\|\dot{\boldsymbol{\xi}}_0(t_0)\|)\sigma(\tau).$$

Note that f_1 is continuous for small τ and $\lim_{\tau\to 0} f_1(\tau)=0$.

Now we estimate

$$\begin{aligned}
&\|\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{x}_0+\boldsymbol{h})-\Phi^{\boldsymbol{F}}(t_1,t_0,\boldsymbol{x}_0)-\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0+\tau,\boldsymbol{x}_0)}}(t_1,t_0+\tau)\cdot(\boldsymbol{x}_0+\boldsymbol{h}-\boldsymbol{\xi}_0(t_0+\tau))\| \\
&\qquad= \|\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{x}_0+\boldsymbol{h})-\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{\xi}_0(t_0+\tau)) \\
&\qquad\quad -\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0+\tau,\boldsymbol{x}_0)}}(t_1,t_0+\tau)\cdot(\boldsymbol{x}_0+\boldsymbol{h}-\boldsymbol{\xi}_0(t_0+\tau))\| \\
&\qquad\le \|\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{x}_0+\boldsymbol{h})-\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{x}_0)-\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0+\tau,\boldsymbol{x}_0)}}(t_1,t_0+\tau)\cdot\boldsymbol{h}\| \\
&\qquad\quad + \|\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{\xi}_0(t_0+\tau))-\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{x}_0) \\
&\qquad\quad -\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0+\tau,\boldsymbol{x}_0)}}(t_1,t_0+\tau)\cdot(\boldsymbol{\xi}_0(t_0+\tau)-\boldsymbol{x}_0)\| \\
&\qquad\le D(\tau,\boldsymbol{h})(|\tau|+\|\boldsymbol{h}\|)+D(\tau,\boldsymbol{\xi}_0(t_0+\tau)-\boldsymbol{x}_0)(|\tau|+\|\boldsymbol{\xi}_0(t_0+\tau)-\boldsymbol{x}_0\|).
\end{aligned}$$

By Taylor's Theorem, we have

$$\boldsymbol{\xi}_0(t_0+\tau)-\boldsymbol{x}_0=\tau(R(\tau)+\dot{\boldsymbol{\xi}}(t_0))$$

for a continuous function R for which $\lim_{\tau\to 0} R(\tau)=0$. Thus, for small τ,

$$\begin{aligned}
&\|\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{x}_0+\boldsymbol{h})-\Phi^{\boldsymbol{F}}(t_1,t_0,\boldsymbol{x}_0)-\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0+\tau,\boldsymbol{x}_0)}}(t_1,t_0+\tau)\cdot(\boldsymbol{x}_0+\boldsymbol{h}-\boldsymbol{\xi}_0(t_0+\tau))\| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\le f_2(\tau,\boldsymbol{h})(|\tau|+\|\boldsymbol{h}\|),
\end{aligned}$$

where

$$f_2(\tau,\boldsymbol{h})=D(\tau,\boldsymbol{h})+(1+\|\dot{\boldsymbol{\xi}}(t_0)\|)D(\tau,\boldsymbol{\xi}_0(t_0+\tau)-\boldsymbol{x}_0).$$

We note that f_2 is continuous and that $\lim_{(\tau,\|\boldsymbol{h}\|)} f_2(\tau,\boldsymbol{h})=0$.

Combining the preceding two estimates we have

$$\begin{aligned}
&\|\Phi^{\boldsymbol{F}}(t_1,t_0+\tau,\boldsymbol{x}_0+\boldsymbol{h})-\Phi^{\boldsymbol{F}}(t_1,t_0,\boldsymbol{x}_0)-\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}}(t_1,t_0)\cdot(\boldsymbol{h}-\tau\dot{\boldsymbol{\xi}}_0(t_0))\| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\le (f_1(\tau)+f_2(\tau,\boldsymbol{h}))(|\tau|+\|\boldsymbol{h}\|).
\end{aligned}$$

We thus conclude this part of the theorem. ∎

The proof of the theorem immediately gives the following result.

5.1.9 Corollary (Flow of ordinary differential equations of class C^1) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times X \to \mathbb{R}^n.$$

If $\widehat{\mathbf{F}}$ *is of class* C^1*, then* $\Phi^{\mathbf{F}}\colon D_{\mathbf{F}} \to X$ *is of class* C^1.

Proof From the proof of part (vi) of the preceding theorem, we have

$$\|\Phi^F(t_1, t_0+\tau_0, x_0+h) - \Phi^F(t_1,t_0,x_0) - \mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1,t_0)\cdot(h-\tau_0\dot{\xi}_0(t_0))\| \leq f(\tau_0,h)(|\tau_0|+\|h\|)$$

for a continuous function f satisfying $\lim_{(\tau_0,h)\to(0,0)} f(\tau_0,h) = 0$. Note that, under the hypotheses of the corollary, this conclusion holds for every $(t_1,t_0,x_0) \in D_F$ since solutions for F are of class C^1 in this case.

Now we have

$$\begin{aligned}\Phi^F(t_1+\tau_1,t_0+\tau_0,x_0+h) &- \Phi^F(t_1,t_0,x_0)\\ &= \Phi^F(t_1+\tau_1,t_0+\tau_0,x_0+h) - \Phi^F(t_1+\tau_1,t_0,x_0)\\ &\quad + \Phi^F(t_1+\tau_1,t_0,x_0) - \Phi^F(t_1,t_0,x_0).\end{aligned}$$

This then gives

$$\begin{aligned}&\|\Phi^F(t_1+\tau_1,t_0+\tau_0,x_0+h) - \Phi^F(t_1,t_0,x_0) - \mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1,t_0)\cdot(h-\tau_0\dot{\xi}_0(t_0)) - \tau_1\dot{\xi}_0(t_1)\|\\ &\leq \|\Phi^F(t_1+\tau_1,t_0+\tau_0,x_0+h) - \Phi^F(t_1+\tau_1,t_0,x_0) - \mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1+\tau_1,t_0)\cdot(h-\tau_0\dot{\xi}_0(t_0))\|\\ &\quad + \left|\left|\left|\mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1+\tau_1,t_0) - \mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1+\tau_1,t_0)\right|\right|\right| \|h-\tau_0\dot{\xi}_0(t_0)\|\\ &\quad + \|\Phi^F(t_1+\tau_1,t_0,x_0) - \Phi^F(t_1,t_0,x_0) - \tau_1\dot{\xi}_0(t_1)\|\end{aligned}$$

Arguments like those from the proof of part (vi) of the preceding theorem then give

$$\|\Phi^F(t_1+\tau_1,t_0+\tau_0,x_0+h) - \Phi^F(t_1,t_0,x_0) - \mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1,t_0)\cdot(h-\tau_0\dot{\xi}_0(t_0)) - \tau_1\dot{\xi}_0(t_1)\| \leq f(\tau_1,\tau_0,h)(|\tau_1|+|\tau_0|+\|h\|),$$

where f is a continuous function satisfying

$$\lim_{(\tau_1,\tau_0,h)\to(0,0,0)} f(\tau_1,\tau_0,h) = 0.$$

From this we conclude that the Φ^F is differentiable, and, moreover, that the derivative at $(t_1,t_0,x_0) \in D_F$ is given by the linear map

$$(\sigma_1,\sigma_0,v) \mapsto \mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1,t_0)\cdot(v-\sigma_0\dot{\xi}_0(t_0)) - \sigma_1\dot{\xi}_0(t_1).$$

In the proof of part (iv) of the preceding theorem we showed that $(t_1,t_0,x_0) \mapsto \mathbf{\Phi}_{A_{(t_0,x_0)}}(t_1,t_0)$ is continuous. Since the map

$$(t_1,t_0,x_0) \mapsto \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=t_1} \Phi^F(t,t_0,x_0) = \widehat{F}(t_1,\Phi^F(t_1,t_0,x_0))$$

is also continuous, we conclude in this case that Φ^F is *continuously* differentiable. ■

The next construction is a natural one, intuitively; it involves "wiggling" the initial data for an ordinary differential equation.

5.1.10 Definition (Variation of initial data) Let $\boldsymbol{F}$ be an ordinary differential equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

and let $\xi_0\colon \mathbb{T}' \to X$ be a solution for $\boldsymbol{F}$ satisfying $\xi_0(t_0) = \boldsymbol{x}_0$ for some $t_0 \in \mathbb{T}'$ and $\boldsymbol{x}_0 \in X$. A ***variation*** of the initial data $(t_0, \boldsymbol{x}_0)$ in the direction of $(\tau, \boldsymbol{v}) \in \mathbb{R} \times \mathbb{R}^n$ is the curve

$$s \mapsto (t_0 + s\tau, \boldsymbol{x}_0 + s\boldsymbol{v}),$$

which we assume takes values in $\mathbb{T} \times X$ for small $s \in \mathbb{R}_{>0}$. •

For s small, one can then consider "perturbations" of the solution $t \mapsto \xi_0(t) = \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}_0)$, by which we mean the solutions $t \mapsto \Phi^{\boldsymbol{F}}(t, t_0 + s\tau, \boldsymbol{x}_0 + s\boldsymbol{v})$. Note, by Theorem 3.2.13(ix), that if $(t, t_0, \boldsymbol{x}_0) \in D_{\boldsymbol{F}}$, then $(t, t_0 + s\tau, \boldsymbol{x}_0 + s\boldsymbol{v}) \in D_{\boldsymbol{F}}$ for s sufficiently small. Thus we can ask for the "first-order effect" of the variation of the initial data on the solution at the final time t. Precisely, this is

$$\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} \Phi^{\boldsymbol{F}}(t, t_0 + s\tau, \boldsymbol{x}_0 + s\boldsymbol{v}) \in \mathbb{R}^n.$$

This is sufficiently interesting a quantity that we give it a name.

5.1.11 Definition (Infinitesimal variation corresponding to variation of initial data) Let $\boldsymbol{F}$ be an ordinary differential equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

and let $\xi_0\colon \mathbb{T}' \to X$ be a solution for $\boldsymbol{F}$ satisfying $\xi_0(t_0) = \boldsymbol{x}_0$ for some $t_0 \in \mathbb{T}'$ and $\boldsymbol{x}_0 \in X$. The ***infinitesimal variation*** associated with the variation of the initial data $(t_0, \boldsymbol{x}_0)$ in the direction of $(\tau, \boldsymbol{v}) \in \mathbb{R} \times \mathbb{R}^n$ is

$$\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} \Phi^{\boldsymbol{F}}(t, t_0 + s\tau, \boldsymbol{x}_0 + s\boldsymbol{v}) \in \mathbb{R}^n,$$

when the derivative exists. •

The following result, which is an immediate consequence of Theorem 5.1.8, gives the formula for this first-order effect.

5.1.12 Corollary (The infinitesimal variation corresponding to a variation of initial data) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{X} \to \mathbb{R}^{\mathrm{n}}$$

satisfying the hypotheses of Theorem 5.1.8, and let $\xi_0\colon \mathbb{T}' \to \mathrm{X}$ *be a solution for* $\mathbf{F}$ *satisfying* $\xi_0(\mathrm{t}_0) = \mathbf{x}_0$ *for some* $\mathrm{t}_0 \in \mathbb{T}'$ *and* $\mathbf{x}_0 \in \mathrm{X}$. *The infinitesimal variation associated with the variation of the initial data* $(\mathrm{t}_0, \mathbf{x}_0)$ *in the direction of* $(\tau, \mathbf{v}) \in \mathbb{R} \times \mathbb{R}^{\mathrm{n}}$ *is given by*

$$\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} \Phi^{\mathbf{F}}(\mathrm{t}, \mathrm{t}_0 + \mathrm{s}\tau, \mathbf{x}_0 + \mathrm{s}\mathbf{v}) = \boldsymbol{\Phi}_{\mathbf{A}_{(\mathrm{t}_0, \mathrm{x}_0)}}(\mathrm{t}, \mathrm{t}_0) \cdot \mathbf{v} - \tau \boldsymbol{\Phi}_{\mathbf{A}_{(\mathrm{t}_0, \mathrm{x}_0)}}(\mathrm{t}, \mathrm{t}_0) \cdot \widehat{\mathbf{F}}(\mathrm{t}_0, \mathbf{x}_0).$$

Proof This follows from Theorem 5.1.8 and the Chain Rule. ■

5.1.1.4 While we're at it: ordinary differential equations of class C^m In the previous section we considered ordinary differential equations depending continuously differentiably on state (Theorem 5.1.8) and on state and time (Corollary 5.1.9). In this section we extend these results to case where we assume more differentiability.

Let us start with just differentiability in state.

5.1.13 Theorem (Higher-order differentiability of flows for ordinary differential equations) *Let $\mathbf{F}$ be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times X\to\mathbb{R}^n,$$

let $m\in\mathbb{Z}_{>0}$, and make the following assumptions:

(i) the map $t\mapsto\widehat{\mathbf{F}}(t,\mathbf{x})$ is measurable for each $\mathbf{x}\in X$;

(ii) the map $\mathbf{x}\mapsto\widehat{\mathbf{F}}(t,\mathbf{x})$ is of class C^m for each $t\in\mathbb{T}$;

(iii) for each $(t,\mathbf{x})\in\mathbb{T}\times X$, there exist $r,\rho\in\mathbb{R}_{>0}$ and

$$g_0,g_1,\dots,g_m\in L^1([t_0-\rho,t_0+\rho];\mathbb{R}_{\geq 0})$$

such that

(a) $\|\widehat{\mathbf{F}}(s,\mathbf{y})\|\leq g_0(s)$ for $(s,\mathbf{y})\in([t_0-\rho,t_0+\rho]\cap\mathbb{T})\times B(r,\mathbf{x})$ and

(b) $\left|\dfrac{\partial^l\widehat{F}_j}{\partial x_{k_1}\cdots\partial x_{k_l}}(s,\mathbf{y})\right|\leq g_l(s)$ for $(t,\mathbf{y})\in([t_0-\rho,t_0+\rho]\cap\mathbb{T})\times B(r,\mathbf{x})$, $j,k_1,\dots,k_l\in\{1,\dots,n\}$, and $l\in\{1,\dots,m\}$.

Then, for $t,t_0\in\mathbb{T}$, $\Phi^{\mathbf{F}}_{t,t_0}\colon D_{\mathbf{F}}(t,t_0)\to X$ is a C^m-diffeomorphism onto its image.

Proof It suffices to prove the theorem locally, since once this is done, one can use an argument like that in the proof of Theorem 5.1.8(iv) to get the global result.

We recursively define ordinary differential equations $\boldsymbol{F}_{L,m}$, $m\in\mathbb{Z}_{>0}$, by $\boldsymbol{F}_{L,1}=\boldsymbol{F}_L$ and then by $\boldsymbol{F}_{L,m+1}=(\boldsymbol{F}_{L,m})_L$. We have

$$\widehat{\boldsymbol{F}}_{L,m}\colon\mathbb{T}\times X\times(\mathbb{R}^n)^{2m-1}\to(\mathbb{R}^n)^{2m},$$

and one can verify that the components of

$$\widehat{\boldsymbol{F}}_{L,m}(t,\boldsymbol{x},\boldsymbol{v}_1,\dots,\boldsymbol{v}_{2m-1})$$

are linear combinations of expressions of the form

$$D^k\widehat{\boldsymbol{F}}(\boldsymbol{x})\cdot(\boldsymbol{v}_{j_1},\dots,\boldsymbol{v}_{j_k})$$

for some $j_1,\dots,j_k\in\{1,\dots,2m-1\}$. Let us draw the important conclusion from this. In the proof of Theorem 5.1.8, we saw that if the hypotheses of the present theorem hold for $m=1$, then the hypotheses of Theorem 3.2.13 hold for $\boldsymbol{F}_{L,1}$, and so the mapping

$$(\boldsymbol{x},\boldsymbol{v}_1)\mapsto\Phi^{\boldsymbol{F}_{L,1}}_{t,t_0}(\boldsymbol{x},\boldsymbol{v}_1))=(\Phi^{\boldsymbol{F}}_{t,t_0}(\boldsymbol{x}),D\Phi^{\boldsymbol{F}}_{t,t_0}(\boldsymbol{x})\cdot\boldsymbol{v})$$

is a locally bi-Lipschitz homeomorphism. This shows, in particular, that Φ^F_{t,t_0} is of class C^1. Similarly, if F satisfies the hypotheses of the present theorem for $m = 2$, then $F_{L,1}$ satisfies the hypotheses of Theorem 5.1.8 and so the mapping

$$\begin{aligned}(x, v_1, v_2, v_3) &\mapsto \Phi^{F_{L,2}}_{t,t_0}(x, v_1, v_2, v_3)\\ &= (\Phi^{F_{L,1}}_{t,t_0}(x, v_1), D\Phi^{F_{L,1}}_{t,t_0}(x, v_1) \cdot (v_2, v_3))\\ &= (\Phi^F_{t,t_0}(x), D\Phi^F_{t,t_0}(x) \cdot v_1, D^2\Phi^F_{t,t_0}(x) \cdot (v_1, v_2), D\Phi^F_{t,t_0}(x) \cdot v_3).\end{aligned}$$

is a locally bi-Lipschitz homeomorphism. This shows that Φ^F_{t,t_0} is of class C^2. One can continue this process for arbitrary m. ∎

Using the preceding theorem, it is fairly easy to characterise the flows of ordinary differential equations that depend regularly *jointly* on time and state.

5.1.14 Corollary (Higher-order differentiability in state and time of flows for ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times X \to \mathbb{R}^n$$

and suppose that $\widehat{\mathbf{F}}$ *is of class* C^m *for some* $m \in \mathbb{Z}_{>0}$. *Then the mapping* $\Phi^{\mathbf{F}} \colon D_{\mathbf{F}} \to X$ *is of class* C^m.

Proof The inductive constructions from the proof of Theorem 5.1.13, using the regularity conclusions of Corollary 5.1.9 in place of those of Theorem 5.1.8, give the result in this case. ∎

Immediate consequences of the preceding two results are the following.

5.1.15 Corollary (Infinite differentiability of flows for ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times X \to \mathbb{R}^n$$

and make the following assumptions:

(i) *the map* $t \mapsto \widehat{\mathbf{F}}(t, \mathbf{x})$ *is measurable for each* $\mathbf{x} \in X$;

(ii) *the map* $\mathbf{x} \mapsto \widehat{\mathbf{F}}(t, \mathbf{x})$ *is of class* C^∞ *for each* $t \in \mathbb{T}$;

(iii) *for each* $(t, \mathbf{x}) \in \mathbb{T} \times X$, *there exist* $r, \rho \in \mathbb{R}_{>0}$ *and*

$$g_j \in L^1([t_0 - \rho, t_0 + \rho]; \mathbb{R}_{\geq 0}), \qquad j \in \mathbb{Z}_{\geq 0},$$

such that

(a) $\|\widehat{\mathbf{F}}(s, \mathbf{y})\| \leq g_0(s)$ *for* $(s, \mathbf{y}) \in (([t_0 - \rho, t_0 + \rho] \cap \mathbb{T}) \times B(r, \mathbf{x})$ *and*

(b) $\left|\dfrac{\partial^l \widehat{F}_j}{\partial x_{k_1} \cdots \partial x_{k_l}}(s, \mathbf{y})\right| \leq g_l(s)$ *for* $(s, \mathbf{y}) \in ([t_0 - \rho, t_0 + \rho] \cap \mathbb{T}) \times B(r, \mathbf{x}), j, k_1, \ldots, k_l \in \{1, \ldots, n\}$, *and* $l \in \mathbb{Z}_{>0}$.

Then, for $t, t_0 \in \mathbb{T}$, $\Phi^{\mathbf{F}}_{t,t_0} \colon D_{\mathbf{F}}(t, t_0) \to X$ *is a* C^∞*-diffeomorphism onto its image.*

5.1.16 Corollary (Infinite differentiability in state and time of flows for ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{X} \to \mathbb{R}^{\mathrm{n}}$$

and suppose that $\widehat{\mathbf{F}}$ *is of class* C^∞*. Then* $\Phi^{\mathbf{F}}\colon \mathrm{D}_{\mathbf{F}} \to \mathrm{X}$ *is of class* C^∞*.*

5.1.2 Linearisation of ordinary difference equations

Now we turn to the linearisation of systems of ordinary difference equations. We shall mirror the constructions of the preceding sections for differential equations, but in the discrete-time there are far fewer difficulties since one does not have to fuss with the precise nature of time dependence as one does in the continuous-time case.

5.1.2.1 Linearisation along solutions Suppose that we have a system of ordinary difference equations $\boldsymbol{F}$ with right-hand side $\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n$ and that we have a solution $\xi_0\colon \mathbb{T}' \to X$ for $\boldsymbol{F}$. We wish to understand what happens to solutions "nearby" this fixed solution ξ_0.

To do this, we suppose that the map

$$\begin{aligned}\widehat{\boldsymbol{F}}_t\colon X &\to \mathbb{R}^n\\ \boldsymbol{x} &\mapsto \widehat{\boldsymbol{F}}(t, \boldsymbol{x})\end{aligned}$$

is of class C^1. We denote

$$\boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) = \boldsymbol{D}\widehat{\boldsymbol{F}}_t(\boldsymbol{x}), \qquad t \in \mathbb{T}.$$

We then suppose that we have a solution $\xi\colon \mathbb{T}' \to X$ for $\boldsymbol{F}$ for which the deviation $\boldsymbol{\nu} \triangleq \xi - \xi_0$ is small. Let us try to understand the behaviour of $\boldsymbol{\nu}$. Naïvely, we can do this as follows:

$$\xi(t + h) = (\xi_0 + \boldsymbol{\nu})(t + h) = \widehat{\boldsymbol{F}}(t, \xi_0(t) + \boldsymbol{\nu}(t)) = \widehat{\boldsymbol{F}}(t, \xi_0(t)) + \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \xi_0(t)) \cdot \boldsymbol{\nu}(t) + \cdots .$$

We will not here try to be precise about what "$\cdots$" might mean, but merely say that the idea of the preceding equation is that we approximate using the constant and first-order terms in the Taylor expansion, and then pray that this gives us something meaningful. Note that, since ξ_0 is a solution for $\boldsymbol{F}$, the approximation we arrive at is

$$\boldsymbol{\nu}(t + h) \approx \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \xi_0) \cdot \boldsymbol{\nu}(t).$$

Meaningful or not, the preceding naïve calculations give rise to the following definition.

5.1.17 Definition (Linearisation of an ordinary difference equation along a solution) Let $\boldsymbol{F}$ be an ordinary difference equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

supposing that $\widehat{\boldsymbol{F}}_t$ is of class C^1 for every $t \in \mathbb{T}$. For a solution $\xi_0\colon \mathbb{T}' \to X$ for $\boldsymbol{F}$, the ***linearisation of*** $\mathbf{F}$ ***along*** $\boldsymbol{\xi_0}$ is the linear ordinary difference equation $\boldsymbol{F}_{L,\xi_0}$ with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}_{L,\xi_0}\colon \mathbb{T}' \times \mathbb{R}^n &\to \mathbb{R}^n\\ (t, v) &\mapsto \widehat{\boldsymbol{DF}}(\xi_0(t)) \cdot v.\end{aligned}$$

•

Note that a solution $t \mapsto \boldsymbol{\nu}(t)$ for the linearisation of $\boldsymbol{F}$ along ξ_0 satisfies

$$\boldsymbol{\nu}(t + h) = \boldsymbol{A}(t)(\boldsymbol{\nu}(t)),$$

where

$$\boldsymbol{A}(t) = \widehat{\boldsymbol{DF}}(t, \xi_0(t)).$$

This is indeed a linear ordinary difference equation. We note that, even when $\boldsymbol{F}$ is autonomous, the linearisation will generally be nonautonomous, due to the dependence of the reference solution ξ_0 on time.

Note that there is an alternative view of linearisation that can be easily developed, one where linearisation is of the *equation*, not just along a solution. The construction we make is the following.

5.1.18 Definition (Linearisation of an ordinary difference equation) Let $\boldsymbol{F}$ be an ordinary difference equation with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

supposing that $\widehat{\boldsymbol{F}}_t$ is of class C^1 for every $t \in \mathbb{T}$. The ***linearisation*** of $\boldsymbol{F}$ is the ordinary difference equation $\boldsymbol{F}_L$ with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}_L\colon \mathbb{T} \times (X \times \mathbb{R}^n) &\to \mathbb{R}^n \oplus \mathbb{R}^n\\ (t, (x, v)) &\mapsto (\widehat{\boldsymbol{F}}(t, x), \widehat{\boldsymbol{DF}}(t, x)(v)).\end{aligned}$$

•

Solutions of the linearisation of $\boldsymbol{F}$ are then mappings $t \mapsto (\xi(t), \boldsymbol{\nu}(t))$ satisfying

$$\begin{aligned}\xi(t + h) &= \widehat{\boldsymbol{F}}(t, \xi(t)),\\ \boldsymbol{\nu}(t + h) &= \widehat{\boldsymbol{DF}}(t, \xi(t)) \cdot \boldsymbol{\nu}(t).\end{aligned}$$

We see, then, that in this version of linearisation we carry along the original difference equation $\boldsymbol{F}$ as part of the linearisation. This is, in no way, incompatible with the definition of linearisation along a solution ξ_0, since one needs $\boldsymbol{F}$ to provide the solution.

5.1.2.2 Linearisation about equilibria In this section we consider what amounts to a special case of linearisation about a solution. The solution we consider is a very particular sort of solution, as given by the following definition.

5.1.19 Definition (Equilibrium state for an ordinary difference equation) Let F be an ordinary difference equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^n.$$

A state $x_0 \in X$ is an ***equilibrium state*** if $\widehat{F}(t, x_0) = x_0$ for every $t \in \mathbb{T}$. •

The following result gives the relationship between equilibrium states and solutions.

5.1.20 Proposition (Equilibrium states and constant solutions) *Let* $\mathbf{F}$ *be an ordinary difference equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{X} \to \mathbb{R}^{\mathrm{n}}.$$

A state $\mathbf{x}_0 \in \mathrm{X}$ *is an equilibrium state if and only if the constant function* $\mathrm{t} \mapsto \mathbf{x}_0$ *is a solution for* $\mathbf{F}$*.*

Proof Let us denote by ξ_0 the constant function $t \mapsto x_0$.

First suppose that x_0 is an equilibrium state. Then $\xi_0(t + h) = x_0$ for every $t \in \mathbb{T}$ and $\widehat{F}(t, \xi_0(t)) = x_0$ and so

$$\xi_0(t + h) = \widehat{F}(t, \xi_0(t)), \qquad t \in \mathbb{T},$$

and thus ξ_{x_0} is a solution.

Next suppose that ξ_0 is a solution. Then

$$x_0 = \xi_0(t + h) = \widehat{F}(t, \xi_0(t)) = \widehat{F}(t, x_0), \qquad t \in \mathbb{T},$$

so giving that x_0 is an equilibrium state. ■

Note that, as a consequence of the preceding simple result, we can linearise about the constant solution $t \mapsto x_0$ in the event that x_0 is an equilibrium state. Let us, however, use some particular language in this case.

5.1.21 Definition (Linearisation of an ordinary difference equation about an equilibrium state) Let F be an ordinary difference equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

supposing that $\widehat{F}_t$ is of class C^1 for every $t \in \mathbb{T}$, and let x_0 be an equilibrium state. The ***linearisation of*** $\mathbf{F}$ ***about*** $\mathbf{x_0}$ is the linear ordinary difference equation F_{L,x_0} with right-hand side

$$\begin{aligned} \widehat{F}_{L,x_0}\colon \mathbb{T} \times \mathbb{R}^n &\to \mathbb{R}^n \\ (t, v) &\mapsto D\widehat{F}(t, x_0) \cdot v. \end{aligned}$$

•

A solution $t \mapsto \boldsymbol{\nu}(t)$ for $\boldsymbol{F}_{L,\boldsymbol{x}_0}$ satisfies

$$\boldsymbol{\nu}(t+h) = \boldsymbol{A}(t)(\boldsymbol{v}),$$

where

$$\boldsymbol{A}(t) = \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \boldsymbol{x}_0).$$

Thus we see that the linearisation about an equilibrium point is indeed a linear ordinary difference equation, just as it should be since the same is true of the linearisation about an arbitrary solution. What is special here, however, is that the linearisation is autonomous if $\boldsymbol{F}$ is autonomous. Thus the linearisation when $\boldsymbol{F}$ is autonomous is a linear ordinary difference equation with constant coefficients.

5.1.2.3 The flow of the linearisation In this section, in contrast with the preceding sections, we give a very precise characterisation of linearisation. In contrast to the situation with differential equations, the conclusions of the results in this section follow almost immediately from their hypotheses.

We shall first investigate thoroughly the properties of the flow of an ordinary difference equation that has more regularity properties than are required for the basic existence and uniqueness theorem, Theorem 3.4.2. In order to state the result we want, we will make use of some ideas that we will not develop fully until Section 5.6. Let us suppose that we have a system of ordinary difference equations $\boldsymbol{F}$ with right-hand side

$$\widehat{\boldsymbol{F}} \colon \mathbb{T} \times X \to \mathbb{R}^n,$$

and let $(t_0, \boldsymbol{x}_0) \in \mathbb{T} \times X$. We then have the solution

$$t \mapsto \boldsymbol{\xi}_0(t) \triangleq \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}_0)$$

defined for $t \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) \cap \mathbb{T}_{\geq t_0}$. We then define

$$\begin{aligned} \boldsymbol{A}_{(t_0,\boldsymbol{x}_0)} \colon J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) \cap \mathbb{T}_{\geq t_0} &\to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n) \\ t &\mapsto \boldsymbol{D}\widehat{\boldsymbol{F}}(t, \Phi^{\boldsymbol{F}}(t, t_0, \boldsymbol{x}_0)). \end{aligned}$$

Now consider the linear time-varying difference equation $\boldsymbol{F}^T_{(t_0,\boldsymbol{x}_0)}$ with right-hand side

$$\begin{aligned} \widehat{\boldsymbol{F}}^T_{(t_0,\boldsymbol{x}_0)} \colon J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) \cap \mathbb{T}_{\geq t_0} \times \mathbb{R}^n &\to \mathbb{R}^n \\ (t, \boldsymbol{v}) &\mapsto \boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}(t) \cdot \boldsymbol{v}. \end{aligned}$$

To describe solutions of this linear ordinary difference equation, we consider first the following ordinary difference equation. For $t \in J_{\boldsymbol{F}}(t_0, \boldsymbol{x}_0) \cap \mathbb{T}_{\geq t_0} \times \mathbb{R}^n$, we consider the following initial value problem:

$$\boldsymbol{\Psi}(s+h) = \boldsymbol{A}_{(t_0,\boldsymbol{x}_0)}(s) \circ \boldsymbol{\Psi}(s), \quad \boldsymbol{\Psi}(t) = \boldsymbol{I}_n.$$

As a linear ordinary difference equation, this initial value problem has solutions defined for all $s \in J_F(t_0, x_0) \cap \mathbb{T}_{\geq t}$. Moreover, we denote the solution at time s by $\mathbf{\Phi}_{A_{(t_0,x_0)}}(s,t)$; the associated map

$$\mathbf{\Phi}_{A_{(t_0,x_0)}} \colon P_F(t_0, x_0) \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$$

is what we shall call the "state transition map" in Section 5.6.1.2, and we shall use some of the results from this section in the proof below. Here we denote

$$P_F(t_0, x_0) = \{(s,t) \in J_F(t_0, x_0) \cap \mathbb{T}_{\geq t_0} \times J_F(t_0, x_0) \cap \mathbb{T}_{\geq t_0} \mid s \geq t\}.$$

In particular, we shall use the fact that the solution to the initial value problem

$$v(s+h) = A_{(t_0,x_0)}(s) \cdot v(s), \quad v(t) = v_0$$

is

$$v(s) = \mathbf{\Phi}_{A_{(t_0,x_0)}}(s,t) \cdot v_0, \qquad s \in J_F(t_0, x_0) \cap \mathbb{T}_{\geq t}.$$

With the preceding background, we can now state the theorem.

5.1.22 Theorem (Differentiability of flows for ordinary difference equations) *Let* $\mathbf{F}$ *be an ordinary difference equation with right-hand side*

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times \mathrm{X} \to \mathbb{R}^{\mathrm{n}},$$

and make the following assumption:

(i) the map $\mathbf{x} \mapsto \widehat{\mathbf{F}}(\mathrm{t}, \mathbf{x})$ *is continuously differentiable for each* $\mathrm{t} \in \mathbb{T}$*;*

Then the following statements hold:

(ii) for $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}$, $\Phi^{\mathbf{F}}_{\mathrm{t},\mathrm{t}_0} \colon \mathrm{D}_{\mathbf{F}}(\mathrm{t}, \mathrm{t}_0) \to \mathrm{X}$ *is of class* C^1 *and its derivative is given by* $\mathbf{D}\Phi^{\mathbf{F}}_{\mathrm{t},\mathrm{t}_0}(\mathbf{x}_0) = \mathbf{\Phi}_{\mathbf{A}_{(\mathrm{t}_0,\mathbf{x}_0)}}(\mathrm{t}, \mathrm{t}_0)$*;*

(iii) the map

$$\begin{aligned} \mathbf{D}\Phi^{\mathbf{F}} \colon \mathrm{D}_{\mathbf{F}} &\to \mathrm{L}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{n}}) \\ (\mathrm{t}, \mathrm{t}_0, \mathbf{x}) &\mapsto \mathbf{D}\Phi^{\mathbf{F}}_{\mathrm{t},\mathrm{t}_0}(\mathbf{x}) \end{aligned}$$

is continuous.

Proof (ii) We note that

$$\Phi^{F}_{t,t_0} = \widehat{F}_{t-h} \circ \cdots \circ \widehat{F}_{t_0+h} \circ \widehat{F}_{t_0}.$$

In a similar vein,

$$\mathbf{\Phi}_{A_{(t_0,x_0)}}(t, t_0) = A_{(t_0,x_0)}(t-h) \circ \cdots \circ A_{(t_0,x_0)}(t_0+h) \circ A_{(t_0,x_0)}(t_0).$$

Since

$$A_{(t_0,x_0)}(t) = D\widehat{F}(t, \Phi^{F}_{t,t_0}),$$

this part of the result follows from the Chain Rule (Theorem II-1.4.49).

(iii) In the present discrete-time case, this simply follows since $x \mapsto D\Phi^{F}_{t,t_0}(x)$ is continuous, since all functions defined on a topological space with the discrete topology are continuous (). ■ ref

Note that we have no analogue of Corollary 5.1.9 here since differentiability is not a meaningful concept for discrete-time functions.

The next construction is a natural one, intuitively; it involves "wiggling" the initial data for an ordinary differential equation. Note here that we cannot "wiggle" the initial condition in time in the discrete-time case/

5.1.23 Definition (Variation of initial state) Let F be an ordinary difference equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

and let $\xi_0\colon \mathbb{T}' \to X$ be a solution for F satisfying $\xi_0(t_0) = x_0$ for some $t_0 \in \mathbb{T}'$ and $x_0 \in X$. A ***variation*** of the initial state x_0 in the direction of $v \in \mathbb{R}^n$ is the curve

$$s \mapsto x_0 + sv,$$

which we assume takes values in $\mathbb{T} \times X$ for small $s \in \mathbb{R}_{>0}$. •

For s small, one can then consider "perturbations" of the solution $t \mapsto \xi_0(t) = \Phi^F(t, t_0, x_0)$, by which we mean the solutions $t \mapsto \Phi^F(t, t_0, x_0 + sv)$. Thus we ask for the "first-order effect" of the variation of the initial state on the solution at the final time t. Precisely, this is

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \Phi^F(t, t_0, x_0 + sv) \in \mathbb{R}^n.$$

This is sufficiently interesting a quantity that we give it a name.

5.1.24 Definition (Infinitesimal variation corresponding to variation of initial state) Let F be an ordinary difference equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

and let $\xi_0\colon \mathbb{T}' \to X$ be a solution for F satisfying $\xi_0(t_0) = x_0$ for some $t_0 \in \mathbb{T}'$ and $x_0 \in X$. The ***infinitesimal variation*** associated with the variation of the initial state x_0 in the direction of $v \in \mathbb{R}^n$ is

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \Phi^F(t, t_0, x_0 + sv) \in \mathbb{R}^n,$$

when the derivative exists. •

The following result, which is an immediate consequence of Theorem 5.1.22, gives the formula for this first-order effect.

5.1.25 Corollary (The infinitesimal variation corresponding to a variation of initial state) *Let* $\mathbf{F}$ *be an ordinary difference equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n$$

satisfying the hypotheses of Theorem 5.1.22, and let $\xi_0\colon \mathbb{T}' \to X$ *be a solution for* $\mathbf{F}$ *satisfying* $\xi_0(t_0) = \mathbf{x}_0$ *for some* $t_0 \in \mathbb{T}'$ *and* $\mathbf{x}_0 \in X$*. The infinitesimal variation associated with the variation of the initial state* $\mathbf{x}_0$ *in the direction of* $\mathbf{v} \in \mathbb{R}^n$ *is given by*

$$\frac{d}{ds}\Big|_{s=0} \Phi^{\mathbf{F}}(t, t_0, \mathbf{x}_0 + s\mathbf{v}) = \mathbf{\Phi}_{\mathbf{A}_{(t_0,x_0)}}(t, t_0) \cdot \mathbf{v}.$$

Proof This follows from Theorem 5.1.22 and the Chain Rule. ■

5.1.2.4 While we're at it: ordinary difference equations of class C^m In the previous section we considered ordinary differential equations depending continuously differentiably on state (Theorem 5.1.22). In this section we extend this result to case where we assume more differentiability.

Let us start with just differentiability in state.

5.1.26 Theorem (Higher-order differentiability of flows for ordinary difference equations) *Let* $\mathbf{F}$ *be an ordinary difference equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n,$$

let $m \in \mathbb{Z}_{>0}$*, and make the following assumption:*

(i) the map $\mathbf{x} \mapsto \widehat{\mathbf{F}}(t, \mathbf{x})$ *is of class* C^m *for each* $t \in \mathbb{T}$*.*

Then, for $t, t_0 \in \mathbb{T}$, $\Phi^{\mathbf{F}}_{t,t_0}\colon D_{\mathbf{F}}(t, t_0) \to X$ *is of class* C^m*.*

Proof Theorem 5.1.22 follows since compositions of C^1-maps are of class C^1. Similarly, compositions of C^m-maps are of class C^m (Theorem II-1.4.51). ■

An immediate consequence of the preceding theorem is the following.

5.1.27 Corollary (Infinite differentiability of flows for ordinary difference equations) *Let* $\mathbf{F}$ *be an ordinary difference equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times X \to \mathbb{R}^n.$$

and make the following assumption:

(i) the map $\mathbf{x} \mapsto \widehat{\mathbf{F}}(t, \mathbf{x})$ *is of class* C^∞ *for each* $t \in \mathbb{T}$*.*

Then, for $t, t_0 \in \mathbb{T}$, $\Phi^{\mathbf{F}}_{t,t_0}\colon D_{\mathbf{F}}(t, t_0) \to X$ *is of class* C^∞*.*

Exercises

5.1.1 Let F be a linear homogeneous ordinary differential equation with constant coefficients in a finite-dimensional $\mathbb{R}$-vector space X and with right-hand side $\widehat{F}(t, x) = \mathsf{A}(x)$, as in Section 5.2.2. Answer the following questions.

(a) Write an explicit formula for the flow for F.

(b) Determine the linearisation F_L of F as in Definition 5.1.2.

(c) Write an explicit formula for the flow of F_L.

(d) What can you say about the linearisation about the equilibrium 0, i.e., about the zero trajectory?

5.1.2 Let F be a kth-order scalar ordinary differential equation with right-hand side

$$\widehat{F} \colon \mathbb{T} \times X \times \mathrm{L}^{\leq k-1}_{\mathrm{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}.$$

Let $\boldsymbol{F}_1$ be the first-order ordinary differential equation with k states, as in Exercise 3.1.23. Denote the state for F by $x \in X$ and the state for $\boldsymbol{F}_1$ by $\boldsymbol{y} \in X_1 = X \times \mathbb{R}^{k-1}$, as in Exercise 3.1.23.

(a) Argue that the correct definition of an equilibrium state for the kth-order ordinary differential equation F is a state $x_0 \in X$ such that

$$\widehat{F}(t, x_0, 0, \dots, 0) = 0.$$

(b) Show that $x_0 \in X$ is an equilibrium for F as in part (a) if and only if $(x_0, 0, \dots, 0)$ is an equilibrium state for $\boldsymbol{F}_1$.

Now let $x_0 \in X$ be an equilibrium state for F, as in part (a), with $\boldsymbol{y}_0 = (x_0, 0, \dots, 0) \in X_1$ the associated equilibrium state for $\boldsymbol{F}_1$.

(c) Determine the linearisation of $\boldsymbol{F}_1$ about an equilibrium state $\boldsymbol{y}_0 = (x_0, 0, \dots, 0)$.

(d) Show that the linearisation of $\boldsymbol{F}_1$ is a first-order linear ordinary differential equation with k states that comes from a kth-order scalar linear ordinary differential equation, and determine explicitly the coefficients in this scalar equation in terms of $\widehat{\boldsymbol{F}}$.

5.1.3 For the ordinary differential equations $\boldsymbol{F}$ with the given time-domains, state spaces, and right-hand sides, determine their equilibrium states and the linearisations about these equilibrium states:

(a) $\mathbb{T} = \mathbb{R}$, $X = \mathbb{R}$, and $\widehat{F}(t, x) = x - x^3$;

(b) $\mathbb{T} = \mathbb{R}$, $X = \mathbb{R}$, and $\widehat{F}(t, x) = a(t)x$, $a \in \mathsf{C}^0(\mathbb{T}; \mathbb{R})$ not identically zero;

(c) $\mathbb{T} = \mathbb{R}$, $X = \mathbb{R}$, and $\widehat{F}(t, x) = \cos(x)$;

(d) $\mathbb{T} = \mathbb{R}$, $X = \mathbb{R}^2$, and $\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = (x_2, x_1 - x_1^3)$;

(e) $\mathbb{T} = \mathbb{R}$, $X = \mathbb{R}^2$, and $\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = (x_2, a(t)x_1)$, $a \in \mathsf{C}^0(\mathbb{T}; \mathbb{R})$ not identically zero;

(f) $\mathbb{T} = \mathbb{R}$, $X = \mathbb{R}^2$, and $\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = (x_2, \cos(x_1))$;

(g) $\mathbb{T} = \mathbb{R}$, $X = \mathbb{R}^2_{>0}$, and $\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = (\alpha x_1 - \beta x_1 x_2, \delta x_1 x_2 - \gamma x_2)$, $\alpha, \beta, \delta, \gamma \in \mathbb{R}_{>0}$.

5.1.4 Let F be a linear homogeneous ordinary difference equation with constant coefficients in a finite-dimensional $\mathbb{R}$-vector space X and with right-hand side $\widehat{F}(t, x) = \mathsf{A}(x)$, as in Section 5.6.2. Answer the following questions.

(a) Write an explicit formula for the flow for F.

(b) Determine the linearisation F_L of F as in Definition 5.1.18.

(c) Write an explicit formula for the flow of F_L.

(d) What can you say about the linearisation about the equilibrium 0, i.e., about the zero trajectory?

more exercises for difference equations

Section 5.2

Systems of linear homogeneous ordinary differential equations

In this section we shall begin our study of systems of linear ordinary differential equations by working with homogeneous systems. We will follow our development of Section 3.1.3.3 and consider the state space to be a finite-dimensional $\mathbb{R}$-vector space V. Thus we work with a system of linear homogeneous ordinary differential equations F in a finite-dimensional $\mathbb{R}$-vector space X, whose right-hand side, therefore, takes the form

$$\begin{aligned}\widehat{F}\colon \mathbb{T}\times\mathsf{X} &\to \mathsf{X}\\ (t,x) &\mapsto \mathsf{A}(t)(x)\end{aligned} \tag{5.3}$$

for a map $\mathsf{A}\colon \mathbb{T}\to \mathrm{L}(\mathsf{X};\mathsf{X})$. Thus we are looking at differential equations whose solutions $t\mapsto \xi(t)$ satisfy

$$\dot{\xi}(t) = \mathsf{A}(t)(\xi(t)).$$

Our treatment will be structured in the same way as was the treatment in Section 4.2 for scalar equations, to emphasise the similarities between the two theories.

Do I need to read this section? This material is fundamental to the study of linear system theory. •

5.2.1 Equations with time-varying coefficients

We begin by a consideration of general systems with time-varying coefficients, i.e., for which A is not a constant function of time.

5.2.1.1 Solutions and their properties

First let us verify that the basic existence and uniqueness result holds for the differential equations we are considering.

5.2.1 Proposition (Local existence and uniqueness of solutions for systems of linear homogeneous ordinary differential equations) *Consider the system of linear homogeneous ordinary differential equations* F *with right-hand side (5.3) and suppose that* $\mathsf{A}\in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{X};\mathsf{X}))$. *Let* $(\mathrm{t}_0,\mathrm{x}_0)\in\mathbb{T}\times\mathsf{X}$. *Then there exists an interval* $\mathbb{T}'\subseteq\mathbb{T}$ *and* $\xi\in \mathsf{AC}_{\mathrm{loc}}(\mathbb{T}';\mathsf{X})$ *that is a solution for* F *and which satisfies* $\xi(\mathrm{t}_0)=\mathrm{x}_0$. *Moreover, if* $\tilde{\mathbb{T}}'\subseteq\mathbb{T}$ *is another subinterval and if* $\tilde{\xi}\in\mathsf{AC}_{\mathrm{loc}}(\tilde{\mathbb{T}}';\mathsf{X})$ *is another solution for* F *satisfying* $\tilde{\xi}(\mathrm{t}_0)=\mathrm{x}_0$, *then* $\tilde{\xi}(\mathrm{t})=\xi(\mathrm{t})$ *for every* $\mathrm{t}\in\tilde{\mathbb{T}}'\cap\mathbb{T}'$.

Proof By choosing a basis for X, we can take $\mathsf{X}=\mathbb{R}^n$ so that A is an $n\times n$ matrix-valued function, which we denote as $\boldsymbol{A}$ in the usual way. (This is legitimate by Exercise 5.2.1.) We denote the components of $\boldsymbol{A}(t)$ by $A^j_k(t)$, $j,k\in\{1,\dots,n\}$. Let

$g \in \mathsf{L}^1_{\text{loc}}(\mathbb{T}; \mathbb{R}_{\geq 0})$ be such that $|A^j_k(t)| \leq g(t)$ for $t \in \mathbb{T}$. We recall the following formula from Proposition II-1.1.11(i): for $v \in \mathbb{R}^n$,

$$\sum_{j=1}^n |v_j| \leq \sqrt{n} \left(\sum_{j=1}^n |v_j|^2 \right)^{1/2}.$$

Now let $a, b \in \mathbb{T}$, $a < b$, be such that $t_0 \in [a, b]$. The following estimate will be useful for us: for any $x_1, x_2 \in \mathbb{R}^n$ and $t \in [a, b]$,

$$\begin{aligned}
\|\widehat{F}(t, x_1) - \widehat{F}(t, x_2)\| &= \|A(t)(x_1) - A(t)(x_2)\| = \|A(t)(x_1 - x_2)\| \\
&= \left(\sum_{j=1}^n \left(\sum_{k=1}^n A^j_k(t)(x_{1,k} - x_{2,k}) \right)^2 \right)^{1/2} \\
&\leq \left(\sum_{j=1}^n \left(\sum_{k=1}^n |A^j_k(t)(x_{1,k} - x_{2,k})| \right)^2 \right)^{1/2} \\
&\leq \left(\sum_{j=1}^n \left(g(t) \sum_{k=1}^n |x_{1,k} - x_{2,k}| \right)^2 \right)^{1/2} = \left(g(t)^2 \sum_{j=1}^n \left(\sum_{k=1}^n |x_{1,k} - x_{2,k}| \right)^2 \right)^{1/2} \\
&\leq \left(g(t)^2 \sum_{j=1}^n \sum_{k=1}^n |x_{1,k} - x_{2,k}|^2 \right)^{1/2} \leq \left(n g(t)^2 \sum_{k=1}^n |x_{1,k} - x_{2,k}|^2 \right)^{1/2} \\
&= \sqrt{n} g(t) \left(\sum_{k=1}^n |x_{1,k} - x_{2,k}|^2 \right)^{1/2} = \sqrt{n} g(t) \|x_1 - x_2\|.
\end{aligned}$$

Let us take $h(t) = \sqrt{n} g(t)$, noting that h is locally integrable. We consider the Banach space $\mathsf{C}^0([a, b]; \mathbb{R}^n)$ with the norm

$$\|f\|_{\infty, h, t_0} = \sup \left\{ \left\| f(t) \mathrm{e}^{-2 \int_{t_0}^t h(s) \, \mathrm{d}s} \right\| \;\middle|\; t \in [a, b] \right\}.$$

Let us define

$$F_+ \colon \mathsf{C}^0([a, b]; \mathbb{R}^n) \to \mathsf{C}^0([a, b]; \mathbb{R}^n)$$

by

$$F_+(\xi)(t) = x_0 + \int_{t_0}^t A(s)(\xi(s)) \, \mathrm{d}s.$$

We now estimate, for $t \in [a, b]$,

$$\begin{aligned}
\|F_+(\xi_1)(t) - F_+(\xi_2)(t)\| &= \left\| \int_{t_0}^t A(s)(\xi_1(s) - \xi_2(s)) \, \mathrm{d}s \right\| \\
&\leq \int_{t_0}^t \|A(s)(\xi_1(s) - \xi_2(s))\| \, \mathrm{d}s \\
&\leq \int_{t_0}^t \|\xi_1(s) - \xi_2(s)\| \mathrm{e}^{-2\int_{t_0}^s h(\tau)\,\mathrm{d}\tau} h(s) \mathrm{e}^{2\int_{t_0}^s h(\tau)\,\mathrm{d}\tau} \, ds \\
&\leq \frac{1}{2} \|\xi_1 - \xi_2\|_{\infty,h,t_0} \int_{t_0}^t \frac{\mathrm{d}}{\mathrm{d}s} \mathrm{e}^{2\int_{t_0}^s h(\tau)\,\mathrm{d}\tau} \, \mathrm{d}s \\
&\leq \frac{1}{2} \|\xi_1 - \xi_2\|_{\infty,h,t_0} \mathrm{e}^{2\int_{t_0}^t h(s)\,\mathrm{d}s}.
\end{aligned}$$

From this we conclude that

$$\|F_+(\xi_1) - F_+(\xi_2)\|_{\infty,L} \leq \frac{1}{2} \|\xi_1 - \xi_2\|_{\infty,L}.$$

Now one argues just as in the proof of Theorem 3.2.8(ii), using the Contraction Mapping Theorem to conclude the existence of a unique solution ξ_+ for F on $[a, b]$. Moreover, since

$$\xi(t) = x_0 + \int_{t_0}^t A(s)(\xi(s)) \, \mathrm{d}s,$$

we see that ξ is locally absolutely continuous and satisfies the initial conditions. ■

Next, as for scalar linear ordinary differential equations, we show that solutions exist for all time.

5.2.2 Proposition (Global existence of solutions for systems of linear homogeneous ordinary differential equations) *Consider the system of linear homogeneous ordinary differential equations* F *with right-hand side (5.3) and suppose that* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{X}; \mathsf{X}))$. *If* $\xi \colon \mathbb{T}' \to \mathsf{X}$ *is a solution for* F, *then there exists a solution* $\overline{\xi} \colon \mathbb{T} \to \mathsf{X}$ *for which* $\overline{\xi}|\mathbb{T}' = \xi$.

Proof Note that in the proof of Proposition 5.2.1 we showed that solutions of the initial value problem

$$\dot{\xi}(t) = \mathsf{A}(t)(\xi(t)), \quad \xi(t_0) = x_0,$$

exist on any interval $[a, b] \subseteq \mathbb{T}$ containing t_0. So let $t \in \mathbb{T}$ and let $[a, b]$ be an interval containing both t_0 and t. We then have a solution for the initial value problem that is defined at t. Since $t \in \mathbb{T}$ is arbitrary, the result follows. ■

Now we can discuss the set of all solutions of a system of linear homogeneous ordinary differential equation F with right-hand side

$$\begin{aligned}
\widehat{F} \colon \mathbb{T} \times \mathsf{X} &\to \mathsf{X} \\
(t, x) &\mapsto \mathsf{A}(t)(x).
\end{aligned}$$

To this end, we denote by

$$\mathrm{Sol}(F) = \left\{ \xi \in \mathsf{AC}_{\mathrm{loc}}(\mathbb{T}; \mathsf{X}) \;\middle|\; \dot{\xi}(t) = \mathsf{A}(t)(\xi(t)), \text{ a.e. } t \in \mathbb{T} \right\}$$

the set of solutions for F. The following result is then the main structural result about the set of solutions to a system of linear homogeneous ordinary differential equations.

5.2.3 Theorem (Vector space structure of sets of solutions) *Consider the system of linear homogeneous ordinary differential equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.3) and suppose that* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$*. Then* Sol(F) *is an* n*-dimensional subspace of* $\mathsf{AC}_{\mathrm{loc}}(\mathbb{T}; \mathsf{X})$*.*

Proof Fix $t_0 \in \mathbb{T}$ and define

$$\begin{aligned} \sigma_{t_0} \colon \mathrm{Sol}(F) &\to \mathsf{X} \\ \xi &\mapsto \xi(t_0). \end{aligned}$$

We claim that σ_{t_0} is an isomorphism of vector spaces. First, the verification of the linearity of σ_{t_0} follows from the equalities

$$(\xi_1 + \xi_2)(t_0) = \xi_1(t_0) + \xi_2(t_0), \quad (a\xi)(t_0) = a(\xi(t_0)),$$

which themselves follow from the definition of the vector space structure in $\mathsf{AC}_{\mathrm{loc}}(\mathbb{T}; \mathsf{X})$. Next let us show that σ_{t_0} is injective by showing that $\ker(\sigma_{t_0}) = \{0\}$. Indeed, suppose that $\sigma_{t_0}(\xi) = 0$. Then, by the uniqueness assertion of Proposition 5.2.1, it follows that $\xi(t) = 0$ for every $t \in \mathbb{T}$, as desired. To show that σ_{t_0} is surjective, let $x_0 \in \mathsf{X}$. Then, by the existence assertion of Proposition 5.2.1, there exists $\xi \in \mathrm{Sol}(F)$ such that $\xi(t_0) = x_0$, i.e., such that $\sigma_{t_0}(\xi) = x_0$. ■

The following corollary, immediate from the proof of the theorem, gives an easy check on the linear independence of subsets of Sol(F).

5.2.4 Corollary (Linear independence in Sol(F)) *Consider the system of linear homogeneous ordinary differential equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.3) and suppose that* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$*. Then a subset* $\{\xi_1, \dots, \xi_k\} \subseteq \mathrm{Sol(F)}$ *is linearly independent if and only if, for some* $\mathsf{t} \in \mathbb{T}$*, the subset* $\{\xi_1(\mathsf{t}), \dots, \xi_k(\mathsf{t})\} \subseteq \mathsf{X}$ *is linearly independent.*

As with scalar linear homogeneous ordinary differential equations, the theorem allows us to give a special name to a basis for Sol(F).

5.2.5 Definition (Fundamental set of solutions) Consider the system of linear homogeneous ordinary differential equations F in the n-dimensional $\mathbb{R}$-vector space X with right-hand side (5.3) and suppose that $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$. A set $\{\xi_1, \dots, \xi_n\}$ of linearly independent elements of Sol(F) is a ***fundamental set of solutions*** for F. •

5.2.1.2 The continuous-time state transition map We now present a particular way of organising a fundamental set of solutions into one object that, for all intents and purposes, completely characterises Sol(F). This we organise as the following theorem.

5.2.6 Theorem (Existence of, and properties of, the continuous-time state transition map) *Consider the system of linear homogeneous ordinary differential equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.3) and suppose that* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$. *Then there exists a unique map* $\Phi^{\mathrm{c}}_{\mathsf{A}} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ *with the following properties:*

(i) for each $t_0 \in \mathbb{T}$*, the function*

$$\begin{aligned} \Phi^{\mathrm{c}}_{\mathsf{A},t_0} \colon \mathbb{T} &\to \mathrm{L}(\mathsf{X}; \mathsf{X}) \\ t &\mapsto \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0) \end{aligned}$$

is locally absolutely continuous and satisfies the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t} \Phi^{\mathrm{c}}_{\mathsf{A},t_0}(t) = \mathsf{A}(t) \circ \Phi^{\mathrm{c}}_{\mathsf{A},t_0}(t), \quad \Phi^{\mathrm{c}}_{\mathsf{A},t_0}(t_0) = \mathrm{id}_{\mathsf{X}};$$

(ii) the solution to the initial value problem

$$\dot{\xi}(t) = \mathsf{A}(t)(\xi(t)), \quad \xi(t_0) = x_0,$$

is $t \mapsto \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)(x_0)$*;*

(iii) $\det(\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)) = \mathrm{e}^{\int_{t_0}^{t} \operatorname{tr}(\mathsf{A}(s))\, \mathrm{d}s}$ *(the* ***Abel–Jacobi–Liouville formula****);*

(iv) for $t, t_0, t_1 \in \mathbb{T}$, $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_1) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}(t_1, t_0)$*;*

(v) for each $t, t_0 \in \mathbb{T}$, $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$ *is invertible and* $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)^{-1} = \Phi^{\mathrm{c}}_{\mathsf{A}}(t_0, t)$.

Proof First of all, we define $\Phi^{\mathrm{c}}_{\mathsf{A}}$ by the condition in part (i). That is to say, we define $\Phi^{\mathrm{c}}_{\mathsf{A}}$ by

$$\frac{\partial \Phi^{\mathrm{c}}_{\mathsf{A}}}{\partial t}(t, t_0) = \mathsf{A}(t) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0), \quad \Phi^{\mathrm{c}}_{\mathsf{A}}(t_0, t_0) = \mathrm{id}_{\mathsf{X}}.$$

Note that this is an initial value problem associated with the system of linear homogeneous ordinary differential equations F_{A} in $\mathrm{L}(\mathsf{X}; \mathsf{X})$ with right-hand side

$$\begin{aligned} \widehat{F}_{\mathsf{A}} \colon \mathbb{T} \times \mathrm{L}(\mathsf{X}; \mathsf{X}) &\to \mathrm{L}(\mathsf{X}; \mathsf{X}) \\ \Phi^{\mathrm{c}}_{\mathsf{A}} &\mapsto \mathsf{A}(t) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}; \end{aligned}$$

note the mapping $\Phi^{\mathrm{c}}_{\mathsf{A}} \mapsto \mathsf{A}(t) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}$ is linear, cf. Proposition I-5.4.16. Thus, by Proposition 5.2.1, it possesses a unique solution which, by Proposition 5.2.2, exists in all of $\mathbb{T}$. This proves the existence and uniqueness and part (i).

(ii) We compute

$$\frac{\mathrm{d}}{\mathrm{d}t} \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)(x_0) = \frac{\partial \Phi^{\mathrm{c}}_{\mathsf{A}}}{\partial t}(t, t_0)(x_0) = \mathsf{A}(t) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)(x_0)$$

and $\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0, t_0)(x_0) = x_0$, which shows that $t \mapsto \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)(x_0)$ solves the stated initial value problem. By uniqueness of such solutions, this part of the theorem follows.

(iii) We start with a lemma.

1 Lemma *Let $\mathbb{T} \subseteq \mathbb{R}$ be an interval and let $\mathbf{A}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ be a locally absolutely continuous map. For $j, k \in \{1, \dots, n\}$, let $C_{jk}(t)$ be the (j,k)th cofactor of $\mathbf{A}(t)$, i.e., $(-1)^{j+k}$ times the determinant of the $(n-1)\times(n-1)$ matrix formed by deleting the jth row and kth column from $\mathbf{A}(t)$. Then*

$$\frac{d(\det \mathbf{A})}{dt}(t) = \sum_{j,k=1}^{n} C_{jk}(t)\dot{A}^j_k(t)$$

for almost every $t \in \mathbb{T}$.

Proof The row/column expansion rule for determinants gives

$$\det \boldsymbol{A}(t) = \sum_{k=1}^{n} A^j_k(t) C_{jk}(t)$$

for any $j \in \{1, \dots, n\}$. Using the Chain Rule,

$$\frac{d(\det \boldsymbol{A})}{dt}(t) = \sum_{j,k=1}^{n} \frac{\partial(\det \boldsymbol{A})}{\partial A^j_k} \dot{A}^j_k(t) = \sum_{j,k=1}^{n} C_{jk}(t)\dot{A}^j_k(t),$$

because C_{jk} does not depend on the (j,k)th component of $\boldsymbol{A}$. ▼

We choose a basis $\{e_1, \dots, e_n\}$ for X and denote by $\boldsymbol{A}(t)$ the matrix representative of $\mathsf{A}(t)$ and by $\Phi^c_{\boldsymbol{A}}(t, t_0)$ the matrix representative of $\Phi^c_{\mathsf{A}}(t, t_0)$. (That we can reduce to $\mathsf{X} = \mathbb{R}^n$ is justified by Exercises 5.2.1 and 5.2.2.) For $j, k \in \{1, \dots, n\}$, denote by $C_{jk}(t, t_0)$ the (j,k)th cofactor of $\Phi^c_{\boldsymbol{A}}(t, t_0)$, i.e., $(-1)^{j+k}$ times the determinant of the matrix $\Phi^c_{\boldsymbol{A}}(t, t_0)$ with the jth row and kth column removed. Also let $\boldsymbol{C}(t, t_0)$ be the matrix formed from these cofactors. Denote by $\Phi_{jk}(t, t_0)$ the (j,k)th component of $\Phi^c_{\boldsymbol{A}}$. Using the lemma,

$$\begin{aligned}
\frac{d}{dt} \det \Phi^c_{\boldsymbol{A}}(t, t_0) &= \sum_{j,k=1}^{n} C_{jk}(t, t_0) \frac{d}{dt} \Phi_{jk}(t, t_0) \\
&= \operatorname{tr}\left(\boldsymbol{C}(t, t_0)^T \frac{d}{dt} \Phi^c_{\boldsymbol{A}}(t, t_0) \right) \\
&= \operatorname{tr}(\Phi^c_{\boldsymbol{A}}(t, t_0) \boldsymbol{C}(t, t_0)^T \boldsymbol{A}(t)),
\end{aligned}$$

using part (i), the definition of trace and transpose, and the easily verified fact that $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$ for $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Now we note that

$$\Phi^c_{\boldsymbol{A}} \boldsymbol{C}(t, t_0)^T = \det \Phi^c_{\boldsymbol{A}} \boldsymbol{I}_n$$

using Cramer's Rule for matrix inversion. Thus we arrive at

$$\frac{d}{dt} \det \Phi^c_{\boldsymbol{A}}(t, t_0) = \det \Phi^c_{\boldsymbol{A}}(t, t_0) \boldsymbol{A}(t).$$

This equation is a first-order scalar linear homogeneous ordinary differential equation, and we have seen how to solve these in Example 4.2.5. Applying the computations there to the present equation, and using the fact that $\det \Phi^c_{\boldsymbol{A}}(t, t_0) = \det \boldsymbol{I}_n = 1$, we get this part of the theorem.

(iv) We compute

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_1)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_1,t_0)) = \mathsf{A}(t)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_1,t_0)$$

and

$$\Phi^{\mathrm{c}}_{\mathsf{A}}(t_1,t_1)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_1,t_0) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t_1,t_0).$$

We also have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0) = \mathsf{A}(t)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0).$$

That is to say, both $t\mapsto\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)$ and $t\mapsto\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_1)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_1,t_0)$ satisfy the initial problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t) = \mathsf{A}(t)\circ\Phi(t),\quad \Phi(t_1) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t_1,t_0).$$

By uniqueness of solutions for systems of linear homogeneous ordinary differential equations, we conclude that $\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_1)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_1,t_0)$, as desired.

(v) The invertibility of $\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)$ follows from part (iii). The specific formula for the inverse follows from the formula

$$\mathrm{id}_{\mathsf{X}} = \Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t_0) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0),$$

which itself follows from part (iv). ■

Let us formally name the mapping $\Phi^{\mathrm{c}}_{\mathsf{A}}$ defined in the theorem.

5.2.7 Definition (Continuous-time state transition map) Consider the system of linear homogeneous ordinary differential equations F in the n-dimensional $\mathbb{R}$-vector space X with right-hand side (5.3) and suppose that $\mathsf{A}\in\mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{X};\mathsf{X}))$. The map $\Phi^{\mathrm{c}}_{\mathsf{A}}\colon\mathbb{T}\times\mathbb{T}\to\mathrm{L}(\mathsf{X};\mathsf{X})$ from Theorem 5.2.6 is the ***continuous-time state transition map***. •

One imagines that it is possible to compute the continuous-time state transition map if one is given a fundamental set of solutions. The following procedure gives an explicit means of doing this.

5.2.8 Procedure (Determining the continuous-time state transition map from a fundamental set of solutions) Given a system of linear homogeneous ordinary differential equations F in the n-dimensional $\mathbb{R}$-vector space X with right-hand side equation

$$\widehat{F}(t,x) = \mathsf{A}(t)(x),$$

with $\mathsf{A}\in\mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{X};\mathsf{X}))$, and given a fundamental set of solutions $\{\xi_1,\dots,\xi_n\}$, do the following.

1. Choose a basis $\{e_1,\dots,e_n\}$.
2. Let $\xi_j\colon\mathbb{T}\to\mathbb{R}^n$ be the components of ξ_j, $j\in\{1,\dots,n\}$, i.e.,

$$\xi_j(t) = \xi_{1,j}(t)e_1+\cdots+\xi_{j,n}(t)e_n.$$

If $\mathsf{X}=\mathbb{R}^n$, one can just take the components of ξ_j, $j\in\{1,\dots,n\}$, in the standard basis, as usual.

3. Assemble the matrix function $\Xi\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ by making the components of $\xi_1(t), \dots, \xi_j(t)$ the columns of $\Xi(t)$:

$$\Xi(t) = \begin{bmatrix} \xi_{1,1}(t) & \xi_{2,1}(t) & \cdots & \xi_{n,1}(t) \\ \xi_{1,2}(t) & \xi_{2,2}(t) & \cdots & \xi_{n,2}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{1,n}(t) & \xi_{2,n}(t) & \cdots & \xi_{n,n}(t) \end{bmatrix}.$$

   (Be sure you understand that $\xi_{j,k}(t)$ is the kth component of $\xi_j(t)$.) We call the matrix-valued function $\Xi\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ a ***fundamental matrix*** for F.
4. Define $\mathbf{\Phi}(t, t_0) = \Xi(t)\Xi(t_0)^{-1}$.
5. Then $\mathbf{\Phi}(t, t_0)$ is the matrix representative of $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$ in the basis $\{e_1, \dots, e_n\}$. •

Let us verify that the preceding procedure does indeed yield the continuous-time state transition map.

5.2.9 Proposition (Determining the continuous-time state transition map from a fundamental set of solutions) *Consider the system of linear homogeneous ordinary differential equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.3) and suppose that* $\mathsf{A} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$. *Then Procedure 5.2.8 will produce the continuous-time state transition map.*

Proof By choosing a basis $\{e_1, \dots, e_n\}$ as in Procedure 5.2.8, we can assume that $\mathsf{X} = \mathbb{R}^n$. (This is legitimate by virtue of Exercises 5.2.1 and 5.2.2.) Let us denote by $A(t)$ the matrix representative of $\mathsf{A}(t)$. Defining $\mathbf{\Phi}(t, t_0)$ as in the given procedure, we have

$$\frac{\partial \mathbf{\Phi}}{\partial t}(t, t_0) = \dot{\Xi}(t)\Xi(t_0)^{-1}.$$

Noting that each of ξ_j, $j \in \{1, \dots, n\}$, is a solution for F, we have

$$\dot{\xi}_{j,k}(t) = \sum_{l=1}^{n} A^k_l(t)\xi_{j,l}(t), \qquad j \in \{1, \dots, n\},\ t \in \mathbb{T}.$$

Therefore, in matrix notation,

$$\begin{bmatrix} \dot{\xi}_1(t) \mid \cdots \mid \dot{\xi}(t) \end{bmatrix} = A(t) \begin{bmatrix} \xi_1(t) \mid \cdots \mid \xi(t) \end{bmatrix} \implies \dot{\Xi}(t) = A(t)\Xi(t), \qquad t \in \mathbb{T}.$$

Therefore,

$$\frac{\partial \mathbf{\Phi}}{\partial t}(t, t_0) = A(t)\Xi(t)\Xi(t_0)^{-1} = A(t)\mathbf{\Phi}(t, t_0).$$

Moreover, $\mathbf{\Phi}(t_0, t_0) = \boldsymbol{I}_n$. Thus $t \mapsto \mathbf{\Phi}(t, t_0)$ satisfies the matrix representative of the initial value problem satisfied by $t \mapsto \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$, i.e., $\mathbf{\Phi}(t, t_0)$ is the matrix representative of $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$. ■

In general, it cannot be expected to find the continuous-time state transition map for a system of linear homogeneous ordinary differential equations. However, to illustrate Procedure 5.2.8, let us give a "cooked" example.

5.2.10 Example (Computing the continuous-time state transition map) We take the system of linear homogeneous ordinary differential equations $\boldsymbol{F}$ in $\mathbb{R}^2$ with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}\colon (0,\infty)\times\mathbb{R}^2 &\to \mathbb{R}^2\\ (t,(x_1,x_2)) &\mapsto \Big(\frac{1}{t}x_1 - x_2, \frac{1}{t^2}x_1 + \frac{2}{t}x_2\Big).\end{aligned}$$

Solutions $t \mapsto (x_1(t), x_2(t))$ satisfy

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \underbrace{\begin{bmatrix}\frac{1}{t} & -1\\ \frac{1}{t^2} & \frac{2}{t}\end{bmatrix}}_{\boldsymbol{A}(t)}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}.$$

A direct verification shows that the functions $\xi_1, \xi_2\colon (0,\infty) \to \mathbb{R}^2$ defined by

$$\xi_1 = (t^2, -t), \quad \xi_2(t) = (-t^2\ln(t), t + t\ln(t))$$

are solutions of $\boldsymbol{F}$. To verify that these are linearly independent we compute

$$\det\begin{bmatrix}t^2 & -t^2\ln(t)\\ -t & t + t\ln(t)\end{bmatrix} = t^3.$$

As this determinant is nowhere zero, we conclude the desired linear independence.

Now we determine the continuous-time state transition map in this case. In the notation of Procedure 5.2.8, we have

$$\Xi(t) = \begin{bmatrix}t^2 & -t^2\ln(t)\\ -t & t + t\ln(t)\end{bmatrix},$$

and then a tedious computation gives

$$\Phi^{\mathrm{c}}_A(t,t_0) = \Xi(t)\Xi(t_0)^{-1} = \begin{bmatrix}-\frac{t^2(\ln(t/t_0)-1)}{t_0^2} & -\frac{t^2\ln(t/t_0)}{t_0}\\ \frac{t\ln(t/t_0)}{t_0^2} & \frac{t(\ln(t/t_0)+1)}{t_0}\end{bmatrix}.$$ •

5.2.1.3 The Peano–Baker series In this section we will provide a series representation for the continuous-time state transition map for a system of linear ordinary differential equations. This is presented for two reasons: (1) as an illustration of series methods in ordinary differential equations, as these arise in many important contexts; (2) as an illustration, in an elementary setting, of iterative procedure used in the proof of Theorem 3.2.8. It is by no means being suggested that the series representation we give for the continuous-time state transition map is useful for computation.

We let $\mathbb{T} \subseteq \mathbb{R}$ be an interval and let $\mathsf{A} \in \mathsf{L}^1_{\text{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$. By its definition, the continuous-time state transition map $(t, t_0) \mapsto \Phi^{\text{c}}_{\mathsf{A}}(t, t_0)$ is determined from the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t) = \mathsf{A}(t) \circ \Phi(t), \quad \Phi(t_0) = \mathrm{id}_{\mathsf{X}}.$$

Let us fix $t, t_0 \in \mathbb{T}$ and take $t > t_0$, for concreteness. By the Fundamental Theorem of Calculus (in the form of Theorem III-2.9.33), this is equivalent to

$$\Phi(t) = \mathrm{id}_{\mathsf{X}} + \int_{t_0}^{t} \mathsf{A}(\tau) \circ \Phi(s)\, \mathrm{d}s. \tag{5.4}$$

Let us informally iterate to find a solution. We define $\Phi_0 \colon [t_0, t] \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ by $\Phi_0(\tau) = \mathrm{id}_{\mathsf{X}}$. This will, generally, not satisfy the integral equation (5.4). So, let us substitute this zeroth-order approximation into the same integral equation to get (hopefully) a better approximation $\Phi_1 \colon [t_0, t] \to \mathrm{L}(\mathsf{X}; \mathsf{X})$:

$$\Phi_1(\tau) = \Phi_0(\tau) + \int_{t_0}^{t} \mathsf{A}(\tau) \circ \Phi_0(\tau)\, \mathrm{d}\tau = \mathrm{id}_{\mathsf{X}} + \int_{t_0}^{t} \mathsf{A}(\tau)\, \mathrm{d}\tau.$$

We now continue this process iteratively, assuming that, if we have defined $\Phi_k \colon [t_0, t] \to \mathrm{L}(\mathsf{X}; \mathsf{X})$, we define $\Phi_{k+1} \colon [t_0, t] \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ by

$$\Phi_{k+1}(\tau) = \Phi_k(\tau) + \int_{t_0}^{t} \mathsf{A}(\tau) \circ \Phi_k(\tau)\, \mathrm{d}\tau.$$

It is pretty clear that

$$\Phi_k(t) - \Phi_{k-1}(t) = \underbrace{\int_{t_0}^{t} \int_{t_0}^{t_1} \cdots \int_{t_0}^{t_{k-1}} \mathsf{A}(t_1) \circ \mathsf{A}(t_2) \circ \cdots \circ \mathsf{A}(t_k)\, \mathrm{d}t_k \cdots \mathrm{d}t_2 \mathrm{d}t_1}_{I_k(t, t_0)}.$$

Thus we can make the following definition.

5.2.11 Definition (Peano–Baker series) For an interval $\mathbb{T} \subseteq \mathbb{R}$, for $t_0 \in \mathbb{T}$, and for $\mathsf{A} \in \mathsf{L}^\infty_{\text{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$, the series

$$I_\infty(t, t_0) = \mathrm{id}_{\mathsf{X}} + \sum_{k=1}^{\infty} I_k(t, t_0)$$

is the $\mathbf{t_0}$***-Peano–Baker series*** for A. •

Of course, the definition is quite meaningless without addressing whether the series converges. The main result of this section is now the following.

5.2.12 Theorem (Convergence of the Peano–Baker series) *Let* X *be a finite-dimensional* $\mathbb{R}$*-vector space. For an interval* $\mathbb{T} \subseteq \mathbb{R}$*, for* $t_0 \in \mathbb{T}$*, and for* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{X}; \mathsf{X}))$*, the* t_0*-Peano–Baker series converges uniformly on every compact subinterval of* $\mathbb{T}$*, and, moreover,* $\mathrm{I}_\infty(t, t_0) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$.

Proof Let $T_+ > t_0$. We will show that the t_0-Peano–Baker series converges uniformly to $t \mapsto \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$ on $[t_0, T_+]$. A similar proof can be concocted for $T_- < t_0$. Then, given a compact subinterval $\mathbb{T}' \subseteq \mathbb{T}$. the theorem follows by taking T_- and T_+ such that $\mathbb{T}' \subseteq [T_-, T_+]$.

We let $\{e_1, \dots, e_n\}$ be a basis for X. We let $\boldsymbol{A}(t)$ be the matrix representative of $\mathsf{A}(t)$ and let $\boldsymbol{I}_k(t, t_0)$ be the matrix representative for $I_k(t, t_0)$. Note that, because the matrix representation for a composition of linear maps is the product of the matrix representations (Theorem I-5.4.22), we have

$$\int_{t_0}^{t} \int_{t_0}^{t_1} \cdots \int_{t_0}^{t_{k-1}} \boldsymbol{A}(t_1)\boldsymbol{A}(t_2) \cdots \boldsymbol{A}(t_k) \, \mathrm{d}t_k \cdots \mathrm{d}t_2 \mathrm{d}t_1.$$

For $\boldsymbol{B} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ denote by $\|\boldsymbol{B}\|_{\mathrm{Fr}}$ the Frobenius norm of $\boldsymbol{B}$. Recall from Proposition II-1.1.16(vi) that

$$\|\boldsymbol{B}\boldsymbol{C}\| \leq \|\boldsymbol{B}\|\|C\|. \tag{5.5}$$

We also use the following lemma.

1 Lemma *If* $\mathbb{T} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$*, then*

$$\int_{t_0}^{t} \int_{t_0}^{t_1} \cdots \int_{t_0}^{t_{k-1}} \mathrm{f}(t_1)\mathrm{f}(t_2) \cdots \mathrm{f}(t_k) \, \mathrm{d}t_k \cdots \mathrm{d}t_2 \mathrm{d}t_1 = \frac{1}{k!} \left(\int_{t_0}^{t} \mathrm{f}(\tau) \, \mathrm{d}\tau \right)^k.$$

Proof For $k = 1$ the claim is a tautology. So suppose the claim true for $k = m$ and compute

$$\begin{aligned}
\int_{t_0}^{t} \int_{t_0}^{t_1} \cdots \int_{t_0}^{t_m} & f(t_1)f(t_2) \cdots f(t_{m+1}) \, \mathrm{d}t_{m+1} \cdots \mathrm{d}t_2 \mathrm{d}t_1 \\
&= \int_{t_0}^{t} \frac{1}{k!} \left(\int_{t_0}^{t_1} f(\tau) \, \mathrm{d}\tau \right)^m \mathrm{d}t_1 \\
&= \int_{t_0}^{t} \frac{1}{(m+1)!} \frac{\mathrm{d}}{\mathrm{d}t_1} \left(\int_{t_0}^{t_1} f(\tau) \, \mathrm{d}\tau \right)^{m+1} \mathrm{d}t_1 \\
&= \frac{1}{(m+1)!} \left(\int_{t_0}^{t} f(\tau) \, \mathrm{d}\tau \right)^{m+1},
\end{aligned}$$

as claimed. ▼

Let $\epsilon \in \mathbb{R}_{>0}$. Since the series

$$\sum_{k=0}^{\infty} \frac{1}{k!} \left(\int_{t_0}^{t} \|\boldsymbol{A}(\tau)\|_{\mathrm{Fr}} \, \mathrm{d}\tau \right)^k$$

converges uniformly on $[t_0, T_+]$ (converging to $\mathrm{e}^{\int_{t_0}^t \|\boldsymbol{A}(\tau)\|_{\mathrm{Fr}}\,\mathrm{d}\tau}$), there exists $N \in \mathbb{Z}_{>0}$ such that, if $r, s \geq N$ with $r > s$, then

$$\sum_{k=s+1}^{r} \frac{1}{k!}\left(\int_{t_0}^{t} \|\boldsymbol{A}(\tau)\|_{\mathrm{Fr}}\,\mathrm{d}\tau\right)^k < \epsilon, \qquad t \in [t_0, T_+].$$

Therefore, for $r, s \geq N$ with $r > s$, we have

$$\begin{aligned}
\left\|\sum_{k=1}^{r} \boldsymbol{I}_k(t, t_0) - \sum_{k=1}^{s} \boldsymbol{I}_k(t, t_0)\right\|_{\mathrm{Fr}} &\leq \sum_{k=s+1}^{r} \|\boldsymbol{I}_k(t, t_0)\|_{\mathrm{Fr}} \\
&\leq \sum_{k=s+1}^{r} \int_{t_0}^{t}\int_{t_0}^{t_1}\cdots\int_{t_0}^{t_{k-1}} \|\boldsymbol{A}(t_1)\boldsymbol{A}(t_2)\cdots\boldsymbol{A}(t_k)\|_{\mathrm{Fr}}\,\mathrm{d}t_k\cdots\mathrm{d}t_2\mathrm{d}t_1 \\
&\leq \sum_{k=s+1}^{r} \int_{t_0}^{t}\int_{t_0}^{t_1}\cdots\int_{t_0}^{t_{k-1}} \|\boldsymbol{A}(t_1)\|_{\mathrm{Fr}}\,\|\boldsymbol{A}(t_2)\|_{\mathrm{Fr}}\cdots\|\boldsymbol{A}(t_k)\|_{\mathrm{Fr}}\,\mathrm{d}t_k\cdots\mathrm{d}t_2\mathrm{d}t_1 \\
&\leq \sum_{k=s+1}^{r} \int_{t_0}^{t}\int_{t_0}^{t}\cdots\int_{t_0}^{t} \|\boldsymbol{A}(t_1)\|_{\mathrm{Fr}}\,\|\boldsymbol{A}(t_2)\|_{\mathrm{Fr}}\cdots\|\boldsymbol{A}(t_k)\|_{\mathrm{Fr}}\,\mathrm{d}t_k\cdots\mathrm{d}t_2\mathrm{d}t_1 \\
&\leq \sum_{k=s+1}^{r} \frac{1}{k!}\left(\int_{t_0}^{t} \|\boldsymbol{A}(\tau)\|_{\mathrm{Fr}}\,\mathrm{d}\tau\right)^k < \epsilon
\end{aligned}$$

using (5.5) and Lemma 1. This shows that the sequence of functions

$$t \mapsto \mathrm{id}_X + \sum_{k=1}^{m} \boldsymbol{I}_k(t, t_0), \qquad m \in \mathbb{Z}_{>0},$$

is uniformly Cauchy, and so uniformly convergent, cf. Theorem I-3.6.5.

Finally, we show that $I_\infty(t, t_0) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$. By the Fundamental Theorem of Calculus (in the form of Theorem III-2.9.33) the function

$$t \mapsto \boldsymbol{I}_k(t, t_0)$$

is locally absolutely continuous. Moreover, a direct calculation using the definitions gives

$$\dot{\boldsymbol{I}}_{k+1}(t, t_0) = \boldsymbol{A}(t)\boldsymbol{I}_k(t, t_0), \qquad k \in \mathbb{Z}_{>0}.$$

Therefore, the series

$$\sum_{k=1}^{\infty} \dot{\boldsymbol{I}}_k(t, t_0) = \boldsymbol{A}(t)\sum_{k=1}^{\infty} \boldsymbol{I}_{k-1}(t, t_0),$$

with the convention that $\boldsymbol{I}_0(t, t_0) = \boldsymbol{I}_n$, converges uniformly. Thus the series of term-by-term derivatives converges uniformly, and so term-by-term differentiation of $\boldsymbol{I}_\infty(t, t_0)$ is permissible. Moreover,

$$\dot{\boldsymbol{I}}_\infty(t, t_0) = \boldsymbol{A}(t)\boldsymbol{I}_\infty(t, t_0)$$

and $\boldsymbol{I}_\infty(t_0, t_0) = \boldsymbol{I}_n$. Thus the matrix representative of $t \mapsto I_\infty(t, t_0)$ satisfies the same initial value problem as $t \mapsto \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$, and the uniqueness assertion of Proposition 5.2.1 gives the result, at least for matrix representatives. That the conclusion also holds in X is a consequence of Exercise 5.2.2. ■

5.2.1.4 The adjoint equation In this section we consider a differential equation "dual" to a system of linear homogeneous ordinary differential equations. To do this, we ask the reader to recall from Definition I-5.7.1 the notion of the dual to a vector space and from Definition I-5.7.19 the notion of the dual of a linear map between vector spaces. With these notions as backdrop, we can now define the adjoint equation.

5.2.13 Definition (Adjoint of a system of linear homogeneous ordinary differential equations) Consider the system of linear homogeneous ordinary differential equations F in the n-dimensional $\mathbb{R}$-vector space X with right-hand side (5.3). The ***adjoint equation*** for F is the system F^* of linear homogeneous ordinary differential equations in X^* with right-hand side

$$\begin{aligned}\widehat{F^*}\colon \mathbb{T}\times\mathsf{X}^* &\to \mathsf{X}^*\\ (t,p) &\mapsto -\mathsf{A}^*(t)(p).\end{aligned}$$

•

Thus solutions $t \mapsto p(t)$ for the adjoint equation satisfy

$$\dot{p}(t) = -\mathsf{A}^*(t)(p(t)).$$

Let us give the continuous-time state transition map for the adjoint equation.

5.2.14 Proposition (Continuous-time state transition map for the adjoint equation) *Consider the system of linear homogeneous ordinary differential equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.3) and suppose that* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{X};\mathsf{X}))$. *Then* $\mathsf{A}^* \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{X}^*;\mathsf{X}^*))$ *and the continuous-time state transition map for the adjoint equation is defined by* $\Phi^{\mathrm{c}}_{-\mathsf{A}^*}(\mathrm{t},\mathrm{t}_0) = \Phi^{\mathrm{c}}_{\mathsf{A}}(\mathrm{t}_0,\mathrm{t})^*$ *for* $\mathrm{t},\mathrm{t}_0 \in \mathbb{T}$.

Proof The local integrability of A^* follows from choosing a basis for X so that A becomes the matrix-valued function $\boldsymbol{A}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$. In this case, $\mathsf{A}^*(t)$ has the matrix representative $\boldsymbol{A}(t)^T$ (Theorem I-5.7.22(iii)), which shows that the matrix representative of A is locally integrable if and only if the matrix representative of A^* is locally integrable.

By Theorem 5.2.6(v) we have

$$\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t) = \mathrm{id}_{\mathsf{X}}.$$

Differentiating this with respect to time we get

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t) = \Big(\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)\Big)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t) + \Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)\circ\Big(\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t)\Big),$$

from which we derive

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t) &= -\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t)\circ\Big(\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)\Big)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t)\\ &= -\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t)\circ\mathsf{A}(t)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t)\\ &= -\Phi^{\mathrm{c}}_{\mathsf{A}}(t_0,t)\circ\mathsf{A}(t). \qquad (5.6)\end{aligned}$$

Taking the dual of this equation, and using Proposition I-5.7.20(ii), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_{\mathsf{A}}^{\mathrm{c}}(t_0,t)^* = -\mathsf{A}^*(t)\circ\Phi_{\mathsf{A}}^{\mathrm{c}}(t_0,t)^*.$$

Since $(\Phi_{\mathsf{A}}^{\mathrm{c}})^*(t_0,t_0) = \mathrm{id}_{\mathsf{X}^*}$, we thus see that $t \mapsto (\Phi_{\mathsf{A}}^{\mathrm{c}})^*(t,t_0)$ satisfies the initial value problem that defines the continuous-time state transition map for the adjoint equation, and so the uniqueness assertion of Proposition 5.2.1 gives the result. ■

We have not yet addressed the important question, "Why should one care about the adjoint equation?" We convert this question into another question with the following result.

5.2.15 Proposition (A property of the adjoint equation) *Consider the system of linear homogeneous ordinary differential equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.3) and suppose that* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{X};\mathsf{X}))$. *Let* $\mathrm{t}_0 \in \mathbb{T}$, $\mathrm{x}_0 \in \mathsf{X}$, *and* $\mathrm{p}_0 \in \mathsf{X}^*$, *and denote* $\mathrm{x}(\mathrm{t}) = \Phi_{\mathsf{A}}^{\mathrm{c}}(\mathrm{t},\mathrm{t}_0)(\mathrm{x}_0)$ *and* $\mathrm{p}(\mathrm{t}) = (\Phi_{\mathsf{A}}^{\mathrm{c}})^*(\mathrm{t}_0,\mathrm{t})(\mathrm{p}_0)$. *Then*

$$\langle \mathrm{p}(\mathrm{t});\mathrm{x}(\mathrm{t})\rangle = \langle \mathrm{p}_0;\mathrm{x}_0\rangle.$$

Proof We compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle p(t);x(t)\rangle &= \langle \dot{p}(t);x(t)\rangle + \langle p(t);\dot{x}(t)\rangle\\ &= -\langle \mathsf{A}^*(t)(p(t));\Phi_{\mathsf{A}}^{\mathrm{c}}(t,t_0)(x_0)\rangle + \langle (\Phi_{\mathsf{A}}^{\mathrm{c}})^*(t_0,t)(p_0);\mathsf{A}(t)(x(t))\rangle\\ &= -\langle \mathsf{A}^*(t)\circ(\Phi_{\mathsf{A}}^{\mathrm{c}})^*(t_0,t)(p_0);\Phi_{\mathsf{A}}^{\mathrm{c}}(t,t_0)(x_0)\rangle + \langle (\Phi_{\mathsf{A}}^{\mathrm{c}})^*(t_0,t)(p_0);\mathsf{A}(t)\circ\Phi_{\mathsf{A}}^{\mathrm{c}}(t,t_0)(x_0)\rangle\\ &= 0.\end{aligned}$$

Since the function $t \mapsto \langle p(t);x(t)\rangle$ is locally absolutely continuous, it follows that this function is constant by Lemma III-2.9.32. ■

When $\alpha \in \mathsf{X}^*$ and $v \in \mathsf{X}$ satisfy $\alpha(v) = 0$, we say that α ***annihilates*** v. This is a sort of "orthogonality condition," although it most definitely is not an actual orthogonality condition, there being no inner product in sight. One of the upshots of the preceding result is the following corollary, saying that the adjoint equation preserves the annihilation condition.

5.2.16 Corollary (The geometric meaning of the adjoint equation) *Consider the system of linear homogeneous ordinary differential equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.3) and suppose that* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{X};\mathsf{X}))$. *Let* $\mathrm{t}_0 \in \mathbb{T}$, $\mathrm{x}_0 \in \mathsf{X}$, *and* $\mathrm{p}_0 \in \mathsf{X}^*$, *and denote* $\mathrm{x}(\mathrm{t}) = \Phi_{\mathsf{A}}^{\mathrm{c}}(\mathrm{t},\mathrm{t}_0)(\mathrm{x}_0)$ *and* $\mathrm{p}(\mathrm{t}) = (\Phi_{\mathsf{A}}^{\mathrm{c}})^*(\mathrm{t}_0,\mathrm{t})(\mathrm{p}_0)$. *If* $\langle \mathrm{p}_0;\mathrm{x}_0\rangle = 0$, *then* $\langle \mathrm{p}(\mathrm{t});\mathrm{x}(\mathrm{t})\rangle = 0$ *for all* $\mathrm{t} \in \mathbb{T}$.

It is this property of the adjoint equation that makes it an important tool in optimal control theory, but this is not a subject into which we shall dwell deeply here.

5.2.2 Equations with constant coefficients

We now consider the special case of systems of linear homogeneous equations with constant coefficients, i.e., those systems of linear ordinary differential equations F in a vector space X with right-hand sides

$$\widehat{F}(t, x) = \mathsf{A}(x), \tag{5.7}$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$. As with the scalar version of such equations that we studied in Section 4.2.2, there is a great deal more that we can say about such equations, beyond the general assertions in the preceding section. Indeed, one can say that, in principle, one can "solve" such equations, and we shall present a procedure for doing so.

Before we do so, however, we reiterate that the ordinary differential equations we are considering in this section are special cases of the time-varying equations of the preceding section, so all of the general statements made there apply here as well. In particular, Propositions 5.2.1 and 5.2.2, and Theorem 5.2.3 hold for equations of the form (5.7).

We have already seen in Theorem 5.2.3 that linear algebra plays a rôle in the theory of systems of linear homogeneous ordinary differential equations. We shall see in this section that this rôle is amplified for equations with constant coefficients. The material required for this was developed comprehensively in Sections I-5.4.9, I-5.4.10, and I-5.8.10, among other places. Let us summarise the essential points of this development as we shall need them. We let X be a finite-dimensional $\mathbb{R}$-vector space and let $\mathsf{L} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$. As in Definition I-4.5.60, we let $\mathsf{X}_{\mathbb{C}}$ be the complexification of X and, as in Definition I-5.4.62, we let $\mathsf{L}_{\mathbb{C}} \in \mathrm{L}(\mathsf{X}_{\mathbb{C}}; \mathsf{X}_{\mathbb{C}})$ be the complexification of L. We suppose we have distinct real eigenvalues

$$\ell_1, \dots, \ell_r$$

for L and distinct complex eigenvalues

$$\lambda_1 = \sigma_1 + \mathrm{i}\omega_1, \dots, \lambda_s = \sigma_s + \mathrm{i}\omega_s,$$

along with their complex conjugates. We let $m_{\mathrm{a}}(\ell_j, \mathsf{L})$, $j \in \{1, \dots, r\}$, and $m_{\mathrm{a}}(\lambda_j, \mathsf{L})$, $j \in \{1, \dots, s\}$, be the algebraic multiplicities (see Definition I-5.4.57).

1. For each $j \in \{1, \dots, r\}$, there is a generalised eigenspace

$$\overline{\mathsf{W}}(\ell_j, \mathsf{L}) = \ker((\ell_j \mathrm{id}_{\mathsf{X}} - \mathsf{L})^{m_{\mathrm{a}}(\ell_j, \mathsf{L})})$$

   of X of $\mathbb{R}$-dimension $m_{\mathrm{a}}(\ell_j, \mathsf{L})$ that is L-invariant (Proposition I-5.8.31).
2. For each $j \in \{1, \dots, s\}$, there is a generalised eigenspace

$$\overline{\mathsf{W}}(\lambda_j, \mathsf{L}_{\mathbb{C}}) = \ker((\lambda_j \mathrm{id}_{\mathsf{X}_{\mathbb{C}}} - \mathsf{L}_{\mathbb{C}})^{m_{\mathrm{a}}(\lambda_j, \mathsf{L})})$$

   of $\mathsf{X}_{\mathbb{C}}$ of $\mathbb{C}$-dimension $m_{\mathrm{a}}(\lambda_j, \mathsf{L})$ that is $\mathsf{L}_{\mathbb{C}}$-invariant (Proposition I-5.8.31).

3. For each $j \in \{1, \dots, s\}$, there is a subspace $\overline{\mathsf{W}}(\lambda_j, \mathsf{L})$ of X of $\mathbb{R}$-dimension $2m_{\mathrm{a}}(\lambda_j, \mathsf{L})$ that is L-invariant (part (I-vi) of Theorem I-5.4.67).
4. We have
$$\mathsf{X} = \overline{\mathsf{W}}(\ell_1, \mathsf{L}) \oplus \cdots \oplus \overline{\mathsf{W}}(\ell_r, \mathsf{L}) \oplus \overline{\mathsf{W}}(\lambda_1, \mathsf{L}) \oplus \cdots \oplus \overline{\mathsf{W}}(\lambda_s, \mathsf{L}) \tag{5.8}$$
(Corollary I-5.8.32). In particular,
$$\sum_{j=1}^{r} m_{\mathrm{a}}(\ell_j, \mathsf{L}) + 2 \sum_{j=1}^{s} m_{\mathrm{a}}(\lambda_j, \mathsf{L}) = \dim_{\mathbb{R}}(\mathsf{X})$$
This decomposition of X into L-invariant subspaces will form the basis for Procedure 5.2.23 where we determine the continuous-time state transition map for a system of linear homogeneous ordinary differential equations with constant coefficients.
5. There exists
   (a) $p_j \in \mathbb{Z}_{>0}$, $j \in \{1, \dots, r\}$,
   (b) $\boldsymbol{k}_j \in \mathbb{Z}_{>0}^{p_j}$, $j \in \{1, \dots, r\}$,
   (c) $q_j \in \mathbb{Z}_{>0}$, $j \in \{1, \dots, s\}$,
   (d) $\boldsymbol{l}_j \in \mathbb{Z}_{>0}^{q_j}$, $j \in \{1, \dots, s\}$, and
   (e) a basis $\mathscr{B}$ for X

   such that
$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \begin{bmatrix} \boldsymbol{J}(\ell_1, \boldsymbol{k}_1) & \cdots & \boldsymbol{0} & \boldsymbol{0} \cdots & \boldsymbol{0} & \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{J}(\ell_r, \boldsymbol{k}_r) & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{J}(\sigma_1, \omega_1, \boldsymbol{l}_1) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{J}(\sigma_s, \omega_s, \boldsymbol{l}_s) \end{bmatrix},$$
   where the top r diagonal blocks are Jordan arrangements for the real eigenvalues (Definition I-5.8.43) and the lower s diagonal blocks are $\mathbb{R}$-Jordan arrangements for the complex eigenvalues (Definition I-5.8.73). Moreover, this form of the matrix representative is unique up to reordering of the diagonal blocks. This is all proved in Theorem I-5.8.74.

5.2.2.1 Complexification of systems of linear ordinary differential equations

In Section 4.2.2.1 we complexified a scalar linear homogeneous ordinary differential equation with constant coefficients. The reason we had to do so was that the characteristic polynomial for such an equation will generally have complex roots, and these complex roots lead naturally to complex solutions of the differential

equation. It is only after taking real and imaginary parts of a complex solution that we recover the real solutions. The same sort of thing happens with systems of linear homogeneous ordinary differential equations with constant coefficients. In this case, the issue that arises is that one will generally have complex eigenvalues.

The process of complexification is an easy one, and requires no words like "everything we have done in the real case also works in the complex case," since we are working with systems defined on abstract $\mathbb{R}$-vector spaces, and $\mathsf{X}_{\mathbb{C}}$ is certainly a $\mathbb{R}$-vector space.

5.2.17 Definition (Complexification of a system of linear ordinary differential equations) Consider the system of linear homogeneous ordinary differential equations F with constant coefficients and with right-hand side (5.7). The ***complexification*** of F is the system of linear homogeneous ordinary differential equations $F_{\mathbb{C}}$ with constant coefficients given by

$$\begin{aligned} F_{\mathbb{C}} \colon \mathbb{T} \times \mathsf{X}_{\mathbb{C}} \times \mathsf{X}_{\mathbb{C}} &\to \mathsf{X}_{\mathbb{C}} \\ (t, z, w) &\mapsto w - \mathsf{A}_{\mathbb{C}}(z). \end{aligned}$$

•

A ***solution*** for $F_{\mathbb{C}}$ is a locally absolutely continuous map $\zeta \colon \mathbb{T} \to \mathsf{X}_{\mathbb{C}}$ that satisfies

$$\dot{\zeta}(t) = \mathsf{A}_{\mathbb{C}}(\zeta(t)).$$

Note that, as $\mathsf{X}_{\mathbb{C}} = \mathsf{X} \times \mathsf{X}$, we can write $\zeta(t) = (\xi(t), \eta(y))$ for locally absolutely continuous maps $\xi, \eta \colon \mathbb{T} \to \mathsf{X}$ that are the ***real part*** and ***imaginary part*** of ζ, respectively.

As in the scalar case, the real and imaginary parts of a solution separately satisfy the uncomplexified differential equation.

5.2.18 Lemma (Real and imaginary parts of complex solutions are solutions) *Consider the system of linear homogeneous ordinary differential equations* F *with constant coefficients, with right-hand side (5.7) and with complexification* $\mathrm{F}_{\mathbb{C}}$*. If* $\zeta \colon \mathbb{T} \to \mathsf{X}_{\mathbb{C}}$ *is a solution for* $\mathrm{F}_{\mathbb{C}}$*, then* $\mathrm{Re}(\zeta)$ *and* $\mathrm{Im}(\zeta)$ *are solutions for* F*.*

Proof Given $\zeta \colon \mathbb{T} \to \mathsf{X}_{\mathbb{C}}$ we write $\zeta(t) = (\xi(t), \eta(t))$ so that $\xi = \mathrm{Re}(\zeta)$ and $\eta = \mathrm{Im}(\zeta)$. Since ζ is a solution for $F_{\mathbb{C}}$, we have

$$\dot{\zeta}(t) = (\dot{\xi}(t), \dot{\eta}(t)) = \mathsf{A}_{\mathbb{C}}(\zeta(t)) = (\mathsf{A}(\xi(t)), \mathsf{A}(\eta(t))$$

by definition of $\mathsf{A}_{\mathbb{C}}$. Equating the second and fourth terms in this string of equalities gives the lemma. ■

5.2.2.2 The operator exponential In this section we consider the constant coefficient version of the continuous-time state transition map.

5.2.19 Definition (Operator exponential) Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, X be a finite-dimensional $\mathbb{F}$-vector space, and let $L \in \mathrm{L}(X; X)$. The ***operator exponential*** of L is the linear map $e^L \in \mathrm{L}(X; X)$ defined by $e^L = \Phi^c_A(1, 0)$, where $A\colon [0, 1] \to \mathrm{L}(X; X)$ is defined by $A(t) = L$ for all $t \in [0, 1]$. •

What we call the "operator exponential" will almost universally be called the "matrix exponential" because it is defined as we have defined it, but in the case where $X = \mathbb{R}^n$ and so L is an $n \times n$ matrix. Since we work with abstract vector spaces, our terminology in perhaps better suited to our setting.

Let us give some alternative characterisations and properties of the operator exponential.

5.2.20 Theorem (Properties of the operator exponential) *Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let X be a finite-dimensional $\mathbb{F}$-vector space, and let $L, M \in \mathrm{L}(X; X)$. Then the following statements hold:*

(i) $e^L = \mathrm{id}_X + \sum_{k=1}^{\infty} \frac{L^k}{k!}$;

(ii) *if $\mathbb{F} = \mathbb{C}$, then e^L is a $\mathbb{C}$-linear map;*

(iii) $\frac{d}{dt} e^{Lt} = L \circ e^{Lt} = e^{Lt} \circ L$;

(iv) $e^0 = \mathrm{id}_X$;

(v) *for $\alpha \in \mathbb{F}$, $e^\alpha \, \mathrm{id}_X = e^{\alpha \, \mathrm{id}_X}$;*

(vi) $e^{Lt} \circ e^{Ms} = e^{Lt+sM}$ *for all $s, t \in \mathbb{R}$ if and only if $L \circ M = M \circ L$;*

(vii) e^L *is invertible and* $(e^L)^{-1} = e^{-L}$;

(viii) *if $U \subseteq X$ is L-invariant, then it is also e^L-invariant;*

(ix) *the solution to the initial value problem*

$$\dot{\xi}(t) = L(\xi(t)), \quad \xi(t_0) = x_0,$$

is $\xi(t) = e^{L(t-t_0)}(x_0)$.

Proof (i) Let $A\colon [0, 1] \to \mathrm{L}(X; X)$ be defined by $A(t) = L$ for $t \in [0, 1]$. Adapting the notation of Section 5.2.1.3, if we define

$$I_k = \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{k-1}} A(t_1) \circ A(t_2) \circ \cdots \circ A(t_k) \, dt_k \cdots dt_2 dt_1,$$

then

$$e^A = \mathrm{id}_X + \sum_{k=1}^{\infty} I_k,$$

and we know the series converges by virtue of Theorem 5.2.12. Note that

$$I_k = L^k \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{k-1}} dt_k \cdots dt_2 dt_1 = \frac{L^k}{k!},$$

which can be proved by an application of Lemma 1 from the proof of Theorem 5.2.12. This part of the result then follows.

(ii) Since e^{L} is $\mathbb{R}$-linear, we have

$$e^{L}(v_1 + v_2) = e^{L}(v_1) + e^{L}(v_2).$$

Now let $v \in X$ and $a \in \mathbb{C}$. We have

$$e^{L}(av) = av + \sum_{k=1}^{\infty} \frac{L^k}{k!}(av) = a\left(v + \sum_{k=1}^{\infty} \frac{L^k}{k!}(v)\right) = a \exp^{L}(v),$$

using part (i) and $\mathbb{C}$-linearity of L, and hence also of L^k for every $k \in \mathbb{Z}_{>0}$.

(iii) As we say in the proof of Theorem 5.2.12, both series

$$\sum_{k=0}^{\infty} \frac{L^k t^k}{k!},$$

and the series

$$\sum_{k=1}^{\infty} \frac{L^k t^{k-1}}{(k-1)!} = L \circ \left(\sum_{k=0}^{\infty} \frac{L^k t^k}{k!}\right) = \left(\sum_{k=0}^{\infty} \frac{L^k t^k}{k!}\right) \circ L.$$

of term-by-term derivatives with respect to t, converge uniformly on any bounded time-domain. Therefore,

$$\frac{d}{dt} e^{Lt} = L \circ e^{Lt} = e^{Lt} \circ L.$$

(iv) This follows from part (i).

(v) By part (i) we have

$$e^{\alpha \operatorname{id}_X} = \operatorname{id}_X + \sum_{k=1}^{\infty} \frac{\operatorname{id}_X^k \alpha^k}{k!} = \left(1 + \sum_{k=1}^{\infty} \frac{\alpha^k}{k!}\right) \operatorname{id}_X = e^{\alpha} \operatorname{id}_X,$$

as desired.

(vi) Suppose that $L \circ M = M \circ L$. This gives

$$(L + M)^k = \sum_{j=0}^{k} \binom{k}{j} L^j M^{k-j},$$

using the Binomial Formula. (Note that this *does* require that $L \circ M = M \circ L$.) Then

$$\begin{aligned} e^{Lt+Ms} &= \operatorname{id}_X + \sum_{k=1} \frac{(Lt + Ms)^k}{k!} \\ &= \operatorname{id}_X + \sum_{k=1}^{\infty} \sum_{j=0}^{k} \frac{L^j t^j M^{k-j} s^{k-j}}{j!(k-j)!} \\ &= \left(\operatorname{id}_X + \sum_{j=1}^{\infty} \frac{L^j t^j}{j!}\right)\left(\operatorname{id}_X + \sum_{k=1}^{\infty} \frac{M^k s^k}{k!}\right) \end{aligned}$$

for all $t \in \mathbb{R}$.

Now suppose that $e^{Lt+Ms} = e^{Lt} \circ e^{Ms}$ for all $s, t \in \mathbb{R}$. We then compute, taking $s = t$,

$$\frac{d}{dt} e^{(L+M)t} = (L + M) \circ e^{(L+M)t}$$

and

$$\frac{d}{dt} e^{Lt} \circ e^{Mt} = L \circ e^{Lt} \circ e^{Mt} + e^{Lt} \circ M \circ e^{Mt}.$$

Next

$$\frac{d^2}{dt^2} e^{(L+M)t} = (L + M)^2 e^{(L+M)t}$$

and

$$\frac{d^2}{dt^2} e^{Lt} \circ e^{Mt} = L^2 \circ e^{Lt} \circ e^{Mt} + L \circ e^{Lt} \circ M \circ e^{Mt} + L \circ e^{Lt} \circ M \circ e^{Mt} + e^{Lt} \circ M^2 \circ e^{Mt}.$$

Evaluating the two second-derivatives at $t = 0$ and equating them gives

$$\begin{aligned} & (L + M)^2 = L^2 + 2L \circ M + M^2 \\ \implies \quad & L^2 + L \circ M + M \circ L + M^2 = L^2 + 2L \circ M + M^2 \\ \implies \quad & M \circ L = L \circ M, \end{aligned}$$

as desired.

(vii) That e^L is invertible is a consequence of its definition and Theorem 5.2.6(v). By parts (iv) and (vi), we have

$$\mathrm{id}_X = e^{L-L} = e^L e^{-L},$$

from which we conclude that $(e^L)^{-1} = e^{-L}$.

(viii) Let U be an L-invariant subspace of X and let $u \in U$. We claim that U is also L^k-invariant for every $k \in \mathbb{Z}_{>0}$. This we prove by induction, it obviously being true when $k = 1$. Suppose it true for $k = m$ and let $u \in U$. Then $L^{m+1}(u) = L \circ L^m(u)$. Since $L^m(u) \in U$ and since U is L-invariant, we immediately have $L^{m+1}(u) \in U$, showing that, indeed, U is L^k-invariant for every $k \in \mathbb{Z}_{>0}$. Using part (i) we then have

$$\left(\mathrm{id}_X + \sum_{k=1}^{m} \frac{L^k}{k!} \right)(u) = u + \sum_{k=1}^{m} \frac{L^k(u)}{k!} \in U.$$

Thus we have the sequence $(u_m)_{m \in \mathbb{Z}_{>0}}$ in X given by

$$u_m = u + \sum_{k=1}^{m} \frac{L^k(u)}{k!}.$$

Since U is closed by Corollary III-3.6.19, we have

$$e^L(u) = u + \sum_{k=1}^{\infty} \frac{L^k(u)}{k!} = \lim_{m \to \infty} u_m \in U,$$

as desired.

(ix) Using (iii) we compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{\mathsf{L}(t-t_0)}(x_0) = \mathsf{L} \circ \mathrm{e}^{\mathsf{L}(t-t_0)}(x_0).$$

We also have $\mathrm{e}^{\mathsf{L}(t-t_0)}(x_0)$, when evaluated at $t = t_0$, is x_0 by part (iv). Thus $t \mapsto \mathrm{e}^{\mathsf{L}(t-t_0)}(x_0)$ does indeed satisfy the stated initial value problem. ■

5.2.21 Remark ($\mathrm{e}^{\mathsf{L}} \circ \mathrm{e}^{\mathsf{M}} = \mathrm{e}^{\mathsf{L}+\mathsf{M}}$ does not imply $\mathsf{L} \circ \mathsf{M} = \mathsf{M} \circ \mathsf{L}$) Let $\mathsf{X} = \mathbb{R}^3$ and define $\mathsf{L}, \mathsf{M} \in \mathrm{L}(\mathbb{R}^3; \mathbb{R}^3)$ by the matrices

$$\begin{bmatrix} 0 & 6\pi & 0 \\ -6\pi & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 8\pi \\ 0 & -8\pi & 0 \end{bmatrix},$$

respectively. Using Procedure 5.2.26 below, we can compute $\mathrm{e}^{\mathsf{L}} = \mathrm{e}^{\mathsf{M}} = \mathrm{e}^{\mathsf{L}+\mathsf{M}} = \mathrm{id}_{\mathbb{R}^3}$, and so $\mathrm{e}^{\mathsf{L}}\mathrm{e}^{\mathsf{M}} = \mathrm{e}^{\mathsf{L}+\mathsf{M}}$. However, we do not have $\mathsf{L} \circ \mathsf{M} = \mathsf{M} \circ \mathsf{L}$, as may be verified by a direct computation. •

Let us consider the representation of the operator exponential in a basis.

5.2.22 Proposition (The matrix representation of the operator exponential is the operator exponential of the matrix representation) *Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let X be an n-dimensional $\mathbb{F}$-vector space, let $\mathsf{L} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$, and let $\mathscr{B} = \{\mathrm{e}_1, \dots, \mathrm{e}_n\}$ be a basis for X. Then*

$$[\mathrm{e}^{\mathsf{L}}]^{\mathscr{B}}_{\mathscr{B}} = \mathrm{e}^{[\mathrm{L}]^{\mathscr{B}}_{\mathscr{B}}}.$$

Proof This follows from the definition of the operator exponential and Exercise 5.2.2. ■

5.2.2.3 Bases of solutions Now, for equations with constant coefficients, we construct "explicitly" a basis for $\mathrm{Sol}(F)$.

5.2.23 Procedure (Basis of solutions for a system of linear homogeneous ordinary differential equations with constant coefficients) Given a system of linear homogeneous ordinary differential equations

$$F \colon \mathbb{T} \times \mathsf{X} \oplus \mathsf{X} \to \mathsf{X}$$

in an n-dimensional $\mathbb{R}$-vector space X and with right-hand side

$$\widehat{F}(t, x) = \mathsf{A}(x),$$

do the following.

1. Choose a basis $\{e_1, \dots, e_n\}$ for X. Let A be the matrix representative of A with respect to this basis. If $\mathsf{X} = \mathbb{R}^n$, one can just take A to be the usual matrix associated with $\mathsf{A} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$.

2. Compute the characteristic polynomial $P_{\mathsf{A}} = \det(X\boldsymbol{I}_n - \boldsymbol{A})$.
3. Compute the roots of P_{A}, i.e., the eigenvalues of $\mathsf{A}_{\mathbb{C}}$, and organise them as follows. We have distinct real eigenvalues

$$\ell_1, \dots, \ell_r$$

and distinct complex eigenvalues

$$\lambda_1 = \sigma_1 + \mathrm{i}\omega_1, \dots, \lambda_s = \sigma_2 + \mathrm{i}\omega_s,$$

$\omega_1, \dots, \omega_s \in \mathbb{R}_{>0}$, along with their complex conjugates.
4. Let $m_j = m_{\mathrm{a}}(\ell_j, \mathsf{A})$, $j \in \{1, \dots, r\}$, and $\mu_j = m_{\mathrm{a}}(\lambda_j, \mathsf{A})$, $j \in \{1, \dots, s\}$, be the algebraic multiplicities.
5. For $j \in \{1, \dots, r\}$, let $\{\boldsymbol{x}_{j,1}, \dots, \boldsymbol{x}_{j,m_j}\}$ be a basis for

$$\overline{\mathsf{W}}(\ell_j, \boldsymbol{A}) = \ker((\ell_j \boldsymbol{I}_n - \boldsymbol{A})^{m_j}).$$

6. For $j \in \{1, \dots, s\}$, let $\{\boldsymbol{z}_{j,1}, \dots, \boldsymbol{z}_{j,\mu_j}\}$ be a basis for

$$\overline{\mathsf{W}}(\lambda_j, \boldsymbol{A}_{\mathbb{C}}) = \ker((\lambda_j \boldsymbol{I}_n - \boldsymbol{A}_{\mathbb{C}})^{\mu_j}).$$

Write $\boldsymbol{z}_{j,k} = \boldsymbol{a}_{j,k} + \mathrm{i}\boldsymbol{b}_{j,k}$ for each $k \in \{1, \dots, \mu_j\}$. Then

$$\{\boldsymbol{a}_{j,1}, \boldsymbol{b}_{j,1}, \dots, \boldsymbol{a}_{j,\mu_j}, \boldsymbol{b}_{j,\mu_j}\}$$

is a basis for $\overline{\mathsf{W}}(\lambda_j, \boldsymbol{A})$.
7. For $j \in \{1, \dots, r\}$ and $k \in \{1, \dots, m_j\}$, define

$$\xi_{j,k}(t) = \mathrm{e}^{\ell_j t}\left(\boldsymbol{I}_n + \frac{(\boldsymbol{A} - \ell_j \boldsymbol{I}_n)t}{1!} + \cdots + \frac{(\boldsymbol{A} - \ell_j \boldsymbol{I}_n)^{m_j - 1} t^{m_j - 1}}{(m_j - 1)!}\right)\boldsymbol{x}_{j,k}.$$

8. For $j \in \{1, \dots, s\}$ and $k \in \{1, \dots, \mu_j\}$, define

$$\begin{aligned}\boldsymbol{\alpha}_{j,k}(t) = \mathrm{e}^{\sigma_j t}&\left(\left(\sum_{m=0}^{\mu_j - 1}\sum_{l=0}^{\lceil \frac{m-1}{2} \rceil} \frac{(-1)^l \omega_j^{2l} t^m}{(2l)!(m-2l)!}(\boldsymbol{A} - \sigma_j \boldsymbol{I}_n)^{m-2l}\right)(\cos(\omega_j t)\boldsymbol{a}_{j,k} - \sin(\omega_j t)\boldsymbol{b}_{j,k})\right.\\ &\left. - \left(\sum_{m=0}^{\mu_j - 1}\sum_{l=0}^{\lfloor \frac{m-1}{2} \rfloor} \frac{(-1)^{l+1} \omega_j^{2l+1} t^m}{(2l+1)!(m-2l-1)!}(\boldsymbol{A} - \sigma_j \boldsymbol{I}_n)^{m-2l-1}\right)(\cos(\omega_j t)\boldsymbol{b}_{j,k} + \sin(\omega_j t)\boldsymbol{a}_{j,k})\right)\end{aligned} \tag{5.9}$$

and

$$\boldsymbol{\beta}_{j,k}(t) = \mathrm{e}^{\sigma_j t}\left(\left(\sum_{m=0}^{\mu_j-1}\sum_{l=0}^{\lceil\frac{m-1}{2}\rceil}\frac{(-1)^l\omega_j^{2l}t^m}{(2l)!(m-2l)!}(\boldsymbol{A}-\sigma_j\boldsymbol{I}_n)^{m-2l}\right)(\cos(\omega_j t)\boldsymbol{b}_{j,k}+\sin(\omega_j t)\boldsymbol{a}_{j,k})\right.$$
$$\left.+\left(\sum_{m=0}^{\mu_j-1}\sum_{l=0}^{\lfloor\frac{m-1}{2}\rfloor}\frac{(-1)^{l+1}\omega_j^{2l+1}t^m}{(2l+1)!(m-2l-1)!}(\boldsymbol{A}-\sigma_j\boldsymbol{I}_n)^{m-2l-1}\right)(\cos(\omega_j t)\boldsymbol{a}_{j,k}-\sin(\omega_j t)\boldsymbol{b}_{j,k})\right), \tag{5.10}$$

where, for $x \in \mathbb{R}$, $\lfloor x \rfloor$ is greatest integer less than or equal to x and $\lceil x \rceil$ is smallest integer greater than or equal to x.

9. For $j \in \{1,\dots,r\}$ and $k \in \{1,\dots,m_j\}$, let $\xi_{j,k}\colon \mathbb{T} \to \mathsf{X}$ be the function whose components with respect to the basis $\{e_1,\dots,e_n\}$ are the components of $\boldsymbol{\xi}_{j,k}$.
10. For $j \in \{1,\dots,s\}$ and $k \in \{1,\dots,\mu_j\}$, let $\alpha_{j,k}, \beta_{j,k}\colon \mathbb{T} \to \mathsf{X}$ be the functions whose components with respect to the basis $\{e_1,\dots,e_n\}$ are the components of $\boldsymbol{\alpha}_{j,k}$ and $\boldsymbol{\beta}_{j,k}$, respectively.
11. Then the n functions

$$\begin{aligned} &\xi_{j,k}, && j \in \{1,\dots,r\},\ k \in \{1,\dots,m_j\},\\ &\alpha_{j,k}, \beta_{j,k}, && j \in \{1,\dots,s\},\ k \in \{1,\dots,\mu_j\}, \end{aligned}$$

are a basis for Sol(F). •

Of course, we should verify that the procedure does, indeed, produce a basis for Sol(F).

5.2.24 Theorem (Basis of solutions for a system of linear homogeneous ordinary differential equations with constant coefficients) *Given a system of linear homogeneous ordinary differential equations*

$$\mathrm{F}\colon \mathbb{T} \times \mathsf{X} \oplus \mathsf{X} \to \mathsf{X}$$

in an n*-dimensional* $\mathbb{R}$*-vector space* X *and with right-hand side*

$$\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}) = \mathsf{A}(\mathrm{x}),$$

define n *functions as in Procedure 5.2.23. Then these functions for a basis for* Sol(F).

Proof By virtue of Exercise 5.2.1, we can choose a basis $\{e_1,\dots,e_n\}$ for X and so assume that $\mathsf{X} = \mathbb{R}^n$.

Let us first fix $j \in \{1,\dots,r\}$ and show that $\boldsymbol{\xi}_{j,k}$, $k \in \{1,\dots,m_j\}$, are solutions for $\boldsymbol{F}$. Let $t_0 \in \mathbb{T}$. Let us also fix $k \in \{1,\dots,m_j\}$. By Theorem 5.2.20(ix), the unique solution to the initial value problem

$$\dot{\xi}(t) = \boldsymbol{A}\xi(t), \quad \xi(t_0) = \mathrm{e}^{\boldsymbol{A}t_0}\boldsymbol{x}_{j,k},$$

is

$$t \mapsto \mathrm{e}^{\boldsymbol{A}(t-t_0)}\mathrm{e}^{\boldsymbol{A}t_0}\boldsymbol{x}_{j,k} = \mathrm{e}^{\boldsymbol{A}t}\boldsymbol{x}_{j,k},$$

using Theorem 5.2.20(vi) and the obvious fact that the matrices $t\boldsymbol{A}$ and $t_0\boldsymbol{A}$ commute. Now we have

$$\mathrm{e}^{\boldsymbol{A}t}\boldsymbol{x}_{j,k} = \mathrm{e}^{\ell_j t\boldsymbol{I}_n}\mathrm{e}^{(\boldsymbol{A}-\ell_j\boldsymbol{I}_n)t} = \mathrm{e}^{\ell_j t}\mathrm{e}^{(\boldsymbol{A}-\ell_j\boldsymbol{I}_n)t}\boldsymbol{x}_{j,k}$$

using parts (v) and (vi) of Theorem 5.2.20. Now, since $\boldsymbol{x}_{j,k} \in \overline{\mathrm{W}}(\ell_j, \boldsymbol{A})$,

$$\mathrm{e}^{\ell_j t}\mathrm{e}^{(\boldsymbol{A}-\ell_j\boldsymbol{I}_n)t}\boldsymbol{x}_{j,k} = \mathrm{e}^{\ell_j t}\sum_{m=0}^{m_j-1}\frac{(\boldsymbol{A}-\ell_j\boldsymbol{I}_n)^m t^m}{m!}\boldsymbol{x}_{j,k},$$

using Theorem 5.2.20(i). However, this last expression is exactly $\boldsymbol{\xi}_{j,k}(t)$, showing that this is indeed a solution for $\boldsymbol{F}$.

Next we show that, still keeping $j \in \{1, \dots, r\}$ fixed, the m_j solutions $\boldsymbol{\xi}_{j,k}$, $k \in \{1, \dots, m_j\}$, are linearly independent. As we have seen,

$$\boldsymbol{\xi}_{j,k}(t_0) = \mathrm{e}^{\boldsymbol{A}t_0}\boldsymbol{x}_{j,k}, \qquad k \in \{1, \dots, m_j\}.$$

Thus, for $c_1, \dots, c_{m_j} \in \mathbb{R}$, we have

$$\begin{aligned}
& c_1\boldsymbol{\xi}_{j,k}(t_0) + \cdots + c_{m_j}\boldsymbol{\xi}_{j,m_j}(t_0) = \boldsymbol{0} \\
\Longrightarrow \quad & c_1\mathrm{e}^{\boldsymbol{A}t_0}\boldsymbol{x}_{j,1} + \cdots + c_{m_j}\mathrm{e}^{\boldsymbol{A}t_0}\boldsymbol{x}_{j,m_j} = \boldsymbol{0} \\
\Longrightarrow \quad & \mathrm{e}^{\boldsymbol{A}t_0}(c_1\boldsymbol{x}_{j,1} + \cdots + c_{m_j}\boldsymbol{x}_{j,m_j}) = \boldsymbol{0} \\
\Longrightarrow \quad & c_1\boldsymbol{x}_{j,1} + \cdots + c_{m_j}\boldsymbol{x}_{j,m_j} = \boldsymbol{0} \\
\Longrightarrow \quad & c_1 = \cdots = c_{m_j} = 0,
\end{aligned}$$

since $\boldsymbol{x}_{j,1}, \dots, x_{j,m_j}$ are constructed as being linearly independent. By Corollary 5.2.4 we conclude that $\boldsymbol{\xi}_{j,1}, \dots, \boldsymbol{\xi}_{j,m_j}$ are indeed linearly independent.

Now we fix $j \in \{1, \dots, s\}$ and work with the complex eigenvalue $\lambda_j = \sigma_j + \mathrm{i}\omega_j$. First of all, let us define $\boldsymbol{\zeta}_{j,k} \colon \mathbb{T} \to \mathbb{C}^n$, $k \in \{1, \dots, \mu_j\}$, by

$$\boldsymbol{\zeta}_{j,k} = \mathrm{e}^{\boldsymbol{A}_{\mathbb{C}}t}\boldsymbol{z}_{j,k}.$$

Then, exactly as above for the real eigenvalues, we have

$$\boldsymbol{\zeta}_{j,k}(t) = \mathrm{e}^{\lambda_j t}\sum_{m=0}^{\mu_j-1}\frac{(\boldsymbol{A}_{\mathbb{C}}-\lambda_j\boldsymbol{I}_n)^m t^m}{m!}\boldsymbol{z}_{j,k}.$$

Moreover, $\boldsymbol{\zeta}_{j,k}$, $k \in \{1, \dots, \mu_j\}$, are solutions for $\boldsymbol{F}_{\mathbb{C}}$. Therefore, by Lemma 5.2.18, the real and imaginary parts of $\boldsymbol{\zeta}_{j,k}$ are solutions for $\boldsymbol{F}$. To determine the real and imaginary parts, we first make use of the following lemma.

1 Lemma *For a* $\mathbb{C}$*-vector space* X*, for* $\mathsf{L} \in \mathrm{L}(\mathsf{X};\mathsf{X})$*, for* $\mathrm{b} \in \mathbb{C}$*, and for* $\mathrm{m} \in \mathbb{Z}_{\geq 0}$*,*

$$(\mathsf{L} + \mathrm{i}\,\mathrm{b}\,\mathrm{id}_{\mathsf{X}})^{\mathrm{m}} = \sum_{\mathrm{j}=0}^{\lceil \frac{\mathrm{m}-1}{2} \rceil} \binom{\mathrm{m}}{2\mathrm{j}} (-1)^{\mathrm{j}} \mathrm{b}^{2\mathrm{j}} \mathsf{L}^{\mathrm{m}-2\mathrm{j}} + \mathrm{i} \sum_{\mathrm{j}=0}^{\lfloor \frac{\mathrm{m}-1}{2} \rfloor} \binom{\mathrm{m}}{2\mathrm{j}+1} (-1)^{\mathrm{j}} (\mathrm{b}^{2\mathrm{j}+1} \mathsf{L}^{\mathrm{m}-2\mathrm{j}-1}).$$

Proof By the Binomial Formula, and since L and id_{X} commute, we have

$$(\mathsf{L} + \mathrm{i}\, b\, \mathrm{id}_{\mathsf{X}}) = \sum_{j=0}^{m} \binom{m}{j} (\mathrm{i}\, b)^j \mathsf{L}^{m-1}.$$

The stated formula is obtained by separating this expression into its real and imaginary parts. ▼

With the lemma, one verifies that

$$\mathrm{Re}\left(\sum_{m=0}^{\mu_j - 1} \frac{(\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m t^m}{m!} \right) = \sum_{m=0}^{\mu_j - 1} \sum_{l=0}^{\lceil \frac{m-1}{2} \rceil} \frac{(-1)^l \omega_j^{2l} t^m}{(2l)!(m-2l)!} (\boldsymbol{A} - \sigma_j \boldsymbol{I}_n)^{m-2l}$$

and

$$\mathrm{Im}\left(\sum_{m=0}^{\mu_j - 1} \frac{(\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m t^m}{m!} \right) = \sum_{m=0}^{\mu_j - 1} \sum_{l=0}^{\lfloor \frac{m-1}{2} \rfloor} \frac{(-1)^{l+1} \omega_j^{2l+1} t^m}{(2l+1)!(m-2l-1)!} (\boldsymbol{A} - \sigma_j \boldsymbol{I}_n)^{m-2l-1}.$$

Some tedious manipulations, one can then verify that

$$\begin{aligned}
\boldsymbol{\alpha}_{j,k}(t) &= \mathrm{Re}\left(\mathrm{e}^{\lambda_j t} \left(\boldsymbol{I}_n + \frac{(\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t}{1!} + \cdots + \frac{((\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t)^{\mu_j - 1}}{(\mu_j - 1)!} \right) \boldsymbol{z}_{j,k} \right), \\
\boldsymbol{\beta}_{j,k}(t) &= \mathrm{Im}\left(\mathrm{e}^{\lambda_j t} \left(\boldsymbol{I}_n + \frac{(\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t}{1!} + \cdots + \frac{((\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t)^{\mu_j - 1}}{(\mu_j - 1)!} \right) \boldsymbol{z}_{j,k} \right)
\end{aligned}$$

for $k \in \{1, \ldots, \mu_j\}$. This shows that $\boldsymbol{\alpha}_{j,k}$ and $\boldsymbol{\beta}_{j,k}$ are solutions for $\boldsymbol{F}$ for $k \in \{1, \ldots, \mu_j\}$.

Now we verify that

$$\boldsymbol{\alpha}_{j,1}, \ldots, \boldsymbol{\alpha}_{j,\mu_j}, \boldsymbol{\beta}_{j,1}, \ldots, \boldsymbol{\beta}_{j,\mu_j}$$

are linearly independent. As above in the real case, the complex solutions $\boldsymbol{\zeta}_{j,1}, \ldots, \boldsymbol{\zeta}_{j,\mu_j}$ for $\boldsymbol{F}_{\mathbb{C}}$ are linearly independent. Now let $t_0 \in \mathbb{T}$ and $c_1, \ldots, c_{\mu_j}, d_1, \ldots, d_{\mu_j} \in \mathbb{R}$, and note

that

$$\begin{aligned}
&\sum_{k=1}^{\mu_j}(c_k\boldsymbol{\alpha}_{j,k}(t_0)+d_k\boldsymbol{\beta}_{j,k}(t_0))=\boldsymbol{0}\\
\implies\quad&\sum_{k=1}^{\mu_j}(c_k\operatorname{Re}(\boldsymbol{\zeta}_{j,k})(t_0)+d_k\operatorname{Im}(\boldsymbol{\zeta}_{j,k})(t_0))=\boldsymbol{0}\\
\implies\quad&\sum_{k=1}^{\mu_j}(c_k\operatorname{Re}(\mathrm{e}^{\boldsymbol{A}_{\mathbb{C}}t_0}\boldsymbol{z}_{j,k})+d_k\operatorname{Im}(\mathrm{e}^{\boldsymbol{A}_{\mathbb{C}}t_0}\boldsymbol{z}_{j,k}))\\
\implies\quad&\sum_{k=1}^{\mu_j}(c_k\mathrm{e}^{\boldsymbol{A}_{\mathbb{C}}t_0}\boldsymbol{a}_{j,k}+d_k\mathrm{e}^{\boldsymbol{A}_{\mathbb{C}}t_0}\boldsymbol{b}_{j,k})=\boldsymbol{0}\\
\implies\quad&\sum_{k=1}^{\mu_j}(c_k\boldsymbol{a}_{j,k}+d_k\boldsymbol{b}_{j,k})=\boldsymbol{0}\\
\implies\quad&c_1=\cdots=c_{\mu_j}=d_1=\cdots=d_{\mu_j}=0,
\end{aligned}$$

using the fact that, since $\boldsymbol{A}$ is real, $\mathrm{e}^{\boldsymbol{A}_{\mathbb{C}}t_0}$ is also real and, using Theorem I-5.4.68(i), this gives the linear independence of

$$\boldsymbol{\alpha}_{j,1},\dots,\boldsymbol{\alpha}_{j,\mu_j},\boldsymbol{\beta}_{j,1},\dots,\boldsymbol{\beta}_{j,\mu_j},$$

as claimed.

Now we have $m_1+\cdots+m_r+2(\mu_1+\cdots+\mu_s)=n$ solutions for $\boldsymbol{F}$. It remains to show that the collection of all of these solutions are linearly independent. Let us suppose that

$$\begin{aligned}
&\underbrace{c_{1,1}\boldsymbol{\xi}_{1,1}(t)+\cdots+c_{1,m_1}\boldsymbol{\xi}_{1,m_1}(t)}_{\in\overline{\mathsf{W}}(\ell_1,\boldsymbol{A})}+\cdots+\underbrace{c_{r,1}\boldsymbol{\xi}_{r,1}(t)+\cdots+c_{r,m_r}\boldsymbol{\xi}_{r,m_r}(t)}_{\in\overline{\mathsf{W}}(\ell_r,\boldsymbol{A})}\\
&\quad+\underbrace{d_{1,1}\boldsymbol{a}_{1,1}(t)+\cdots+d_{1,\mu_1}\boldsymbol{a}_{1,\mu_1}(t)}_{\in\overline{\mathsf{W}}(\lambda_1,\boldsymbol{A})}+\cdots+\underbrace{d_{s,1}\boldsymbol{a}_{s,1}(t)+\cdots+d_{s,\mu_s}\boldsymbol{a}_{s,\mu_s}(t)}_{\in\overline{\mathsf{W}}(\lambda_s,\boldsymbol{A})}\\
&\quad+\underbrace{e_{1,1}\boldsymbol{b}_{1,1}(t)+\cdots+e_{1,\mu_1}\boldsymbol{b}_{1,\mu_1}(t)}_{\in\overline{\mathsf{W}}(\lambda_1,\boldsymbol{A})}+\cdots+\underbrace{e_{s,1}\boldsymbol{b}_{s,1}(t)+\cdots+e_{s,\mu_s}\boldsymbol{b}_{s,\mu_s}(t)}_{\in\overline{\mathsf{W}}(\lambda_s,\boldsymbol{A})}=\boldsymbol{0},
\end{aligned}$$

for suitable scalar coefficients. Since the generalised eigenspaces intersect in $\{0\}$ by Proposition I-5.4.60, and since the generalised eigenspaces are invariant under $\mathrm{e}^{\boldsymbol{A}t}$ for all $t\in\mathbb{T}$ by Theorem 5.2.20(viii), for the preceding equation to hold, each of its components in each of the generalised eigenspaces must be zero, i.e.,

$$c_{j,1}\boldsymbol{\xi}_{j,1}(t)+\cdots+c_{j,m_j}\boldsymbol{\xi}_{j,m_j}(t)=\boldsymbol{0},\qquad j\in\{1,\dots,r\},$$

and

$$d_{j,1}\boldsymbol{a}_{j,1}(t)+\cdots+d_{j,\mu_j}\boldsymbol{a}_{j,\mu_j}(t)+e_{j,1}\boldsymbol{b}_{j,1}(t)+\cdots+e_{j,\mu_j}\boldsymbol{b}_{j,\mu_j}(t)=\boldsymbol{0},\qquad j\in\{1,\dots,s\}.$$

This implies that all coefficients must be zero, since we have already shown the linear independence of the solutions with initial conditions in each of the subspaces $\overline{\mathsf{W}}(\ell_l, \boldsymbol{A})$, $j \in \{1, \dots, r\}$, and $\overline{\mathsf{W}}(\lambda_j, \boldsymbol{A})$, $j \in \{1, \dots, s\}$. Thus we have the desired linear independence, and thus the theorem follows. ■

From the proof of the theorem, we provide the following comment on how one might deal with complex eigenvalues in practice.

5.2.25 Remark (Computing solutions associated with complex eigenvalues) The formulae (5.9) and (5.10) of Procedure 5.2.23, while fun to look at, are typically not the best ways to work out solutions associated with complex eigenvalues. However, the proof of the preceding theorem tells us an alternative that is easier in easy examples (although using a computer algebra package is even easier). Indeed, in the proof we saw that

$$\begin{aligned}\boldsymbol{\alpha}_{j,k}(t) &= \operatorname{Re}\left(\mathrm{e}^{\lambda_j t}\left(\boldsymbol{I}_n + \frac{(\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t}{1!} + \cdots + \frac{((\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t)^{\mu_j - 1}}{(\mu_j - 1)!}\right)\boldsymbol{z}_{j,k}\right),\\ \boldsymbol{\beta}_{j,k}(t) &= \operatorname{Im}\left(\mathrm{e}^{\lambda_j t}\left(\boldsymbol{I}_n + \frac{(\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t}{1!} + \cdots + \frac{((\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t)^{\mu_j - 1}}{(\mu_j - 1)!}\right)\boldsymbol{z}_{j,k}\right)\end{aligned}$$

for $k \in \{1, \dots, \mu_j\}$. Thus, in practice, one might simply compute

$$\boldsymbol{\zeta}_{j,k}(t) = \mathrm{e}^{\lambda_j t}\left(\boldsymbol{I}_n + \frac{(\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t}{1!} + \cdots + \frac{((\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)t)^{\mu_j - 1}}{(\mu_j - 1)!}\right)\boldsymbol{z}_{j,k},$$

$k \in \{1, \dots, s\}$, and simply takes its real and imaginary parts as linearly independent solutions. •

We can now give an algorithm for computing, in principle, the operator exponential. The following procedure, while given for computing e^{A}, obviously may be used as well to compute the continuous-time state transition matrix $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0) = \mathrm{e}^{\mathsf{A}(t-t_0)}$ for a system of linear homogeneous ordinary differential equations with constant coefficients.

5.2.26 Procedure (Operator exponential) Given an n-dimensional $\mathbb{R}$-vector space X and $\mathsf{A} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$, do the following.

1. Choose a basis $\{e_1, \dots, e_n\}$ and let $\boldsymbol{A}$ be the matrix representative of A. If $\mathsf{X} = \mathbb{R}^n$, one can just take $\boldsymbol{A}$ to be the usual matrix associated with $\mathsf{A} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$.
2. Using Procedure 5.2.23, determine a fundamental set of solutions $\xi_1, \dots, \xi_n$, defined on all of $\mathbb{R}$, for the system of linear homogeneous ordinary differential equations $\boldsymbol{F}$ in $\mathbb{R}^n$ with right-hand side

$$\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}.$$

3. Define
$$\Xi(t) = \begin{bmatrix} \xi_{1,1}(t) & \xi_{2,1}(t) & \cdots & \xi_{n,1}(t) \\ \xi_{1,2}(t) & \xi_{2,2}(t) & \cdots & \xi_{n,2}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{1,n}(t) & \xi_{2,n}(t) & \cdots & \xi_{n,n}(t) \end{bmatrix},$$
where $\xi_{j,k}$ is the kth component of ξ_j.
4. Using Procedure 5.2.8 calculate
$$\mathrm{e}^{At} = \Phi_A^{\mathrm{c}}(t, 0) = \Xi(t)\Xi(0)^{-1}.$$
5. We then have e^{A} as the linear map whose matrix representative is e^A. •

5.2.2.4 Some examples Obviously, carrying out Procedure 5.2.23 for a moderately complicated linear transformation A is not something one would want to do more than once a day, and that once a day for at most a week or so. Because I am very manly, I did this four times in one day.

5.2.27 Example (Simple 2 × 2 example) We take $\mathsf{X} = \mathbb{R}^2$ and let $A \in \mathrm{L}(\mathbb{R}^2; \mathbb{R}^2)$ be defined by the matrix
$$A = \begin{bmatrix} -7 & 4 \\ -6 & 3 \end{bmatrix}.$$
The characteristic polynomial for A is
$$P_A = X^2 + 4X + 3 = (X + 1)(X + 3).$$
Thus the eigenvalues for A are $\ell_1 = -1$ and $\ell_2 = -3$. Since the eigenvalues are distinct, the algebraic and geometric multiplicities will be equal, and the generalised eigenvectors will simply be eigenvectors. An eigenvector for $\ell_1 = -1$ is $x_{1,1} = (2, 3)$ and an eigenvector for $\ell_2 = -3$ is $x_{2,1} = (1, 1)$. Procedure 5.2.23 then gives two linearly independent solutions as
$$\xi_{1,1}(t) = \mathrm{e}^{-t}\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \xi_{2,1}(t) = \mathrm{e}^{-3t}\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$
Thus we have determined a fundamental matrix to be
$$\Xi(t) = \begin{bmatrix} 2\mathrm{e}^{-t} & \mathrm{e}^{-3t} \\ 3\mathrm{e}^{-2t} & \mathrm{e}^{-3t} \end{bmatrix},$$
by assembling the linearly independent solutions into the columns of this matrix. It is then an easy calculation to arrive at
$$\mathrm{e}^{At} = \Xi(t)\Xi(0)^{-1} = \begin{bmatrix} 3\mathrm{e}^{-3t} - 2\mathrm{e}^{-t} & -2\mathrm{e}^{-3t} + 2\mathrm{e}^{-t} \\ 3\mathrm{e}^{-3t} - 3\mathrm{e}^{-t} & -2\mathrm{e}^{-3t} + 3\mathrm{e}^{-t} \end{bmatrix}$$
•

5.2.28 Example (A 3 × 3 example with multiplicity) We take $X = \mathbb{R}^3$ with the linear map $A \in L(\mathbb{R}^3; \mathbb{R}^3)$ determined by the matrix A more interesting case is the following:

$$A = \begin{bmatrix} -2 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Since the matrix is upper triangular, the eigenvalues are the diagonal elements: $\ell_1 = -2$ and $\ell_2 = -1$. The algebraic multiplicity of ℓ_1 is 2. However, we readily see that $\dim_{\mathbb{R}}(\ker(\ell_1 I_3 - A)) = 1$ and so the geometric multiplicity is 1. So we need to compute generalised eigenvectors in this case. We have

$$(A - \lambda_1 I_3)^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and the generalised eigenvectors span the kernel of this matrix, and so we may take $x_{1,1} = (1, 0, 0)$ and $x_{1,2} = (0, 1, 0)$ as generalised eigenvectors. Applying Procedure 5.2.23 gives

$$\begin{aligned} \xi_{1,1}(t) &= e^{-2t} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t e^{-2t} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} e^{-2t} \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \xi_{1,2}(t) &= e^{-2t} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + t e^{-2t} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} t e^{-2t} \\ e^{-2t} \\ 0 \end{bmatrix}. \end{aligned}$$

Finally we determine that $x_{2,1} = (0, 0, 1)$ is an eigenvector corresponding to $\ell_2 = -1$, and so this gives the solution

$$\xi_{2,1}(t) = \begin{bmatrix} 0 \\ 0 \\ e^{-t} \end{bmatrix}.$$

Thus we arrive at our three linearly independent solutions. We assemble these into the columns of a matrix to determine a fundamental matrix

$$\Xi(t) = \begin{bmatrix} e^{-2t} & t e^{-2t} & 0 \\ 0 & e^{-2t} & 0 \\ 0 & 0 & e^{-t} \end{bmatrix}.$$

It so happens that in this example we lucked out and $e^{At} = \Xi(t)$ since $\Xi(0) = I_3$. •

5.2.29 Example (A simple example with complex eigenvalues) Here we take $\mathsf{X} = \mathbb{R}^3$ with $\boldsymbol{A} \in \mathrm{L}(\mathbb{R}^3; \mathbb{R}^3)$ determined by the matrix

$$\boldsymbol{A} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix}.$$

The characteristic polynomial is

$$P_{\boldsymbol{A}} = X^3 + 4X^2 + 6X + 4.$$

One ascertains that the eigenvalues are then $\lambda_1 = -1 + \mathrm{i}$, $\bar{\lambda}_1 = -1 - \mathrm{i}$, $\ell_1 = -2$. Let's deal with the complex eigenvalue first, using Remark 5.2.25 rather than the complicated formulae (5.9) and (5.10) of Procedure 5.2.23 for complex eigenvalues. We have

$$\boldsymbol{A} - \lambda_1 \boldsymbol{I}_3 = \begin{bmatrix} -\mathrm{i} & 1 & 0 \\ -1 & -\mathrm{i} & 0 \\ 0 & 0 & -1 - \mathrm{i} \end{bmatrix},$$

from which we glean that an eigenvector is $\boldsymbol{z}_{1,1} = (-\mathrm{i}, 1, 0)$. Using Remark 5.2.25, the complex solution is then

$$\boldsymbol{\zeta}_{1,1}(t) = \mathrm{e}^{(-1+\mathrm{i})t} \begin{bmatrix} -\mathrm{i} \\ 1 \\ 0 \end{bmatrix}.$$

Using Euler's formula, $\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$, we have

$$\boldsymbol{\zeta}_{1,1}(t) = \mathrm{e}^{-t} \begin{bmatrix} -\mathrm{i}\cos t + \sin t \\ \cos t + \mathrm{i}\sin t \\ 0 \end{bmatrix} = \mathrm{e}^{-t} \begin{bmatrix} \sin t \\ \cos t \\ 0 \end{bmatrix} + \mathrm{i}\mathrm{e}^{-t} \begin{bmatrix} -\cos t \\ \sin t \\ 0 \end{bmatrix},$$

thus giving

$$\boldsymbol{\alpha}_{1,1}(t) = \mathrm{e}^{-t} \begin{bmatrix} \sin t \\ \cos t \\ 0 \end{bmatrix}, \quad \boldsymbol{\beta}_{1,1}(t) = \mathrm{e}^{-t} \begin{bmatrix} -\cos t \\ \sin t \\ 0 \end{bmatrix}.$$

Corresponding to the real eigenvalue ℓ_1 we readily determine that

$$\boldsymbol{\xi}_{1,1} = \mathrm{e}^{-2t} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

is a corresponding solution. This gives three linearly independent real solutions $\boldsymbol{\alpha}_{1,1}(t)$, $\boldsymbol{\beta}_{1,1}(t)$, and $\boldsymbol{\xi}_{1,1}(t)$. Putting these into the columns of a matrix gives a fundamental matrix

$$\boldsymbol{\Xi}(t) = \begin{bmatrix} \mathrm{e}^{-t}\sin t & -\mathrm{e}^{-t}\cos t & 0 \\ \mathrm{e}^{-t}\cos t & \mathrm{e}^{-t}\sin t & 0 \\ 0 & 0 & \mathrm{e}^{-2t} \end{bmatrix}.$$

A straightforward computation yields

$$e^{At} = \Xi(t)\Xi(0)^{-1} = \begin{bmatrix} e^{-t}\cos t & e^{-t}\sin t & 0 \\ -e^{-t}\sin t & e^{-t}\cos t & 0 \\ 0 & 0 & e^{-2t} \end{bmatrix}. \qquad \bullet$$

5.2.30 Example (An example of complex eigenvalues with multiplicity) Our final example has $\mathsf{X} = \mathbb{R}^4$ and $A \in \mathrm{L}(\mathbb{R}^4; \mathbb{R}^4)$ determined by the matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}.$$

The eigenvalues are determined to be $\lambda_1 = \mathrm{i}$ and $\bar{\lambda}_1 = -\mathrm{i}$, both with algebraic multiplicity 2. One readily determines that the kernel of $iI_4 - A$ is one-dimensional, and so the geometric multiplicity of these eigenvalues is just 1. Thus we need to compute complex generalised eigenvectors. We compute

$$(A - \mathrm{i}I_4)^2 = 2\begin{bmatrix} -1 & -\mathrm{i} & -\mathrm{i} & 1 \\ \mathrm{i} & -1 & -1 & -\mathrm{i} \\ 0 & 0 & -1 & -\mathrm{i} \\ 0 & 0 & \mathrm{i} & -1 \end{bmatrix}$$

and one checks that $z_{1,1} = (0, 0, -\mathrm{i}, 1)$ and $z_{1,2} = (-\mathrm{i}, 1, 0, 0)$ are two linearly independent generalised eigenvectors. We compute

$$(A - \mathrm{i}I_4)z_{1,1} = \begin{bmatrix} -\mathrm{i} \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad (A - \mathrm{i}I_4)z_{1,2} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We now determine the two linearly independent real solutions corresponding to $z_{1,1}$. We have

$$\zeta_{1,1}(t) = e^{\mathrm{i}t}(u_1 + t(A - \mathrm{i}I_4)z_{1,1}) = e^{\mathrm{i}t}\begin{bmatrix} 0 \\ 0 \\ -\mathrm{i} \\ 1 \end{bmatrix} + te^{\mathrm{i}t}\begin{bmatrix} -\mathrm{i} \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$= (\cos t + \mathrm{i}\sin t)\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + \mathrm{i}\begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} + t\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \mathrm{i}t\begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} t\sin t \\ t\cos t \\ \sin t \\ \cos t \end{bmatrix} + \mathrm{i}\begin{bmatrix} -t\cos t \\ t\sin t \\ -\cos t \\ \sin t \end{bmatrix}.$$

Therefore,

$$\boldsymbol{\alpha}_{1,1}(t) = \begin{bmatrix} t\sin t \\ t\cos t \\ \sin t \\ \cos t \end{bmatrix}, \quad \boldsymbol{\beta}_{1,1}(t) = \begin{bmatrix} -t\cos t \\ t\sin t \\ -\cos t \\ \sin t \end{bmatrix}.$$

For $z_{2,1}$ we have

$$\begin{aligned} \zeta_{1,2}(t) &= \mathrm{e}^{\mathrm{i}t}(\boldsymbol{u}_2 + t(\boldsymbol{A} - \mathrm{i}\boldsymbol{I}_4)\boldsymbol{u}_2) = \mathrm{e}^{\mathrm{i}t}\begin{bmatrix} -\mathrm{i} \\ 1 \\ 0 \\ 0 \end{bmatrix} \\ &= (\cos t + \mathrm{i}\sin t)\left(\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \mathrm{i}\begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} \sin t \\ \cos t \\ 0 \\ 0 \end{bmatrix} + \mathrm{i}\begin{bmatrix} -\cos t \\ \sin t \\ 0 \\ 0 \end{bmatrix}, \end{aligned}$$

and so we have

$$\boldsymbol{\alpha}_{1,2}(t) = \begin{bmatrix} \sin t \\ \cos t \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\beta}_{1,2}(t) = \begin{bmatrix} -\cos t \\ \sin t \\ 0 \\ 0 \end{bmatrix}.$$

Thus we have the four real linearly independent solutions $\boldsymbol{\alpha}_{1,1}$, $\boldsymbol{\alpha}_{1,2}$, $\boldsymbol{\beta}_{1,1}$, and $\boldsymbol{\beta}_{1,2}$. The corresponding fundamental matrix is

$$\Xi(t) = \begin{bmatrix} t\sin t & -t\cos t & \sin t & -\cos t \\ t\cos t & t\sin t & \cos t & \sin t \\ \sin t & -\cos t & 0 & 0 \\ \cos t & \sin t & 0 & 0 \end{bmatrix}.$$

A little manipulation gives

$$\mathrm{e}^{At} = \Xi(t)\Xi(0)^{-1} = \begin{bmatrix} \cos t & \sin t & t\cos t & t\sin t \\ -\sin t & \cos t & -t\sin t & t\cos t \\ 0 & 0 & \cos t & \sin t \\ 0 & 0 & -\sin t & \cos t \end{bmatrix}. \qquad \bullet$$

Exercises

5.2.1 Let X be an n-dimensional $\mathbb{R}$-vector space and let F be a system of linear ordinary differential equations in X with right-hand side

$$\widehat{F}(t, x) = \mathsf{A}(t)(x) + b(t)$$

for $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ and $b\colon \mathbb{T} \to \mathsf{X}$. Let $\{e_1, \dots, e_n\}$ be a basis for X and write

$$b(t) = \sum_{j=1}^{n} b_j(t)e_j, \quad \mathsf{A}(t)(e_j) = \sum_{k=1}^{n} A_j^k(t)e_k, \qquad j \in \{1, \dots, n\},$$

for functions $b_j\colon \mathbb{T} \to \mathbb{R}$, $j \in \{1, \dots, n\}$, and $A_j^k\colon \mathbb{T} \to \mathbb{R}$, $j, k \in \{1, \dots, n\}$. This defines $\boldsymbol{b}\colon \mathbb{T} \to \mathbb{R}^n$ and $\boldsymbol{A}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$. Denote by $\boldsymbol{F}$ the system of linear ordinary differential equations in $\mathbb{R}^n$ given by

$$\boldsymbol{F}(t, \boldsymbol{x}, \boldsymbol{x}^{(1)}) = \boldsymbol{x}^{(1)} - \boldsymbol{A}(t)\boldsymbol{x} - \boldsymbol{b}(t).$$

Answer the following questions.

(a) Show that $\xi\colon \mathbb{T}' \to \mathsf{X}$ is a solution for F if and only if the function $\boldsymbol{\xi}\colon \mathbb{T}' \to \mathbb{R}^n$, defined by

$$\boldsymbol{\xi}(t) = \sum_{j=1}^{n} \xi_j(t) e_j,$$

is a solution for $\boldsymbol{F}$.

Now let $\{\tilde{e}_1, \dots, \tilde{e}_n\}$ be another basis for X and let $\boldsymbol{P}$ be the change of basis matrix defined by

$$\tilde{e}_j = \sum_{k=1}^{n} P_{kj} e_k, \qquad j \in \{1, \dots, n\}.$$

Define $\tilde{\boldsymbol{b}}\colon \mathbb{T} \to \mathbb{R}^n$, $\tilde{\boldsymbol{A}}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$, and $\tilde{\boldsymbol{F}}$ as above, for this new basis.

(b) Show that $\tilde{\boldsymbol{b}}(t) = \boldsymbol{P}\boldsymbol{b}(t)$ and $\tilde{\boldsymbol{A}}(t) = \boldsymbol{P}^{-1}\boldsymbol{A}(t)\boldsymbol{P}$ for every $t \in \mathbb{T}$.
Hint: *Use the change of basis formulae from Proposition I-5.4.26 and from Theorem I-5.4.32.*

(c) Show that, if $\boldsymbol{\xi}\colon \mathbb{T}' \to \mathbb{R}^n$ is a solution for $\boldsymbol{F}$, then $\tilde{\boldsymbol{\xi}}\colon \mathbb{T}' \to \mathbb{R}^n$ is a solution for $\tilde{\boldsymbol{F}}$ if and only if $\tilde{\boldsymbol{\xi}}(t) = \boldsymbol{P}^{-1}\boldsymbol{\xi}(t)$ for every $t \in \mathbb{T}$.

5.2.2 Let X be an n-dimensional $\mathbb{R}$-vector space and let F be a system of linear homogeneous ordinary differential equations with right-hand side

$$\widehat{F}(t, x) = \mathsf{A}(t)(x)$$

for $\mathsf{A} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$. Let $\{e_1, \dots, e_n\}$ be a basis for X and let $\boldsymbol{A}(t)$ be the matrix representative for $\mathsf{A}(t)$, $t \in \mathbb{T}$, and let $\boldsymbol{F}$ be the corresponding system of linear homogeneous ordinary differential equations in $\mathbb{R}^n$ with right-hand side

$$\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) = \boldsymbol{A}(t)\boldsymbol{x}.$$

cf. Exercise 5.2.1.

(a) Show that, for every $t, t_0 \in \mathbb{T}$, the matrix representative of $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)$ is $\Phi^{\mathrm{c}}_{\boldsymbol{A}}(t, t_0)$.

Now let $\{\tilde{e}_1, \dots, \tilde{e}_n\}$ be another basis for X and let $\boldsymbol{P}$ be the change of basis matrix defined by

$$\tilde{e}_j = \sum_{k=1}^{n} P_{kj} e_k, \qquad j \in \{1, \dots, n\}.$$

Define $\tilde{\boldsymbol{A}}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $\tilde{\boldsymbol{F}}$ as above, for this new basis.

(b) Show that, for every $t, t_0 \in \mathbb{T}$,

$$\Phi^{\mathrm{c}}_{\tilde{A}}(t, t_0) = \boldsymbol{P}^{-1}\Phi^{\mathrm{c}}_{A}(t, t_0)\boldsymbol{P}.$$

5.2.3 Consider the system of linear homogeneous ordinary differential equations F with right-hand side equation (5.3) and suppose that $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$. Recall from the proof of Theorem 5.2.3 the maps

$$\begin{aligned} \sigma_t \colon \mathrm{Sol}(F) &\to \mathsf{X} \\ \xi &\mapsto \xi(t) \end{aligned}, \qquad t \in \mathbb{T},$$

that were shown to be isomorphisms.

(a) Show that

$$\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0) = \sigma_t \circ \sigma_{t_0}^{-1}$$

for each $t, t_0 \in \mathbb{T}$.

(b) Use this to give alternative proofs of parts (iv) and (v) of Theorem 5.2.6.

5.2.4 Consider a scalar linear homogeneous ordinary differential equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_0(t)x - a_1(t)x^{(1)} - \cdots - a_{k-1}(t)x^{(k-1)},$$

for $a_0, a_1, \dots, a_{k-1} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$.

(a) Following Exercise 3.1.23, convert this kth order scalar system into a first order system $\boldsymbol{F}_1$ of linear homogeneous ordinary differential equations in $\mathbb{R}^k$, i.e., find the matrix function $\boldsymbol{A} \colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^k; \mathbb{R}^k)$ in this case.

(b) For a solution $t \mapsto \xi(t)$ for F, what is the corresponding solution $t \mapsto \boldsymbol{\xi}(t)$ for $\boldsymbol{F}_1$?

(c) Show that, given a fundamental set of solutions $\{\xi_1, \dots, \xi_k\}$ for F, the solutions $\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_k\}$ for $\boldsymbol{F}_1$ from part (b) are a fundamental set of solutions for $\boldsymbol{F}_1$.

(d) Show that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{\mathrm{c}}_{A}(t, t_0) = \frac{\mathrm{W}(\xi_1, \dots, \xi_n)(t)}{\mathrm{W}(\xi_1, \dots, \xi_n)(t_0)}.$$

(e) Show that

$$\mathrm{W}(\xi_1, \dots, \xi_k)(t) = \mathrm{W}(\xi_1, \dots, \xi_k)(t_0)\mathrm{e}^{-\int_{t_0}^{t} a_{k-1}(\tau)\,\mathrm{d}\tau}.$$

5.2.5 Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let X be an n-dimensional $\mathbb{F}$-vector space, and let $\mathsf{L} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$. Let $\mathscr{B} = \{e_1, \dots, e_n\}$ and $\mathscr{B}' = \{\tilde{e}_1, \dots, \tilde{e}_n\}$ be bases for X and let $\boldsymbol{P}$ be the change of basis matrix defined by

$$\tilde{e}_j = \sum_{k=1}^{n} P_{kj} e_k, \qquad j \in \{1, \dots, n\}.$$

Let $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$ be the matrix representatives for L in the $\mathscr{B}$ and $\tilde{\mathscr{B}}$, respectively.

(a) Use part (b) of Exercise 5.2.2 to show that

$$[\mathrm{e}^{\mathsf{L}}]^{\tilde{\mathscr{B}}}_{\tilde{\mathscr{B}}} = \boldsymbol{P}^{-1}[\mathrm{e}^{\mathsf{L}}]^{\mathscr{B}}_{\mathscr{B}}\boldsymbol{P}.$$

(b) Use Theorem 5.2.20(i) and Proposition 5.2.22 to arrive at the same conclusion.

5.2.6 Consider the first-order scalar linear homogeneous ordinary differential equation with right-hand side $\widehat{F}(t,x) = a(t)x$ for $a \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathbb{R})$. Determine the continuous-time state-transition map in this case, thinking of this as a system of linear homogeneous ordinary differential equations in the one-dimensional vector space $\mathbb{R}$.

5.2.7 Let $\lambda \in \mathbb{F} \in \{\mathbb{R},\mathbb{C}\}$ and consider the linear map $\boldsymbol{A} \in \mathrm{L}(\mathbb{F}^n;\mathbb{F}^n)$ determined by the $n \times n$-matrix

$$\boldsymbol{A} = \left[\begin{array}{ccccc|ccccc} \lambda & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \lambda & 0 & 0 & 0 & \cdots & 0 & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 & \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \lambda \end{array}\right].$$

We suppose the lower right block is a $k \times k$-matrix and the upper left block, therefore, is a $(n-k) \times (n-k)$-matrix.

Answer the following questions.

(a) What are the eigenvalues of $\boldsymbol{A}$?

(b) For each of the eigenvalues of $\boldsymbol{A}$, determine its algebraic multiplicity.

(c) For each of the eigenvalues of $\boldsymbol{A}$, determine its eigenspace.

(d) For each of the eigenvalues of $\boldsymbol{A}$, determine its geometric multiplicity.

(e) For each of the eigenvalues of $\boldsymbol{A}$, determine its generalised eigenspace.

(f) For each of the eigenvalues ℓ of $\boldsymbol{A}$, determine the smallest $m \in \mathbb{Z}_{>0}$ for which $\overline{\mathsf{W}}(\ell,\boldsymbol{A}) = \ker((\boldsymbol{A}-\ell\boldsymbol{I}_n)^m)$.

5.2.8 For each of the following linear maps $\boldsymbol{A} \in \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$, given by an $n \times n$-matrix, determine the

1. eigenvalues,
2. eigenvectors,
3. generalised eigenvectors,

4. algebraic multiplicities of each eigenvalue, and
5. geometric multiplicities of each eigenvalue.

Here are the linear maps:

(a) $A = \begin{bmatrix} 2 & -5 \\ 0 & 3 \end{bmatrix}$;

(b) $A = \begin{bmatrix} -1 & -2 \\ 1 & -3 \end{bmatrix}$;

(c) $A = \begin{bmatrix} 4 & -1 \\ 4 & 0 \end{bmatrix}$;

(d) $A = \begin{bmatrix} 5 & 0 & -6 \\ 0 & 2 & 0 \\ 3 & 0 & -4 \end{bmatrix}$;

(e) $A = \begin{bmatrix} 5 & 0 & -6 \\ 1 & 2 & -1 \\ 3 & 0 & -4 \end{bmatrix}$;

(f) $A = \begin{bmatrix} 4 & 2 & -4 \\ 2 & 0 & -4 \\ 2 & 2 & -2 \end{bmatrix}$;

(g) $A = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 3 & -1 & 3 \\ 0 & 1 & 2 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$;

(h) $A = \begin{bmatrix} -7 & 0 & 0 & -4 \\ -13 & -2 & -1 & -8 \\ 6 & 1 & 0 & 4 \\ 15 & 1 & 0 & 9 \end{bmatrix}$;

(i) $A = \begin{bmatrix} 1 & 4 & -2 & 0 & 9 \\ 0 & -2 & 1 & 2 & -6 \\ -2 & 4 & -1 & 3 & 0 \\ -9 & 4 & 1 & 0 & 2 \\ 4 & 0 & 3 & -1 & 3 \end{bmatrix}$.

5.2.9 For each of the following $\mathbb{R}^n$-valued functions ξ of time, indicate whether they can be the solution of a system of linear homogeneous ordinary differential equations with constant coefficients. If they can be, find a matrix A for which the function satisfies $\dot{\xi}(t) = A\xi(t)$. If they cannot be, explain why not.

(a) $\xi(t) = (e^t, e^{-t})$;
(b) $\xi(t) = (\cos(t) - \sin(t), \cos(t) + \sin(t))$;
(c) $\xi(t) = (e^t + e^{2t}, 0, 0)$;
(d) $\xi(t) = (t, 0, 1)$;
(e) $\xi(t) = (e^t, e^t + e^{2t}, 0)$;
(f) $\xi(t) = (\cos(t) + \sin(t), \cos(t) + \sin(t))$.

5.2.10 Let F be a scalar linear homogeneous ordinary differential equation with right-hand side

$$\widehat{F}(t, x, x^{(1)}, \dots, x^{(k-1)}) = -a_{k-1}x^{(k-1)} - \cdots - a_1 x^{(1)} - a_0 x,$$

for $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$.

(a) Following Exercise 3.1.23, convert F into a first-order system of linear homogeneous ordinary differential equations $\boldsymbol{F}_1$ in $\mathbb{R}^k$ and with right-hand side

$$\widehat{\boldsymbol{F}}_1(t, \boldsymbol{x}) = A\boldsymbol{x},$$

explicitly identifying $A \in \mathrm{L}(\mathbb{R}^k; \mathbb{R}^k)$.

(b) Show that the characteristic polynomial P_F of F is the same as the characteristic polynomial P_A of $\boldsymbol{A}$.

5.2.11 Determine $\mathrm{e}^{\boldsymbol{A}t}$ for the linear transformations $\boldsymbol{A} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ from Exercise 5.2.8.

5.2.12 For the linear transformations $\boldsymbol{A} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ of Exercise 5.2.11, determine the solution to the initial value problem

$$\dot{\xi}(t) = \boldsymbol{A}\xi(t), \quad \xi(0) = x_0,$$

with x_0 as follows:

(a) $x_0 = (0, 1)$;
(b) $x_0 = (2, -3)$;
(c) $x_0 = (1, 1)$;
(d) $x_0 = (-3, -1, 0)$;
(e) $x_0 = (1, 0, 1)$;
(f) $x_0 = (4, 1, 2)$;
(g) $x_0 = (1, -1, 0, 1)$;
(h) $x_0 = (-1, -1, 3, -2)$;
(i) $x_0 = (0, 0, 0, 0, 0)$.

5.2.13 For the scalar linear homogeneous ordinary differential equations of Exercise 4.2.10, do the following:

(a) convert these to a first-order system of linear homogeneous ordinary differential equations, explicitly identifying A;
(b) using the fundamental solutions obtained during the solution of the problems from Exercise 4.2.10, compute $\mathrm{e}^{\mathsf{A}t}$;
(c) solve the initial value problems from Exercise 4.2.10 using the operator exponential.

5.2.14 Let $\ell \in \mathbb{R}$ and $k \in \mathbb{Z}_{>0}$. Consider the Jordan block

$$\boldsymbol{J}(\ell, k) = \begin{bmatrix} \ell & 1 & 0 & \cdots & 0 & 0 \\ 0 & \ell & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \ell & 1 \\ 0 & 0 & 0 & \cdots & 0 & \ell \end{bmatrix}.$$

Do the following.

(a) Solve the initial value problems

$$\dot{\xi}_j(t) = \boldsymbol{J}(\ell, k)\xi_j(t), \quad \xi_j(0) = e_j, \qquad \mathrm{IVP}_j$$

$j \in \{1, \ldots, k\}$, recursively, first by solving IVP_k, then by solving IVP_{k-1}, and so on. At each stage you should be solving a scalar linear, possibly inhomogeneous, ordinary differential equation, and so the methods of Sections 4.2.2 and 4.3.2 can be used.

(b) Use your calculations to determine $\mathrm{e}^{\boldsymbol{J}(\ell,k)t}$.

Alternatively, compute $\mathrm{e}^{\boldsymbol{J}(\ell,k)t}$ as follows.

(c) What are the eigenvalues of $\boldsymbol{J}(\ell, k)$?

(d) What are the geometric and algebraic multiplicities of the eigenvalues?

(e) Compute

$$(\boldsymbol{J}(\ell, k) - \ell \boldsymbol{I}_n)^j, \qquad j \in \{0, 1, \ldots, k-1\},$$

probably using mathematical induction on j.

(f) Use your answers from the preceding three questions to explicitly compute $\mathrm{e}^{\boldsymbol{J}(\ell,k)t}$ using Procedure 5.2.23.

Section 5.3

Systems of linear inhomogeneous ordinary differential equations

In this section we extend our discussion of homogeneous equations in Section 5.2 to inhomogeneous equations. Thus we are talking about systems of linear ordinary differential equations F in a finite-dimensional $\mathbb{R}$-vector space X with right-hand side

$$\begin{aligned} \widehat{F}\colon \mathbb{T} \times \mathsf{X} &\to \mathsf{X} \\ (t, x) &\mapsto \mathsf{A}(t)(x) + b(t) \end{aligned} \tag{5.11}$$

for maps $b\colon \mathbb{T} \to \mathsf{X}$ and $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$. In our treatment of scalar equations in Section 5.3, we gave no fewer than three methods for working with inhomogeneous equations, two general methods (using Wronskians in Section 4.3.1.2 and the theory of continuous-time Green's function in Section 4.3.1.3) and one method that only works for inhomogeneous terms that are pretty uninteresting (the "method of undetermined coefficients in Section 4.3.2.1). We shall not be so expansive for systems of linear inhomogeneous equations, and shall really only consider "the" method for working with such equations, since this method is as tractable as any other method in practice (which is to say, not very tractable at all, barring the use of a computer algebra package), and is exceptionally powerful in developing the theory of systems of linear ordinary differential equations.

As we have done in all preceding developments of linear ordinary differential equations, we work first in the general time-varying case, and then in the case of constant coefficients.

Do I need to read this section? The material in this section is fundamental to the theory of linear systems. •

5.3.1 Equations with time-varying coefficients

We state the, by now, more or less obvious results concerning existence and uniqueness, now for systems of linear inhomogeneous ordinary differential equations.

5.3.1 Proposition (Local existence and uniqueness of solutions for systems of linear inhomogeneous ordinary differential equations) *Consider the system of linear inhomogeneous ordinary differential equations* F *with right-hand side (5.11) and suppose that* $\mathsf{b} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{X})$ *and* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$. *Let* $(\mathrm{t}_0, \mathrm{x}_0) \in \mathbb{T} \times \mathsf{X}$. *Then there exists an interval* $\mathbb{T}' \subseteq \mathbb{T}$ *and a map* $\xi \in \mathsf{AC}_{\mathrm{loc}}(\mathbb{T}'; \mathsf{X})$ *that is a solution for* F *and which satisfies* $\xi(\mathrm{t}_0) = \mathrm{x}_0$. *Moreover, if* $\tilde{\mathbb{T}}' \subseteq \mathbb{T}$ *is another subinterval and if* $\tilde{\xi} \in \mathsf{AC}_{\mathrm{loc}}(\tilde{\mathbb{T}}'; \mathsf{X})$ *is another solution for* F *satisfying* $\tilde{\xi}(\mathrm{t}_0) = \mathrm{x}_0$, *then* $\tilde{\xi}(\mathrm{t}) = \xi(\mathrm{t})$ *for every* $\mathrm{t} \in \tilde{\mathbb{T}}' \cap \mathbb{T}'$.

Proof By Proposition 5.2.1, there exists a compact interval $\mathbb{T}' \subseteq \mathbb{T}$ and a solution $\xi_h \colon \mathbb{T} \to X$ for F_h satisfying $\xi_h(t_0) = x_0$. Moreover, $\xi_h(t) = \Phi^c_A(t, t_0)(x_0)$. Now define

$$\begin{aligned} \xi \colon \mathbb{T}' &\to X \\ t &\mapsto \Phi^c_A(t, t_0)(x_0) + \int_{t_0}^{t} \Phi^c_A(t, \tau)(b(\tau))\, d\tau. \end{aligned}$$

Note that the integral defining ξ exists since both $\tau \mapsto \Phi^c_A(t, \tau)$ is continuous (and so bounded on $[t_0, t]$) and since $\tau \mapsto b(\tau)$ is locally integrable, the first holding for every $t \in \mathbb{T}'$. In order to verify that ξ so defined is a solution for F, we will use the following lemma.

1 Lemma *Let $\mathbb{T} \subseteq \mathbb{R}$ be a compact interval and let $\mathbf{g} \colon \mathbb{T} \times \mathbb{T} \to \mathbb{R}^n$ have the following properties:*

(i) for almost every $t \in \mathbb{T}$, the map $\tau \mapsto \mathbf{g}(t, \tau)$ is locally integrable;

(ii) for almost every $\tau \in \mathbb{T}$, the map $t \mapsto \mathbf{g}(t, \tau)$ is absolutely continuous;

(iii) the mapping $(t, \tau) \mapsto \mathbf{g}(t, \tau)$ is locally integrable;

(iv) the mapping $(t, \tau) \mapsto \left\| \frac{\partial \mathbf{g}}{\partial t}(t, \tau) \right\|$ is locally integrable.

Then, for any $t_0 \in \mathbb{T}$, the function

$$\begin{aligned} \mathbf{G} \colon \mathbb{T} &\to \mathbb{R}^n \\ t &\mapsto \int_{t_0}^{t} \mathbf{g}(t, \tau)\, d\tau \end{aligned}$$

is locally absolutely continuous and

$$\frac{d\mathbf{G}}{dt}(t) = \int_{t_0}^{t} \frac{\partial \mathbf{g}}{\partial t}(t, \tau)\, d\tau + \mathbf{g}(t, t).$$

Proof Local integrability of $\tau \mapsto g(t, \tau)$ ensures that the integral in the definition of G exists. Consider the function

$$\begin{aligned} \tilde{G} \colon \mathbb{T} \times \mathbb{T} &\to \mathbb{R}^n \\ (t_1, t_2) &\mapsto \int_{t_0}^{t_1} g(t_2, \tau)\, d\tau. \end{aligned}$$

By the Fundamental Theorem of Calculus in the form of Theorem III-2.9.33, $t_1 \mapsto \tilde{G}(t_1, t_2)$ is locally absolutely continuous for each $t_2 \in \mathbb{T}$ and

$$\frac{\partial \tilde{G}}{\partial t_1}(t_1, t_2) = g(t_2, t_1), \qquad \text{a.e. } t_1 \in \mathbb{T},\ t_2 \in \mathbb{T}.$$

Our hypotheses ensure that one can use Theorem III-2.9.17 to assert that, for almost every fixed t_1, $t_2 \mapsto \tilde{G}(t_1, t_2)$ is locally absolutely continuous and we can differentiate $\tilde{G}$ with respect to t_2 inside the integral:

$$\frac{\partial \tilde{G}}{\partial t_2}(t_1, t_2) = \int_{t_0}^{t_1} \frac{\partial g}{\partial t_2}(t_2, \tau)\, d\tau.$$

Now define

$$\begin{aligned}\delta\colon \mathbb{T} &\to \mathbb{T}\times\mathbb{T}\\ t &\mapsto (t,t).\end{aligned}$$

Clearly δ is differentiable and

$$G(t) = \tilde{G}\circ\delta(t).$$

Thus G is locally absolutely continuous by Exercise II-1.10.6. Using the Chain Rule,

$$\begin{aligned}\frac{\mathrm{d}G}{\mathrm{d}t}(t) &= \frac{\partial\tilde{G}}{\partial t_1}(\delta(t))\circ\frac{\mathrm{d}\delta_1}{\mathrm{d}t}(t) + \frac{\partial\tilde{G}}{\partial t_t}(\delta(t))\circ\frac{\mathrm{d}\delta_2}{\mathrm{d}t}(t)\\ &= g(t,t) + \int_{t_0}^{t}\frac{\partial g}{\partial t}(t,\tau)\,\mathrm{d}\tau,\end{aligned}$$

as claimed. ▼

Let us verify that the hypotheses of the lemma hold for $(t,\tau)\mapsto\Phi^{\mathrm{c}}_{\mathsf{A}}(t,\tau)(b(\tau))$. First of all, we certainly have the first two hypotheses of the lemma. Moreover, writing $\Phi^{\mathrm{c}}_{\mathsf{A}}(t,\tau)(b(\tau)) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(\tau,t_0)^{-1}(b(\tau))$ and noting that (1) $\tau\mapsto b(\tau)$ is locally integrable (and so integrable on the compact interval $\mathbb{T}'$), (2) $t\mapsto\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)$ is continuous (and so also bounded on the compact interval $\mathbb{T}'$), and (3) $\tau\mapsto\Phi^{\mathrm{c}}_{\mathsf{A}}(\tau,t_0)^{-1}$ is also continuous (and so also locally bounded), we conclude that the third hypothesis of the lemma holds. Finally, using Theorem 5.2.6(i), we have $\frac{\partial\Phi^{\mathrm{c}}_{\mathsf{A}}}{\partial t}(t,\tau) = \mathsf{A}(t)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t,\tau)$. Now local integrability of b and A and the continuity of $\Phi^{\mathrm{c}}_{\mathsf{A}}$ ensure the fourth of the hypotheses of the lemma. Thus we can use the lemma to calculate

$$\begin{aligned}\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) &= \mathsf{A}(t)\circ\Phi^{\mathrm{c}}_{\mathsf{A}}(t,t_0)(x_0) + \mathsf{A}(t)\circ\int_{t_0}^{t}\Phi^{\mathrm{c}}_{\mathsf{A}}(t,\tau)(b(\tau))\,\mathrm{d}\tau + b(t)\\ &= \mathsf{A}(t)(\xi(t)) + b(t),\end{aligned}$$

i.e., ξ is a solution of F. Moreover, we also clearly have $\xi(t_0) = x_0$.

To conclude uniqueness, suppose that we have two solutions ξ_1 and ξ_2 defined on the same interval $\mathbb{T}'$. Then

$$\frac{\mathrm{d}\xi_1}{\mathrm{d}t}(t) = \mathsf{A}(t)(\xi_1(t)) + b(t),\quad \frac{\mathrm{d}\xi_2}{\mathrm{d}t}(t) = \mathsf{A}(t)(\xi_2(t)) + b(t),$$

and $\xi_1(t_0) = \xi_2(t_0) = x_0$. Therefore,

$$\frac{\mathrm{d}(\xi_1-\xi_2)}{\mathrm{d}t}(t) = \mathsf{A}(t)(\xi_1(t)-\xi_2(t)),\quad (\xi_1-\xi_2)(t_0) = 0.$$

By the uniqueness assertion of Proposition 5.2.1, we conclude that $\xi_1-\xi_2 = 0$, i.e., $\xi_1 = \xi_2$. ■

We also have a global existence result in this case, just as for homogeneous systems.

5.3.2 Proposition (Global existence of solutions for systems of linear inhomogeneous ordinary differential equations) *Consider the system of linear inhomogeneous ordinary differential equations* F *with right-hand side (5.11) and suppose that* $\mathsf{b} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{X})$ *and* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{X}; \mathsf{X}))$. *If* $\xi \colon \mathbb{T}' \to \mathsf{X}$ *is a solution for* F, *then there exists a solution* $\overline{\xi} \colon \mathbb{T} \to \mathsf{X}$ *for which* $\overline{\xi}|\mathbb{T}' = \xi$.

Proof In the proof of Proposition 5.3.1, we showed that a unique solution exists on any compact interval containing t_0. Just as in the proof of Proposition 5.2.2, this implies that a solution exists at any $t \in \mathbb{T}$. ■

Since, in the proof of Proposition 5.3.1, we gave an explicit formula for solutions to initial value problems, it is worth extracting this explicit formula.

5.3.3 Corollary (An explicit solution for systems of linear inhomogeneous ordinary differential equations) *Consider the system of linear inhomogeneous ordinary differential equations* F *with right-hand side (5.11) and suppose that* $\mathsf{b} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{X})$ *and* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{X}; \mathsf{X}))$. *Given* $\mathrm{t}_0 \in \mathbb{T}$ *and* $\mathrm{x}_0 \in \mathsf{X}$, *the unique solution* $\xi \colon \mathbb{T} \to \mathsf{X}$ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \mathsf{A}(\mathrm{t})(\xi(\mathrm{t})) + \mathsf{b}(\mathrm{t}), \quad \xi(\mathrm{t}_0) = \mathrm{x}_0,$$

is

$$\xi(\mathrm{t}) = \Phi^{\mathrm{c}}_{\mathsf{A}}(\mathrm{t}, \mathrm{t}_0)(\mathrm{x}_0) + \int_{\mathrm{t}_0}^{\mathrm{t}} \Phi^{\mathrm{c}}_{\mathsf{A}}(\mathrm{t}, \tau)(\mathsf{b}(\tau))\, \mathrm{d}\tau, \qquad \mathrm{t} \in \mathbb{T}. \tag{5.12}$$

The formula (5.12) for solutions to systems of linear inhomogeneous ordinary differential equations is often called the ***variation of constants formula***.

We note that this solution bears a strong resemblance in form to the continuous-time Green's function solution for scalar systems given in Theorem 4.3.7; indeed, one can think of the continuous-time state transition map as playing the rôle of a continuous-time Green's function in this case. In particular, given $b \in \mathsf{X}$ (a constant vector, note) the physical interpretation of Remark 4.3.9–2 applies to the map $t \mapsto \Phi^{\mathrm{c}}_{\mathsf{A}}(t, \tau)(b)$, and leads us to think of this as being the result of applying an impulse at time τ with (vector) magnitude b. This leads to the important notion in system theory of the impulse response.

Now we can discuss the set of all solutions of a system of linear inhomogeneous ordinary differential equation F with right-hand side

$$\begin{aligned} \widehat{F} \colon \mathbb{T} \times \mathsf{X} &\to \mathsf{X} \\ (t, x) &\mapsto \mathsf{A}(t)(x). \end{aligned}$$

To this end, we denote by

$$\mathrm{Sol}(F) = \left\{ \xi \in \mathsf{AC}_{\mathrm{loc}}(\mathbb{T}; \mathsf{X}) \;\middle|\; \dot{\xi}(t) = \mathsf{A}(t)(\xi(t)) \right\}$$

the set of solutions for F. While $\mathrm{Sol}(F)$ was a vector space in the homogeneous case, in the inhomogeneous case this is no longer the case. However, the set

of all solutions for the homogeneous case plays an important rôle, even in the inhomogeneous case. To organise this discussion, we let F_{h} be the "homogeneous part" of F. Thus the right-hand side of F_{h} is

$$\widehat{F_{\mathrm{h}}}(t, x) = \mathsf{A}(t)(x).$$

As in Theorem 4.3.3, $\mathrm{Sol}(F_{\mathrm{h}})$ is a $\mathbb{R}$-vector space of dimension $\dim_{\mathbb{R}}(\mathsf{X})$. The following result is then the main structural result about the set of solutions to a system of linear inhomogeneous ordinary differential equations, mirroring Theorem 4.3.3 for scalar systems.

5.3.4 Theorem (Affine space structure of sets of solutions) *Consider the system of linear inhomogeneous ordinary differential equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (4.11) and suppose that* $\mathsf{b} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{X})$ *and* $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{X}; \mathsf{X}))$. *Let* $\xi_{\mathrm{p}} \in \mathrm{Sol}(\mathrm{F})$. *Then*

$$\mathrm{Sol}(\mathrm{F}) = \{\xi + \xi_{\mathrm{p}} \mid \xi \in \mathrm{Sol}(\mathrm{F}_{\mathrm{h}})\}.$$

Proof First note that, by Theorem 5.2.3, $\mathrm{Sol}(F) \neq \varnothing$ and so there exists some $\xi_{\mathrm{p}} \in \mathrm{Sol}(F)$. We have, of course,

$$\frac{\mathrm{d}\xi_{\mathrm{p}}}{\mathrm{d}t}(t) = \mathsf{A}(t)(\xi_{\mathrm{p}}(t)) + b(t). \tag{5.13}$$

Next let $\xi \in \mathrm{Sol}(F)$ so that

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) = \mathsf{A}(t)(\xi(t)) + b(t). \tag{5.14}$$

Subtracting (5.13) from (5.14) we get

$$\frac{\mathrm{d}(\xi - \xi_{\mathrm{p}})}{\mathrm{d}t}(t) = \mathsf{A}(t)(\xi(t) - \xi_{\mathrm{p}}(t)),$$

i.e., $\xi - \xi_{\mathrm{p}} \in \mathrm{Sol}(F_{\mathrm{h}})$. In other words, $\xi = \tilde{\xi} + \xi_{\mathrm{p}}$ for $\tilde{\xi} \in \mathrm{Sol}(F_{\mathrm{h}})$.

Conversely, suppose that $\xi = \tilde{\xi} + \xi_{\mathrm{p}}$ for $\tilde{\xi} \in \mathrm{Sol}(F_{\mathrm{h}})$. Then

$$\frac{\mathrm{d}\tilde{\xi}}{\mathrm{d}t}(t) = \mathsf{A}(t)(\tilde{\xi}(t)). \tag{5.15}$$

Adding (5.13) and (5.15) we get

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) = \mathsf{A}(t)(\xi(t)) + b(t),$$

i.e., $\xi \in \mathrm{Sol}(F)$. ∎

As with scalar linear inhomogeneous ordinary differential equations, there is an insightful correspondence to be made between the situation described in Theorem 5.3.4 and that of systems of linear algebraic equations described in Proposition I-5.4.48.

5.3.5 Remark (Comparison of Theorem 5.3.4 with systems of linear algebraic equations) Let us compare here the result of Theorem 5.3.4 with the situation in Proposition I-5.4.48 concerning linear algebraic equations of the form $L(u) = v_0$, for vector spaces U and W, a linear map $L \in \mathrm{L}(\mathsf{U}; \mathsf{W})$, and a fixed $w_0 \in \mathsf{W}$. In the setting of systems of linear inhomogeneous ordinary differential equations in a $\mathbb{R}$-vector space X, we have

$$\begin{aligned} \mathsf{U} &= \mathsf{AC}_{\mathrm{loc}}(\mathbb{T}; \mathsf{X}), \\ \mathsf{W} &= \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{X}), \\ L(f)(t) &= \dot{f}(t) - \mathsf{A}(t)(f(t)), \\ w_0 &= b. \end{aligned}$$

Then Propositions 5.3.1 and 5.3.2 tells us that L is surjective, and so $w_0 \in \mathrm{image}(L)$. Thus we are in case (I-ii) of Proposition I-5.4.48, which exactly the statement of Theorem 5.3.4. Note that L is not injective, since Theorem 5.2.3 tells us that $\dim_{\mathbb{R}}(\ker(L)) = \dim_{\mathbb{R}}(\mathsf{X})$. •

5.3.6 Remark (What happened to the Wronskian?) In Section 4.3.1.2 we described how the Wronskian can be used for scalar linear inhomogeneous ordinary differential equations to generate a particular solution. A similar development is possible for systems of equations, but we shall not pursue it here. It is worth recording the reasons for not doing so.

1. In Corollary 5.3.3 we produce a specific and natural "particular solution" for a system of linear inhomogeneous ordinary differential equations, namely the function that assigns to the inhomogeneous term "b," the solution

$$\xi_{\mathrm{p}}(t) = \int_{t_0}^{t} \Phi^{\mathrm{c}}_{\mathsf{A}}(t, \tau)(b(\tau)) \, \mathrm{d}\tau.$$

   Then the form of the solution of Corollary 5.3.3 is $\xi = \xi_{\mathrm{h}} + \xi_{\mathrm{p}}$, where $\xi_{\mathrm{h}} \in \mathrm{Sol}(F_{\mathrm{h}})$ satisfies the initial conditions. This is just so cool. . . why would you want to do more?

2. In Section 4.2.1 we discussed the notion of a fundamental set of solutions for scalar linear homogeneous ordinary differential equations. There is no really distinguished fundamental set of solutions, and the Wronskian-related constructions were developed for an *arbitrary* fundamental set of solutions. This has its benefits in this setting, as the results are general in this respect.

   However, in Section 5.2.1.2 we saw that there was *one* object that naturally describes the solutions for a system of linear homogeneous ordinary differential equations, the continuous-time state transition map. Note that in Procedure 5.2.8 we indicate how to build the continuous-time state transition map from a fundamental set of solutions for a system of equations, through the

fundamental matrix-function Ξ that we build after choosing a basis. It is the fundamental matrix, and its determinant, that would be involved in Wronskian-type constructions for systems of equations. However, these are only arrived at after choosing a basis, and so seem quite unnatural in our setting of general vector spaces. •

Given that we will not be pursuing any Wronskian-type constructions, it only remains to illustrate how one might use the about constructions in practice.

5.3.7 Example (System of linear inhomogeneous ordinary differential equations) We take $\mathsf{X} = \mathbb{R}^2$ and the linear inhomogeneous ordinary differential equation $\boldsymbol{F}$ with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}\colon (0,\infty)\times\mathbb{R}^2 &\to \mathbb{R}^2\\ (t,(x_1,x_2)) &\mapsto \left(\frac{1}{t}x_1 - x_2 + t, \frac{1}{t^2}x_1 + \frac{2}{t}x_2 - t^2\right).\end{aligned}$$

A solution $t \mapsto (\xi_1(t), \xi_2(t))$ satisfies

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \underbrace{\begin{bmatrix}\frac{1}{t} & -1\\ \frac{1}{t^2} & \frac{2}{t}\end{bmatrix}}_{\boldsymbol{A}(t)}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} + \underbrace{\begin{bmatrix}t\\ -t^2\end{bmatrix}}_{\boldsymbol{b}(t)}.$$

Note that the homogeneous system $\boldsymbol{F}_{\mathrm{h}}$ was examined in Example 5.2.10, where we computed the continuous-time state transition matrix to be

$$\Phi^{\mathrm{c}}_{\boldsymbol{A}}(t,t_0) = \begin{bmatrix}-\frac{t^2(\ln(t/t_0)-1)}{t_0^2} & -\frac{t^2\ln(t/t_0)}{t_0}\\ \frac{t\ln(t/t_0)}{t_0^2} & \frac{t(\ln(t/t_0)+1)}{t_0}\end{bmatrix}.$$

We then compute[5]

$$\begin{aligned}\int_{t_0}^{t}\Phi^{\mathrm{c}}_{\boldsymbol{A}}(t,\tau)\boldsymbol{b}(\tau)\,\mathrm{d}\tau &= \int_{t_0}^{t}\begin{bmatrix}-\frac{t^2(\ln(t/\tau)-1)}{\tau^2} & -\frac{t^2\ln(t/\tau)}{\tau}\\ \frac{t\ln(t/\tau)}{\tau^2} & \frac{t(\ln(t/\tau)+1)}{\tau}\end{bmatrix}\begin{bmatrix}\tau\\ -\tau^2\end{bmatrix}\mathrm{d}\tau\\ &= \begin{bmatrix}\frac{1}{4}t^2(t^2 - 2t_0^2\ln(t/t_0) - 2\ln(t/t_0)^2 + 4\ln(t/t_0) - t_0^2)\\ \frac{1}{4}t(2\ln(t/t_0)(\ln(t/t_0) + t_0^2) - 3(t-t_0)(t+t_0))\end{bmatrix}.\end{aligned}$$

If we now wish to find the solution for $\boldsymbol{F}$ with initial condition $\boldsymbol{x}_0 = (x_{10}, x_{20})$ at time t_0, we use the explicit form of Corollary 5.3.3:

$$\begin{aligned}\xi(t) &= \Phi^{\mathrm{c}}_{\boldsymbol{A}}(t,t_0)\boldsymbol{x}_0 + \int_{t_0}^{t}\Phi^{\mathrm{c}}_{\boldsymbol{A}}(t,\tau)\boldsymbol{b}(\tau)\,\mathrm{d}\tau\\ &= \begin{bmatrix}-\frac{t^2(\ln(t/t_0)-1)}{t_0^2}x_{10} - \frac{t^2\ln(t/t_0)}{t_0}x_{20} + \frac{1}{4}t^2(t^2 - 2t_0^2\ln(t/t_0) - 2\ln(t/t_0)^2 + 4\ln(t/t_0) - t_0^2)\\ \frac{t\ln(t/t_0)}{t_0^2}x_{10} + \frac{t(\ln(t/t_0)+1)}{t_0}x_{20} + \frac{1}{4}t(2\ln(t/t_0)(\ln(t/t_0) + t_0^2) - 3(t-t_0)(t+t_0))\end{bmatrix}.\end{aligned}$$

[5]Integration courtesy of MATHEMATICA®.

As with pretty much any method for solving systems of linear inhomogeneous (or, indeed, homogeneous) ordinary differential equations, tedious computations and generally impossible integrals render the explicit formula of Corollary 5.3.3 of questionable value as a computational tool. •

5.3.2 Equations with constant coefficients

We now specialise the discussion in the preceding section to systems of linear inhomogeneous ordinary differential equations with constant coefficients. Thus we are looking at a system of linear inhomogeneous ordinary differential equations F in a finite-dimensional $\mathbb{R}$-vector space X and with right-hand side given by

$$\widehat{F}(t, x) = \mathsf{A}(x) + b(t) \tag{5.16}$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$ and $b \colon \mathbb{T} \to \mathsf{X}$. Of course, all general results concerning the existence and uniqueness of solutions (i.e., Propositions 5.3.1 and 5.3.2), and of the structure of the set of solutions (i.e., Theorem 5.3.4) apply in the constant coefficient case. Here, however, we can refine a little the explicit solution of Corollary 5.3.3 because, as per Theorem 5.2.20(ix), $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0) = \mathrm{e}^{\mathsf{A}(t-t_0)}$ in this case. We can thus summarise the situation in the following theorem.

5.3.8 Theorem (An explicit solution for systems of linear inhomogeneous ordinary differential equations with constant coefficients) *Consider the system of linear inhomogeneous ordinary differential equations* F *with constant coefficients and right-hand side (5.16), and suppose that* $\mathrm{b} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{X})$. *Given* $\mathrm{t}_0 \in \mathbb{T}$ *and* $\mathrm{x}_0 \in \mathsf{X}$, *the unique solution* $\xi \colon \mathbb{T} \to \mathsf{X}$ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \mathsf{A}(\xi(\mathrm{t})) + \mathrm{b}(\mathrm{t}), \quad \xi(\mathrm{t}_0) = \mathrm{x}_0,$$

is

$$\xi(\mathrm{t}) = \mathrm{e}^{\mathsf{A}(\mathrm{t}-\mathrm{t}_0)}(\mathrm{x}_0) + \int_{\mathrm{t}_0}^{\mathrm{t}} \mathrm{e}^{\mathsf{A}(\mathrm{t}-\tau)}(\mathrm{b}(\tau))\,\mathrm{d}\tau, \qquad \mathrm{t} \in \mathbb{T}.$$

We comment that our observations Remark 4.3.11 about the particular solution

$$\xi_{\mathrm{p},b} = \int_{t_0}^{t} \mathrm{e}^{\mathsf{A}(t-\tau)}(b(\tau))\,\mathrm{d}\tau$$

for constant coefficient systems and its relation to convolution integrals is also valid here.

5.3.9 Remark (What happened to the "method of undetermined coefficients"?) In Section 4.3.2.1 we spent some time describing a rather *ad hoc* method, the "method of undetermined coefficients," for finding particular solutions for scalar linear inhomogeneous ordinary differential equations with constant coefficients. A similar strategy is possible for systems of linear inhomogeneous ordinary differential equations with constant coefficients, but we shall not pursue it here. Here is why.

1. The rationale of Remark 5.3.6–1 is equally valid here: we have such a nice characterisation in Corollary 5.3.3 of a particular solution that to mess this up with an *ad hoc* procedure that only works for pretty uninteresting functions is simply not a worthwhile undertaking.
2. While for scalar equations it might be argued that there is some reason for being able to quickly bang out particular solutions for specific pretty uninteresting functions—see, particular, the notion of "step response" in Example 4.3.19 and the notion of "frequency response" in Example 4.3.20—for systems of equations the benefit of this is not so clear, given the complexity of doing computation in any example. •

All that remains, since we have discharged ourselves of the responsibility of providing any analogies to the various methods we used for scalar equations in Section 4.2, is to give an example of how to apply the explicit formula of Theorem 5.3.8.

5.3.10 Example (A second-order scalar equation as a system of equations) We consider here the second-order scalar linear inhomogeneous ordinary differential equation F with right-hand side

$$\widehat{F}(t, x, x^{(1)}) = -\omega_0^2 x - 2\zeta\omega_0 x^{(1)} + A\sin(\omega t)$$

that was considered in detail in Example 4.3.20. First we convert this into a system of linear inhomogeneous ordinary differential equations, following Exercise 3.1.23. Thus we introduce the variables $x_1 = x$ and $x_2 = x^{(1)}$ so that

$$\begin{aligned} x_1^{(1)} &= x^{(1)} = x_2, \\ x_2^{(1)} &= x^{(2)} = -\omega_0^2 x - 2\zeta\omega_0 x^{(1)} + A\sin(\omega t) = -\omega_0^2 x_1 - 2\zeta\omega_0 x_2 + A\sin(\omega t). \end{aligned}$$

That is to say

$$\widehat{\boldsymbol{F}}_1(t, (x_1, x_2)) = (x_2, -\omega_0^2 x_1 - 2\zeta\omega_0 x_2 + A\sin(\omega t)).$$

Solutions $t \mapsto (\xi_1(t), \xi_2(t))$ then satisfy

$$\begin{bmatrix} \dot{\xi}_1(t) \\ \dot{\xi}_2(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\zeta\omega_0 \end{bmatrix}}_{\boldsymbol{A}} \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ A\sin(\omega t) \end{bmatrix}}_{\boldsymbol{b}(t)}.$$

To illustrate, we suppose that $\zeta^2 \leq 1$ and $\omega_0 > 0$.

We will first compute $\mathrm{e}^{\boldsymbol{A}t}$ in this case, following Procedure 5.2.26, making use of the notation in Procedure 5.2.23. The characteristic polynomial of $\boldsymbol{A}$ is

$$P_{\boldsymbol{A}} = X^2 + 2\zeta\omega_0 X + \omega_0^2,$$

and so the eigenvalues of $\boldsymbol{A}$ are $\lambda_1 = \omega_0(-\zeta + \mathrm{i}\sqrt{1-\zeta^2})$, along with its complex conjugate $\bar{\lambda}_1$. This eigenvalue necessarily has algebraic and geometric multiplicity 1. We compute that

$$\ker(\boldsymbol{A}^{\mathbb{C}} - \lambda_1 \boldsymbol{I}_2) = \operatorname{span}_{\mathbb{R}}((-\zeta, \omega_0) + \mathrm{i}(\sqrt{1-\zeta^2}, \omega_0)).$$

Thus we take

$$\boldsymbol{\zeta}_{1,1} = (-\zeta, \omega_0) + \mathrm{i}(\sqrt{1-\zeta^2}, \omega_0)$$

and, therefore,

$$\boldsymbol{a}_{1,1} = (-\zeta, \omega_0), \quad \boldsymbol{b}_{1,1} = (-\sqrt{1-\zeta^2}, 0).$$

Thus

$$\boldsymbol{\alpha}_{1,1}(t) = \mathrm{e}^{-\omega_0\zeta t}\cos(\omega_0\sqrt{1-\zeta^2}t)\boldsymbol{a}_{1,1} - \mathrm{e}^{-\omega_0\zeta t}\sin(\omega_0\sqrt{1-\zeta^2}t)\boldsymbol{b}_{1,1}$$

and

$$\boldsymbol{\beta}_{1,1}(t) = \mathrm{e}^{-\omega_0\zeta t}\cos(\omega_0\sqrt{1-\zeta^2}t)\boldsymbol{b}_{1,1} + \mathrm{e}^{-\omega_0\zeta t}\sin(\omega_0\sqrt{1-\zeta^2}t)\boldsymbol{a}_{1,1}.$$

Thus a fundamental matrix is then determined to be

$$\boldsymbol{\Xi}(t) = \mathrm{e}^{-\omega_0\zeta t}\begin{bmatrix} -\zeta\cos(\omega_0\sqrt{1-\zeta^2}t) + \sqrt{1-\zeta^2}\sin(\omega_0\sqrt{1-\zeta^2}t) & \\ \omega_0\cos(\omega_0\sqrt{1-\zeta^2}t) & \\ & -\sqrt{1-\zeta^2}\cos(\omega_0\sqrt{1-\zeta^2}t) - \zeta\sin(\omega_0\sqrt{1-\zeta^2}t) \\ & \omega_0\sin(\omega_0\sqrt{1-\zeta^2}t) \end{bmatrix}.$$

Then we calculate

$$\begin{aligned}\mathrm{e}^{At} &= \boldsymbol{\Xi}(t)\boldsymbol{\Xi}(0)^{-1} \\ &= \mathrm{e}^{-\omega_0\zeta t}\begin{bmatrix} \cos(\omega_0\sqrt{1-\zeta^2}t) + \frac{\zeta\sin(\omega_0\sqrt{1-\zeta^2}t)}{\sqrt{1-\zeta^2}} & \frac{\sin(\omega_0\sqrt{1-\zeta^2}t)}{\omega_0\sqrt{1-\zeta^2}} \\ -\frac{\omega_0\sin(\omega_0\sqrt{1-\zeta^2}t)}{\sqrt{1-\zeta^2}} & \cos(\omega_0\sqrt{1-\zeta^2}t) - \frac{\zeta\sin(\omega_0\sqrt{1-\zeta^2}t)}{\sqrt{1-\zeta^2}} \end{bmatrix}.\end{aligned}$$

Now we can calculate[6]

$$\begin{aligned}\int_0^t \mathrm{e}^{A(t-\tau)}\boldsymbol{b}(\tau)\,\mathrm{d}\tau = \Bigg(&\mathrm{e}^{-\omega_0\zeta t}\frac{2A\omega\omega_0\zeta}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\cos(\omega_0\sqrt{1-\zeta^2}t)\\ &+\mathrm{e}^{-\omega_0\zeta t}\frac{A\omega(\omega^2+\omega_0^2(2\zeta^2-1))}{\sqrt{1-\zeta^2}\omega_0(\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4)}\sin(\omega_0\sqrt{1-\zeta^2}t)\\ &-\frac{2A\omega\omega_0\zeta}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\cos(\omega t)\\ &+\frac{A(\omega_0^2-\omega^2)}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\sin(\omega t),\\ &\mathrm{e}^{-\omega_0\zeta t}\frac{A\omega(\omega^2-\omega_0^2)}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\cos(\omega_0\sqrt{1-\zeta^2}t)\\ &-\mathrm{e}^{-\omega\zeta t}\frac{A\omega\zeta(\omega^2+\omega_0^2)}{\sqrt{1-\zeta^2}(\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4)}\sin(\omega_0\sqrt{1-\zeta^2}t)\\ &+\frac{A\omega(\omega_0^2-\omega^2)}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\cos(\omega t)\\ &+\frac{2A\zeta\omega^2\omega_0}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\sin(\omega t)\Bigg),\end{aligned}\tag{5.17}$$

assuming that $\zeta \neq 0$. If $\zeta = 0$ and $\omega \neq \omega_0$, the preceding expression is still valid. When $\zeta = 0$ and $\omega = \omega_0$, a different computation must be done, and in this case we compute

$$\int_0^t \mathrm{e}^{A(t-\tau)}\boldsymbol{b}(\tau)\,\mathrm{d}\tau = \left(\tfrac{A}{2\omega_0^2}(\sin(\omega_0 t)-\omega_0 t\cos(\omega_0 t)),\tfrac{A}{2}t\sin(\omega_0 t)\right).\tag{5.18}$$

Note that, in all cases, the preceding expressions give the solution to the ordinary differential equation when the initial conditions are $(0,0)$. Let us make some comments on this solution.

1. $\zeta \neq 0$: Note that (5.17) is *not* the steady-state response of the system, as was the particular solution obtained for this problem in Example 4.3.20 using the method of undetermined coefficients. The reason for the disparity is that the expression above has the property that its initial conditions at $t = 0$ are $(0,0)$.

[6]Integration courtesy of MATHEMATICA®.

Note that, as $t \to \infty$, we have

$$\int_0^t \mathrm{e}^{A(t-\tau)}\boldsymbol{b}(\tau)\,\mathrm{d}\tau \approx \Bigg(-\frac{2A\omega\omega_0\zeta}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\cos(\omega t) + \frac{A(\omega_0^2-\omega^2)}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\sin(\omega t), \frac{A\omega(\omega_0^2-\omega^2)}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\cos(\omega t) + \frac{2A\zeta\omega^2\omega_0}{\omega^4+2\omega^2\omega_0^2(2\zeta^2-1)+\omega_0^4}\sin(\omega t)\Bigg).$$

Notice that the first component of this is exactly the particular solution of Example 4.3.20, while the second component is its time-derivative. This is as it should be, given our conversion of the scalar second-order equation into a vector first-order equation.

2. $\zeta = 0$ and $\omega \neq \omega_0$: In this case, there is no steady-state solution since the homogeneous solution does not decay to zero as $t \to \infty$, and is indeed periodic itself. Nonetheless, the solution (5.17) does have two components, one with frequency ω and one with frequency ω_0. While this does not quite disambiguate the particular from the homogeneous solution[7], we can nonetheless see from the expression (5.17) that the particular solution of Example 4.3.20 is comprised on the last two terms in the first component.
3. $\zeta = 0$ and $\omega = \omega_0$: In this case, there is again no steady-state solution; indeed the solution "blows up" as $t \to \infty$. This is as we saw in Example 4.3.20, and is due to the physical phenomenon of "resonance." Moreover, the first component of (5.18) is *not* the particular solution from Example 4.3.20; the particular particular solution (5.18) is prescribed to have initial condition $(0, 0)$, whereas, in the method of undetermined coefficients, it is the *form* of the solution that is determined. •

5.3.3 Equations with distributions as right-hand side

In Section 4.4 we considered in some detail the situation of scalar linear ordinary differential equations with distributional forcing. We saw there that one must fuss a little bit with the manner in which things make sense, especially when dealing with equations with time-varying coefficients. While it is possible to duplicate this for systems of ordinary differential equations, to do so is mainly a matter of notation. To keep things simple and focussed, in this section we consider two aspects of systems of ordinary differential equations with distributional forcing: (1) we give a distributional interpretation of the continuous-time state transition map; (2) we consider in detail the situation for constant coefficient equations.

[7] A periodic function can have more than one frequency.

5.3.3.1 A distributional interpretation of the continuous-time state transition map An interesting connection can be made between the continuous-time state transition map of Theorem 5.2.6 and the inhomogeneous equation with an appropriate delta-function as right-hand side. The following theorem gives the desired result.

5.3.11 Theorem (The continuous-time state transition map as a solution to a distributional differential equation) *Let* F *be a system of linear homogeneous ordinary differential equation with* $\mathbb{T} = \mathbb{R}$*, right-hand side (5.3), and suppose that* $A \in L^1_{loc}(\mathbb{T}; L(X; X))$*. For* $s \in \mathbb{T}$*, let* $\Theta_s \in \mathscr{D}'(\mathbb{R}; L(X; X))$ *be the regular distribution corresponding to the locally integrable function* $t \mapsto \tau_s^* 1_{\geq 0} \Phi_A^c(t, s)$*. Then* Θ_s *is a solution to the distributional equation*

$$\Theta_s^{(1)} = A \circ \Theta_s + \mathrm{id}_X \otimes (\tau_s^* \delta),$$

where, by $A \circ \Theta_s$ *is as defined in Remark IV-3.2.53–3.*

Proof Let $\Xi_s \colon \mathbb{T} \to L(X; X)$ be the solution to the initial value problem

$$\dot{\Xi}_s(t) = A(t)(\Xi_s), \quad \Xi_s(s) = \mathrm{id}_X,$$

so that Θ_s is the distribution associated with the locally integrable function $\tau_s^* 1_{\geq 0} \Xi_s$. We then have

$$(\tau_s^* 1_{\geq 0} \Xi_s)^{(1)} = \tau_s^* 1_{\geq 0} \Xi_s^{(1)} + \tau_s^* \delta \otimes \Xi_s(s) = \tau_s^* 1_{\geq 0} A \circ \Xi_s + \mathrm{id}_X \otimes (\tau_s^* \delta).$$

Thus, replacing $\tau_s^* 1_{\geq 0} \Xi_s$ with the regular distribution Θ_s,

$$\Theta_s^{(1)} = A \circ \Theta_s + \mathrm{id}_X \otimes (\tau_s^* \delta),$$

as claimed. ∎

5.3.3.2 Equations with constant coefficients Next we turn to the consideration of systems of linear ordinary differential equations with distributional forcing. Thus we let X be a finite-dimensional $\mathbb{R}$-vector space, let $A \in L(X; X)$, and let $\beta \in \mathscr{D}'(\mathbb{R}; X)$. We seek solutions $\theta \in \mathscr{D}'(\mathbb{R}; X)$ to the equation

$$\theta^{(1)} = A(\theta) + \beta.$$

Since A is constant, the meaning of $A(\theta)$ is unambiguous, to wit

$$\langle A(\theta); \phi \rangle = A(\langle \theta; \phi \rangle), \qquad \phi \in \mathscr{D}(\mathbb{R}; \mathbb{R}).$$

Let us denote, in this constant coefficient case,

$$\mathrm{Sol}(F, \beta) = \{\theta \in \mathscr{D}(\mathbb{R}; X) \mid \theta^{(1)} = A(\theta) + \beta\}.$$

Given a system of linear homogeneous ordinary differential equations F with right-hand side (5.3), let us define

$$\begin{aligned} L_F \colon \mathscr{D}'(\mathbb{R}; X) &\to \mathscr{D}'(\mathbb{R}; X) \\ \theta &\mapsto \theta^{(1)} - A(\theta) \end{aligned}$$

and, accepting a mild abuse of notation that can be resolved by understanding context,

$$\begin{aligned} L_F \colon \mathscr{D}'(\mathbb{R}; \mathrm{L}(\mathsf{X}; \mathsf{X})) &\to \mathscr{D}'(\mathbb{R}; \mathrm{L}(\mathsf{X}; \mathsf{X})) \\ \Theta &\mapsto \Theta^{(1)} - \mathsf{A} \circ \Theta, \end{aligned}$$

where $\mathsf{A}(\Theta)$ is as defined in Remark IV-3.2.53–1. With this notation, we first have the following preparatory result.

5.3.12 Lemma (Convolution and systems of linear ordinary differential equations with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential equation with constant coefficients and with right-hand side (5.3). If* $\theta \in \mathscr{D}'(\mathbb{R}; \mathsf{X})$, *then*

$$\mathrm{L}_{\mathrm{F}}(\theta) = \mathrm{L}_{\mathrm{F}}(\mathrm{id}_{\mathsf{X}} \otimes \delta) * \theta.$$

Proof We compute need this identity about convolution with $\mathrm{id}_{\mathsf{X}} \otimes \delta$

$$\begin{aligned} L_F(\theta) &= L_F((\mathrm{id}_{\mathsf{X}} \otimes \delta) * \theta) = ((\mathrm{id}_{\mathsf{X}} \otimes \delta) * \theta)^{(1)} - \mathsf{A}((\mathrm{id}_{\mathsf{X}} \otimes \delta) * \theta) \\ &= (\mathrm{id}_{\mathsf{X}} \otimes \delta)^{(1)} * \theta - (\mathsf{A} \circ (\mathrm{id}_{\mathsf{X}} \otimes \delta)) * \theta) = L_F(\mathrm{id}_{\mathsf{X}} \otimes \delta) * \theta, \end{aligned}$$

using and . ■ convolution with delta

derivative of convolution

With the lemma at hand, we can easily prove the basic existence and uniqueness theorem for (F, β), where F has constant coefficients.

5.3.13 Theorem (Existence and uniqueness of solutions for constant coefficient systems of linear inhomogeneous ordinary differential equations with distribution forcing) *Let* F *be a systems of linear homogeneous ordinary differential equation with constant coefficients. Then the following statements hold:*

(i) *if* $\beta \in \mathscr{D}'(\mathbb{R}; \mathsf{X})$, *then* $\mathrm{card}(\mathrm{Sol}(\mathrm{F}, \beta)) \geq 2$;

(ii) *if* $\beta \in \mathscr{D}'_+(\mathbb{R}; \mathsf{X})$, *then* $\mathrm{card}(\mathrm{Sol}(\mathrm{F}, \beta) \cap \mathscr{D}'_+(\mathbb{R}; \mathbb{R})) = 1$;

(iii) *if* $\beta \in \mathscr{D}'_-(\mathbb{R}; \mathsf{X}))$, *then* $\mathrm{card}(\mathrm{Sol}(\mathrm{F}, \beta) \cap \mathscr{D}'_-(\mathbb{R}; \mathbb{R})) = 1$.

Proof Let us first establish the existence of two distribution solutions θ to the equation $L_F(\theta) = \mathrm{id}_{\mathsf{X}} \otimes \delta$. We suppose that

$$\widehat{F}(t, x) = -\mathsf{A}(x).$$

Let $\Xi_0 \in \mathsf{C}^\infty(\mathbb{R}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$ be the solution to the initial value problem

$$\dot{\Xi}_0(t) = \mathsf{A} \circ \Xi_0(t), \quad \Xi_0(0) = \mathrm{id}_{\mathsf{X}},$$

and denote $\Xi_+ = 1_{\geq 0}\Xi_0$ and $\Xi_- = -\sigma^* 1_{\geq 0}\Xi_0$. Note that $\Xi_0(t) = \mathrm{e}^{\mathsf{A}t}$ and that $\theta_{\Xi_+} \in \mathscr{D}'_+(\mathbb{R}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$ and $\theta_{\Xi_-} \in \mathscr{D}'_-(\mathbb{R}; \mathrm{L}(\mathsf{X}; \mathsf{X}))$.

From Theorem 5.3.11 we have $L_F(\theta_{\Xi_+}) = \mathrm{id}_{\mathsf{X}} \otimes \delta$. We claim that $L_F(\theta_{\Xi_-}) = \mathrm{id}_{\mathsf{X}} \otimes \delta$. Since $(\sigma^* 1_{\geq 0})^{(1)} = -\delta$, we have

$$(\sigma^* 1_{\geq 0}\Xi_0)^{(1)} = -\Xi_0(0) \otimes \delta + \sigma^* 1_{\geq 0}\Xi_0^{(1)}.$$

Thus

$$L_F(\theta_{\Xi_-}) = \mathrm{id}_X \otimes \delta - \sigma^* 1_{\geq 0} \Xi_0^{(1)} + A(\sigma^* 1_{\geq 0} \Xi_0) = \sigma^* 1_{\geq 0} L_F(\Xi_0) + \mathrm{id}_X \otimes \delta,$$

as claimed.

Note that this immediately gives

$$L_F(\mathrm{id}_X \otimes \delta) * \theta_{\Xi_+} = L_F((\mathrm{id}_X \otimes \delta) * \theta_{\Xi_+}) = L_F(\theta_{\Xi_+}) = \mathrm{id}_X \otimes \delta.$$

Similarly,

$$L_F(\mathrm{id}_X \otimes \delta) * \theta_{\Xi_-} = \mathrm{id}_X \otimes \delta,$$

showing that both θ_{Ξ_+} and θ_{Ξ_-} are multiplicative inverses of $L_F(\mathrm{id}_X \otimes \delta)$ in the ring $\mathscr{D}'(\mathbb{R}; L(X; X))$ with the convolution product.

need some background to this

Now we proceed with the proof of the theorem, using the notation just introduced.

(i) We have, using Lemma 5.3.12 and the computations above,

$$L_F(\theta_{\Xi_+} * \beta) = L_F(\mathrm{id}_X \otimes \delta) * (\theta_{\Xi_+} * \beta) = (L_F(\mathrm{id}_X \otimes \delta) * \theta_{\Xi_+}) * \beta = \delta * \beta = \beta,$$

and so $\theta_{\Xi_+} * \beta \in \mathrm{Sol}(F, \beta)$. We similarly have $\theta_{\Xi_-} * \beta \in \mathrm{Sol}(F, \beta)$.

(ii) Suppose that $\theta_1, \theta_2 \in \mathrm{Sol}(F, \beta) \cap \mathscr{D}_+(\mathbb{R}; \mathbb{R})$. Then

$$\begin{aligned} & L_F(\theta_1) = \beta,\ L_F(\theta_2) = \beta \\ \implies & L_F(\mathrm{id}_X \otimes \delta) * \theta_1 = \beta,\ L_F(\mathrm{id}_X \otimes \delta) * \theta_2 = \beta \\ \implies & \theta_1 = \theta_{\xi_+} * \beta,\ \theta_2 = \theta_{\xi_+} * \beta \\ \implies & \theta_1 = \theta_2. \end{aligned}$$

what

Here we use the fact that θ_{Ξ_+} is the *unique* inverse of $L_F(\mathrm{id}_X \otimes \delta)$ in $\mathscr{D}'_+(\mathbb{R}; L(X; X))$, by .

(iii) This follows in the same manner as the previous part of the theorem. ■

In summary, distributional equations for constant coefficient equations always have solutions, and if we look for solutions in $\mathscr{D}'_+(\mathbb{R}; X)$ (resp. $\mathscr{D}'_-(\mathbb{R}; X)$) for equations where the forcing is in $\mathscr{D}'_+(\mathbb{R}; X)$ (resp. $\mathscr{D}'_-(\mathbb{R}; X)$), then solutions are unique. Moreover, the proof of the theorem furnishes formulae for the unique solutions in $\mathscr{D}'_+(\mathbb{R}; X)$ and $\mathscr{D}'_-(\mathbb{R}; X)$. Let us present this, outside the stodgy confines of a proof, in the case of $\mathscr{D}'_+(\mathbb{R}; X)$.

5.3.14 Corollary (Solutions to distributional differential equations in $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$) *For a system of linear homogeneous ordinary differential equation* F *with constant coefficients and with right-hand side*

$$\widehat{F}(t, x) = A(x),$$

let

$$\Xi_+(t) = \begin{cases} \Phi_A^c(t, 0), & t \geq 0, \\ 0, & t < 0. \end{cases}$$

Then the unique solution in $\mathscr{D}'_+(\mathbb{R}; X)$ *to* (F, β) *for* $\beta \in \mathscr{D}'_+(\mathbb{R}; X)$ *is* $\theta_{\Xi_+} * \beta$.

Note that the uniqueness of solutions in $\mathscr{D}'_+(\mathbb{R};\mathsf{X})$ and $\mathscr{D}'_-(\mathbb{R};\mathsf{X})$ are in contrast to the situation in Proposition 5.3.2 where, to achieve uniqueness, one needs to also prescribe initial conditions. One might then wonder whether the rôle of initial conditions can be mimicked for distributional differential equations. This is possible, and is presented in Proposition 5.3.17 below.

Next we further connect solutions to distributional equations to their non-distributional counterparts by constructing the distributional solution to a non-distributional equation, including initial conditions.

Let us get started by noting that, if F is a system of linear homogeneous ordinary differential equation with constant coefficients and if $b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R};\mathsf{X})$ satisfies $\inf \operatorname{supp}(b) > -\infty$, then there is a unique solution $\theta \in \mathscr{D}'_+(\mathbb{R};\mathsf{X})$ to (F, θ_b). This is a consequence of Theorem 5.3.13(ii). One might then wonder whether there are other distributional solutions, not in $\mathscr{D}'_+(\mathbb{R};\mathsf{X})$, to (F, θ_b). The following result indicates the constraints on other such solutions.

5.3.15 Proposition (Uniqueness of distributional solutions to non-distributional equations) *Let* F *be a system of linear homogeneous ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{\mathrm{F}}\colon \mathbb{R}\times\mathsf{X} &\to \mathsf{X}\\ (\mathrm{t},\mathrm{x}) &\mapsto \mathsf{A}(\mathrm{x}),\end{aligned}$$

and let $\mathrm{b} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R};\mathsf{X})$ *satisfy* $\inf\operatorname{supp}(\mathrm{b}) > -\infty$. *Then, if* $\theta \in \operatorname{Sol}(\mathrm{F}, \theta_{\mathrm{b}})$, *we have*

$$\langle \theta; \phi\rangle = \langle \theta_\Xi; \phi\rangle, \qquad \phi \in \mathscr{D}((\inf\operatorname{supp}(\mathrm{b}), \infty); \mathbb{R}),$$

where ξ *satisfies*

$$\dot{\xi}(\mathrm{t}) = \mathsf{A}(\xi(\mathrm{t})) + \mathrm{b}(\mathrm{t}), \qquad \mathrm{t} \in (\inf\operatorname{supp}(\mathrm{b}), \infty).$$

In particular,

$$\operatorname{Sol}(\mathrm{F}, 0) = \{\theta_\xi \mid \xi \in \operatorname{Sol}(\mathrm{F})\},$$

i.e., solutions to the homogeneous equation in $\mathscr{D}'(\mathbb{R};\mathsf{X})$ *are exactly the regular distributions associated to the usual solutions of the homogeneous equation.*

Proof The idea of the proof of Proposition 4.4.9 for scalar equations applies in this case as well. ∎

An important consequence of the preceding result is the following complete characterisation of $\operatorname{Sol}(F, \beta)$ for $\beta \in \mathscr{D}'_+(\mathbb{R};\mathsf{X})$. Of course, a similar result holds for $\mathscr{D}'_-(\mathbb{R};\mathsf{X})$.

5.3.16 Corollary (Characterisation of $\operatorname{Sol}(\mathrm{F}, \beta)$ for $\beta \in \mathscr{D}'_+(\mathbb{R};\mathsf{X})$) *Let* F *be a system of linear homogeneous ordinary differential equation with constant coefficients, let* $\beta \in \mathscr{D}'_+(\mathbb{R};\mathsf{X})$, *and let* θ_0 *be the unique solution to* (F, β) *in* $\mathscr{D}'_+(\mathbb{R};\mathsf{X})$, *as in Theorem 5.3.13(ii). Then*

$$\operatorname{Sol}(\mathrm{F}, \beta) = \{\theta_0 + \theta_\xi \mid \xi \in \operatorname{Sol}(\mathrm{F})\}.$$

Proof If $\theta \in \mathrm{Sol}(F, \beta)$ then $L_F(\theta - \theta_0) = 0$. By Proposition 5.3.17, this means that $\theta - \theta_0$ is a regular distribution associated to a solution of the homogeneous equation F. ■

Now let us see how we can resolve the seeming paradox of the uniqueness of solutions asserted in Proposition 4.4.9 with the non-uniqueness arising from Proposition 5.3.2 (due to dependence on initial conditions). We do this by conjuring a distribution as right-hand side that incorporates the initial conditions.

5.3.17 Proposition (Distributional solutions of non-distributional equations with initial conditions) *For a system of linear homogeneous ordinary differential equation* F *with constant coefficients and with right-hand side*

$$\widehat{\mathrm{F}}(\mathsf{t}, \mathsf{x}) = \mathsf{A}(\mathsf{x}),$$

for $\mathsf{b} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathsf{X})$*, and for* $\mathsf{t}_0 \in \mathbb{R}$*, the following statements are equivalent for* $\xi \colon \mathbb{R} \to \mathsf{X}$*:*

(i) $\xi = \tau^*_{\mathsf{t}_0} 1_{\geq 0} \xi_{\mathsf{t}_0}$*, where* ξ_{t_0} *satisfies the initial value problem*

$$\dot{\xi}(\mathsf{t}) = \mathsf{A}(\xi(\mathsf{t})) + \mathsf{b}(\mathsf{t}), \quad \xi(\mathsf{t}_0) = \mathsf{x}_0;$$

(ii) the distribution θ_ξ *is the unique solution in* $\mathscr{D}'_+(\mathbb{R}; \mathsf{X})$ *to* (F, β)*, where*

$$\beta = \theta_{\tau^*_{\mathsf{t}_0} 1_{\geq 0} \mathsf{b}} + (\tau^*_{\mathsf{t}_0} \delta)\mathsf{x}_0.$$

Moreover, θ_ξ *is the unique solution in* $\mathscr{D}'(\mathbb{R}; \mathsf{X})$ *to* (F, β)*.*

Proof The proof is a mild notational adaptation of that for Proposition 4.4.11 in the case of $k = 1$. ■

Exercises

5.3.1 Consider the first-order scalar linear homogeneous ordinary differential equation with right-hand side $\widehat{F}(t, x) = a(t)x + b(t)$ for $a, b \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathbb{R})$. Using your result from Exercise 5.2.6, use Corollary 5.3.3 to determine the solution to the initial value problem

$$\dot{\xi}(t) = a(t)\xi(t) + b(t), \quad \xi(t_0) = x_0,$$

thinking of this as a system of linear inhomogeneous ordinary differential equations in the one-dimensional vector space $\mathbb{R}$.

5.3.2 Consider the scalar linear inhomogeneous ordinary differential equation F given by

$$F(t, x, x^{(1)}, x^{(2)}) = x^{(2)} + \omega^2 x - \sin(\omega t)$$

for $\omega \in \mathbb{R}_{>0}$. Answer the following questions.

(a) Use the method of undetermined coefficients to obtain a particular solution for F.

(b) Convert $\boldsymbol{F}$ into a system of linear inhomogeneous ordinary differential equations $\boldsymbol{F}_1$ in $\mathbb{R}^2$ with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}\colon \mathbb{T}\times\mathbb{R}^2 &\to \mathbb{R}^2\\ (t,(x_1,x_2)) &\mapsto \boldsymbol{A}\begin{bmatrix}x_1\\x_2\end{bmatrix}+\boldsymbol{b}(t),\end{aligned}$$

giving explicit formulae for $\boldsymbol{A}\in \mathrm{L}(\mathbb{R}^2;\mathbb{R}^2)$ and $\boldsymbol{b}\colon \mathbb{T}\to\mathbb{R}^2$.

(c) Show that

$$\mathrm{e}^{\boldsymbol{A}t}=\begin{bmatrix}\cos(\omega t) & \frac{1}{\omega}\sin(\omega t)\\ -\omega\sin(\omega t) & \cos(\omega t)\end{bmatrix}.$$

(d) Compute

$$\xi_{\mathrm{p},b}(t)=\int_0^t \mathrm{e}^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}(\tau)\,\mathrm{d}\tau.$$

Use your answer to give a particular solution for the scalar equation $\boldsymbol{F}$.

(e) Explain how the particular solutions from parts (a) and (d) are the same, and explain how to describe the difference between them.

5.3.3 For the linear transformations $\boldsymbol{A}\in\mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$ of Exercise 5.2.11, use Theorem 5.3.8 to determine the solution to the initial value problem

$$\dot{\xi}(t)=\boldsymbol{A}\xi(t)+\boldsymbol{b}(t),\quad \xi(0)=\boldsymbol{0},$$

with $\boldsymbol{b}$ as follows:

(a) $\boldsymbol{b}(t)=(0,1)$;
(b) $\boldsymbol{b}(t)=(\cos(t),0)$;
(c) $\boldsymbol{b}(t)=(\mathrm{e}^{2t},0)$;
(d) $\boldsymbol{b}(t)=(\sin(t),0,1)$;
(e) $\boldsymbol{b}(t)=(0,\mathrm{e}^{-t},0)$;
(f) $\boldsymbol{b}(t)=(\sin(2t),0,1)$;
(g) $\boldsymbol{b}(t)=(1,0,0,1)$;
(h) $\boldsymbol{b}(t)=(\sin(t),0,0,\cos(t))$;
(i) $\boldsymbol{b}(t)=(0,0,0,0,0)$.

5.3.4 Let V be a finite-dimensional $\mathbb{R}$-vector space, let $\mathsf{A}\in\mathrm{L}(\mathsf{V};\mathsf{V})$, and let $x_0\in\mathsf{V}$. Show, by direct computation, that the unique solution in $\mathscr{D}'_+(\mathbb{R};\mathsf{V})$ to the equation

$$\theta^{(1)}=A(\theta)+x_0\otimes\delta$$

is given by $\theta=\theta_{1_{\geq 0}\xi}$, where $\xi\in\mathsf{C}^\infty(\mathbb{R};\mathsf{V})$ is the solution to the initial value problem

$$\frac{\mathrm{d}\xi}{\mathrm{d}t}(t)=A\circ\xi(t),\quad \xi(0)=x_0.$$

Demonstrate that you understand each part of the computation by pointing to the place in the text where your assertion is defined or shown to make sense.

Section 5.4

Laplace transform methods for systems of ordinary differential equations

In this section we consider an application of the causal CLT to systems of ordinary differential equations. As with our consideration of scalar equations in Section 4.5, we work with linear constant coefficient equations, both homogeneous and inhomogeneous.

Do I need to read this section? Like Section 4.5, one might skip this chapter at a first reading, until one is confronted with the transfer function methods of Chapter 7, and the use of the tool of the causal CLT makes more sense. •

5.4.1 Systems of homogeneous equations

Now we turn to studying systems of equations using the causal CLT, starting with the homogeneous case. As we did in Section 5.2, we shall work with systems whose state space is a finite-dimensional $\mathbb{R}$-vector space V. We refer to Section IV-9.1.7 for a discussion of how the causal CLT work in this setting.

We consider a system of linear ordinary differential equations F with constant coefficients in an n-dimensional $\mathbb{R}$-vector space V, and with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{R}_{\geq 0} \times \mathsf{V} &\to \mathsf{V}\\ x &\mapsto \mathsf{A}(x)\end{aligned}$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$. The associated initial value problem we study is then

$$\dot{\xi}(t) = \mathsf{A}(\xi(t)), \quad \xi(0) = x_0. \tag{5.19}$$

Let us take the causal CLT of this initial value problem.

5.4.1 Proposition (Causal CLT of system of homogeneous equations) *The causal CLT of the solution of the initial value problem (5.19) is*

$$\mathscr{L}^\infty_{\mathrm{C}}(\xi)(\mathrm{z}) = (\mathrm{z}\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}\mathrm{x}_0,$$

and $\mathscr{L}^\infty_{\mathrm{C}}(\xi)$ is defined on

$$\{\mathrm{z} \in \mathbb{C} \mid \ \mathrm{Re}(\mathrm{z}) > \mathrm{Re}(\lambda) \text{ for all } \lambda \in \mathrm{spec}(\mathsf{A})\}.$$

Proof This is a direct computation using Proposition IV-9.1.20:

$$z\mathscr{L}^\infty_{\mathrm{C}}(\xi)(z) - \xi(0) = \mathsf{A}\mathscr{L}^\infty_{\mathrm{C}}(\xi)(z),$$

from which the result follows immediately after noting that $z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A}$ is invertible if the real part of z exceeds the real part of any eigenvalue of A. ■

As with scalar equations, the application of the causal CLT permits a solution for systems of linear homogeneous equations with constant coefficients using just algebraic computations in the transformed variables. In order to understand the inverse $(z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}$, let us think about how one may compute this inverse. We shall suppose that we have a basis $\{e_1, \dots, e_n\}$ for V and let $\boldsymbol{A} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ be the matrix representative for A. Then the matrix representative for $(z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}$ is $(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}$. For $\boldsymbol{B} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$, let us denote by $\mathrm{Cof}(\boldsymbol{B})$ the $n \times n$-matrix whose (j,k)th entry is $(-1)^{j+k} \det \hat{\boldsymbol{B}}(j,k)$, where $\hat{\boldsymbol{B}}(j,k)$ is the $(n-1) \times (n-1)$-matrix obtained by deleting the jth row and kth column from $\boldsymbol{B}$. Then, by Theorem I-5.3.10,

$$\mathrm{Cof}(\boldsymbol{B})^T \boldsymbol{B} = \boldsymbol{B}\,\mathrm{Cof}(\boldsymbol{B})^T = (\det \boldsymbol{B}) \boldsymbol{I}_n.$$

Therefore,

$$(z\boldsymbol{I}_n - \boldsymbol{A})^{-1} = \frac{\mathrm{Cof}(z\boldsymbol{I}_n - \boldsymbol{A})^T}{\det(z\boldsymbol{I}_n - \boldsymbol{A})}.$$

Note that the entries of $\mathrm{Cof}(z\boldsymbol{I} - \boldsymbol{A})$ are determinants of $(n-1)\times(n-1)$-matrices whose entries are polynomials of degree at most 1 in z. Thus the entries of $\mathrm{Cof}(z\boldsymbol{I}_n - \boldsymbol{A})$ are polynomials of degree at most $n-1$. Thus, since $\det(z\boldsymbol{I}_n - \boldsymbol{A})$ is a monic polynomial of degree n in z, the entries of $(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}$ are rational functions in z whose numerator polynomial has degree strictly less than that of the denominator polynomial. Therefore, the inverse causal CLT of $(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}$ can be computed by performing a partial fraction expansion on each of its entries, and then applying the inverse causal CLT of Example IV-9.1.15.

However, the inverse causal CLT of $(z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}$ is known to us already.

5.4.2 Proposition (Causal CLT of operator exponential) *For an* n*-dimensional* $\mathbb{R}$*-vector space* V *and for* $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$*, denote*

$$\begin{aligned} \exp_{\mathsf{A}} \colon \mathbb{R}_{\geq 0} &\to \mathrm{L}(\mathsf{V}; \mathsf{V}) \\ t &\mapsto \mathrm{e}^{\mathsf{A}t}. \end{aligned}$$

Then $\mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}})(z) = (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}$.

Proof By Theorem 5.2.6(i) and since $\exp_{\mathsf{A}}(t) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t, 0)$, we note that $\exp_{\mathsf{A}}$ satisfies the initial value problem

$$\frac{\mathrm{d} \exp_{\mathsf{A}}}{\mathrm{d}t}(t) = \mathsf{A} \circ \exp_{\mathsf{A}}(t), \quad \exp_{\mathsf{A}}(0) = \mathrm{id}_{\mathsf{V}}.$$

Taking the causal CLT of this initial value problem gives

$$z\,\mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}})(z) - \mathrm{id}_{\mathsf{V}} = \mathsf{A} \circ \mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}})(z) \quad \Longrightarrow \quad \mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}})(z) = (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1},$$

as claimed. ■

Let's illustrate this in a simple example.

5.4.3 Example (Operator exponential via the causal CLT) We consider the linear map $A \in \mathrm{L}(\mathbb{R}^2; \mathbb{R}^2)$ considered in Example 5.2.27:

$$A = \begin{bmatrix} -7 & 4 \\ -6 & 3 \end{bmatrix}.$$

We compute

$$(zI_2 - A)^{-1} = \begin{bmatrix} \frac{z-3}{z^2+4z+3} & \frac{4}{z^2+4z+3} \\ -\frac{6}{z^2+4z+3} & \frac{z+7}{z^2+4z+3} \end{bmatrix}.$$

We then use partial fraction expansions:

$$\begin{aligned} \frac{z-3}{z^2+4z+3} &= -\frac{2}{z+1} + \frac{3}{z+3}, \\ \frac{4}{z^2+4z+3} &= \frac{2}{z+1} - \frac{2}{z+3}, \\ -\frac{6}{z^2+4z+3} &= -\frac{3}{z+1} + \frac{3}{z+3}, \\ \frac{z+7}{z^2+4z+3} &= \frac{3}{z+1} - \frac{2}{z+3}. \end{aligned}$$

Using Example IV-9.1.15–2, we apply the inverse transform to get

$$\mathrm{e}^{At} = \begin{bmatrix} 3\mathrm{e}^{-3t} - 2\mathrm{e}^{-t} & -2\mathrm{e}^{-3t} + 2\mathrm{e}^{-t} \\ 3\mathrm{e}^{-3t} - 3\mathrm{e}^{-t} & -2\mathrm{e}^{-3t} + 3\mathrm{e}^{-t} \end{bmatrix},$$

just as in Example 5.2.27. •

It is a matter of taste whether one thinks that using the causal CLT to compute the operator exponential is preferable to Procedure 5.2.26. It is, however, not such an important matter to resolve in favour of one method or the other; actually computing the operator exponential is seldom of interest *per se*. What is certainly true is that with the causal CLT one loses the insight offered by invariant subspaces in Procedure 5.2.26. The benefits of the causal CLT in this context arises in system theory, where complex function techniques offer some genuine insights.

5.4.2 Systems of inhomogeneous equations

Next we consider systems of homogeneous equations. Thus we have an ordinary differential equation with state space V and with right-hand side

$$\begin{aligned} \widehat{F} \colon \mathbb{R}_{\geq 0} \times \mathsf{V} &\to \mathsf{V} \\ x &\mapsto \mathsf{A}(x) + b(t), \end{aligned} \tag{5.20}$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ and for $b \colon \mathbb{R}_{\geq 0} \to \mathsf{V}$. The associated initial value problem we consider is

$$\dot{\xi}(t) = \mathsf{A}(\xi(t)) + b(t), \quad \xi(0) = x_0. \tag{5.21}$$

We can, of course, easily take the causal CLT of this initial value problem to get the following.

5.4.4 Proposition (Causal CLT of system of inhomogeneous equations) *Consider the system of scalar ordinary differential equations with right-hand side (5.20), and suppose that* b *is continuous and satisfies* $\mathsf{b} \in \mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; \mathsf{V})$. *The causal CLT of the solution of the initial value problem (5.21) satisfies*

$$\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) = (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}(\mathrm{x}_0 + \mathscr{L}^{\infty}_{\mathrm{C}}(\mathrm{b})(z)).$$

Proof The proof is an easy adaptation of that of Proposition 5.4.1. ■

As was the case with our discussion of scalar inhomogeneous equations in Section 4.5.2, the preceding result can be interpreted in two ways, one having theoretical value and the other as a means of computing solutions. We shall explore both.

The first result makes a connection with the formula given in Corollary 5.3.3 for solutions to systems of linear inhomogeneous equations, in the general setting of time-varying systems.

5.4.5 Proposition (Causal CLT and convolution for solutions of linear inhomogeneous equations) *Consider the system of scalar ordinary differential equations with right-hand side (5.20), and suppose that* $\mathsf{b} \in \mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; \mathsf{V})$. *Then the solution to the initial value problem (5.21) is*

$$\xi(\mathrm{t}) = \mathrm{e}^{\mathsf{A}\mathrm{t}}(\mathrm{x}_0) + \exp_{\mathsf{A}} * \mathrm{b}(\mathrm{t}).$$

Proof This follows immediately from Corollary 5.3.3, after understanding that

$$\exp_{\mathsf{A}} * b(t) = \int_0^t \mathrm{e}^{\mathsf{A}(t-\tau)}(b(\tau))\,\mathrm{d}\tau.$$

However, here we shall give a proof using the causal CLT, valid when $b \in \mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; \mathsf{V})$.

From Proposition 5.4.4 we have

$$\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) = (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}(x_0) + (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}\mathscr{L}^{\infty}_{\mathrm{C}}(b)(z).$$

By Proposition 5.4.2 we have

$$(z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1} = \mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}})(z).$$

For $x \in \mathsf{V}$, let us denote

$$\begin{aligned} \mathrm{ev}_x \colon \mathrm{L}(\mathsf{V}; \mathsf{V}) &\to \mathsf{V} \\ \mathsf{A} &\mapsto \mathsf{A}(x). \end{aligned}$$

We then have, noting that ev_{x_0} is a linear map,

$$\mathscr{L}^{\infty}_{\mathrm{C}}(\mathrm{ev}_{x_0} \circ \exp_{\mathsf{A}})(z) = \mathrm{ev}_{x_0} \circ \mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}})(z) = (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})(x_0).$$

Also, by Proposition IV-9.1.10,

$$\mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}} * b)(z) = \mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}})(z)\mathscr{L}^{\infty}_{\mathrm{C}}(b)(z) = (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})\mathscr{L}^{\infty}_{\mathrm{C}}(b)(z).$$

Therefore,

$$\mathscr{L}^{\infty}_{\mathrm{C}}(\xi)(z) = \mathrm{ev}_{x_0} \circ \mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}}) + \mathscr{L}^{\infty}_{\mathrm{C}}(\exp_{\mathsf{A}} * b)(z).$$

Taking the inverse causal CLT gives

$$\xi(t) = \mathrm{ev}_{x_0} \circ \mathrm{e}^{\mathsf{A}t} + \exp_{\mathsf{A}} * b(t) = \mathrm{e}^{\mathsf{A}t}(x_0) + \exp_{\mathsf{A}} * b(t),$$

as claimed. ■

Finally, in the case when b is an also pretty interesting function (meaning that, in a basis for V, the components of b are also pretty uninteresting functions), we can use Proposition 5.4.4, and partial fraction expansions, to compute solutions. We only validate this by a simple example since, in reality, this is not something one ever does.

5.4.6 Example (Solving systems of inhomogeneous equations using the causal CLT) We take $\mathsf{V} = \mathbb{R}^2$ and

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix}, \qquad \boldsymbol{b}(t) = \begin{bmatrix} 0 \\ \sin(\omega t) \end{bmatrix}.$$

We then calculate

$$(z\boldsymbol{I}_2 - \boldsymbol{A})^{-1} = \begin{bmatrix} \frac{z}{z^2+\omega^2} & \frac{1}{z^2+\omega^2} \\ -\frac{\omega^2}{z^2+\omega^2} & \frac{z}{z^2+\omega^2} \end{bmatrix}, \qquad \mathscr{L}^\infty_{\mathrm{C}}(b)(z) = \begin{bmatrix} 0 \\ \frac{\omega}{z^2+\omega^2} \end{bmatrix}.$$

Thus, by Proposition 5.4.4,

$$\begin{aligned} \mathscr{L}^\infty_{\mathrm{C}}(\xi)(z) &= \begin{bmatrix} \frac{z}{z^2+\omega^2} & \frac{1}{z^2+\omega^2} \\ -\frac{\omega^2}{z^2+\omega^2} & \frac{z}{z^2+\omega^2} \end{bmatrix} \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix} + \begin{bmatrix} \frac{z}{z^2+\omega^2} & \frac{1}{z^2+\omega^2} \\ -\frac{\omega^2}{z^2+\omega^2} & \frac{z}{z^2+\omega^2} \end{bmatrix} \begin{bmatrix} 0 \\ \frac{\omega}{z^2+\omega^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\omega}{(z^2+\omega^2)^2} + \frac{x_{01}z+x_{02}}{z^2+\omega^2} \\ \frac{\omega z}{(z^2+\omega^2)^2} + \frac{x_{02}z-x_{01}\omega^2}{z^2+\omega^2} \end{bmatrix}. \end{aligned}$$

The last line was arrived at by performing the matrix multiplication, then performing a partial fraction expansion of the entries of the resulting vector. This, then, is a bit of effort that we do not fully illustrate. In any case, one can apply the conclusions of Example IV-9.1.15–4 and Example IV-9.1.15–5 to arrive at

$$\xi(t) = \begin{bmatrix} \frac{1}{2\omega^2}\sin(\omega t) - \frac{t}{2\omega}\cos(\omega t) + x_{01}\cos(\omega t) + \frac{x_{02}}{\omega}\sin(\omega t) \\ \frac{t}{2}\sin(\omega t) + -\omega x_{01}\sin(\omega t) + x_{02}\cos(\omega t) \end{bmatrix}.$$

We encourage the reader to understand the relationship between this answer and the one from Example 4.5.7. •

As with systems of homogeneous equations, the use of the Laplace transform to solve inhomogeneous equations does not have a lot to recommend it from a computational point of view. The advantages it has come more from exploiting the algebraic structure of the differential equation as a function of the transformed independent variable z.

Exercises

5.4.1 Determine the causal CLT of the solution of the initial value problem

$$\dot{\xi}(t) = A\xi(t), \quad \xi(0) = x_0,$$

for the following choices of $A \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $x_0 \in \mathbb{R}^n$:

(a) $A = \begin{bmatrix} 2 & -5 \\ 0 & 3 \end{bmatrix}$, $x_0 = (0, 1)$;

(b) $A = \begin{bmatrix} -1 & -2 \\ 1 & -3 \end{bmatrix}$, $x_0 = (2, -3)$;

(c) $A = \begin{bmatrix} 4 & -1 \\ 4 & 0 \end{bmatrix}$, $x_0 = (1, 1)$;

(d) $A = \begin{bmatrix} 5 & 0 & -6 \\ 0 & 2 & 0 \\ 3 & 0 & -4 \end{bmatrix}$, $x_0 = (-3, -1, 0)$;

(e) $A = \begin{bmatrix} 5 & 0 & -6 \\ 1 & 2 & -1 \\ 3 & 0 & -4 \end{bmatrix}$, $x_0 = (1, 0, 1)$;

(f) $A = \begin{bmatrix} 4 & 2 & -4 \\ 2 & 0 & -4 \\ 2 & 2 & -2 \end{bmatrix}$, $x_0 = (4, 1, 2)$;

(g) $A = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 3 & -1 & 3 \\ 0 & 1 & 2 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$, $x_0 = (1, -1, 0, 1)$;

(h) $A = \begin{bmatrix} -7 & 0 & 0 & -4 \\ -13 & -2 & -1 & -8 \\ 6 & 1 & 0 & 4 \\ 15 & 1 & 0 & 9 \end{bmatrix}$, $x_0 = (-1, -1, 3, -2)$;

(i) $A = \begin{bmatrix} 1 & 4 & -2 & 0 & 9 \\ 0 & -2 & 1 & 2 & -6 \\ -2 & 4 & -1 & 3 & 0 \\ -9 & 4 & 1 & 0 & 2 \\ 4 & 0 & 3 & -1 & 3 \end{bmatrix}$, $x_0 = (0, 0, 0, 0, 0)$.

NB. These are the same initial value problems you worked out in Exercise 5.2.12.

5.4.2 Using partial fraction expansion, compute e^{At} for the linear transformations $A \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ from Exercise 5.4.1.

5.4.3 Determine the causal CLT of the solution of the initial value problem

$$\dot{\xi}(t) = A\xi(t) + b(t), \quad \xi(0) = 0,$$

for the choices of $A \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ from Exercise 5.4.1 and for the following b:

(a) $b(t) = (0, 1)$;

(b) $b(t) = (\cos(t), 0)$;

(c) $b(t) = (\mathrm{e}^{2t}, 0)$;

(d) $b(t) = (\sin(t), 0, 1)$;

(e) $b(t) = (0, \mathrm{e}^{-t}, 0)$;

(f) $b(t) = (\sin(2t), 0, 1)$;

(g) $b(t) = (1, 0, 0, 1)$;

(h) $b(t) = (\sin(t), 0, 0, \cos(t))$;

(i) $b(t) = (0, 0, 0, 0, 0)$.

NB. These are the same initial value problems you worked out in Exercise 5.3.3.

5.4.4 Using partial fraction expansion, determine the solution to the initial value problems from Exercise 5.4.3.

Section 5.5

Phase-plane analysis for differential equations

In this section we consider a way of representing the behaviour of ordinary differential equations whose state space is a subset of $\mathbb{R}^2$ via their "phase portraits." We have already used this method informally on a number of occasions, and in this section we shall be a little more systematic. We begin in Section 5.5.1 by exhaustively examining phase portraits for linear systems in two variables. In Section 5.5.2 we consider phenomenon that can happen for nonlinear systems. In this case, the presentation is essentially example driven, and we give little by way of rigorous methodology. This analysis appears a little *ad hoc*, however, the methods can give more insight into what is "really happening" with a differential equation. Also, the ideas that we encounter in the simple two-dimensional setting suggest techniques that may be profitably applied in higher-dimensions. These ideas are discussed in Section 5.5.3.

5.5.1 Phase portraits for linear systems

We begin our discussion with a consideration of phase portraits for systems of linear ordinary differential equations in $\mathbb{R}^2$ with constant coefficients. Thus we are considering differential equations $\boldsymbol{F}$ with

$$\begin{aligned}\widehat{\boldsymbol{F}}\colon &\mathbb{T}\times\mathbb{R}^2\to\mathbb{R}^2\\ &(t,(x_1,x_2))\mapsto \underbrace{\begin{bmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{bmatrix}}_{A}\begin{bmatrix}x_1\\x_2\end{bmatrix}.\end{aligned}$$

In Section 5.2.2 we learned that the solution to the initial value problem

$$\begin{bmatrix}\dot\xi_1(t)\\\dot\xi_2(t)\end{bmatrix}=\begin{bmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{bmatrix}\begin{bmatrix}\xi_1(t)\\\xi_2(t)\end{bmatrix},\quad\begin{bmatrix}\xi_1(0)\\\xi_2(0)\end{bmatrix}=\boldsymbol{x}_0=\begin{bmatrix}x_{0,1}\\x_{2,0}\end{bmatrix},$$

is $\xi(t)=\mathrm{e}^{At}\boldsymbol{x}_0$. What we shall do in this section is represent these solutions in a particular way, such as we initially discussed in Example 3.1.25. To be specific, we shall plot the solutions as parameterised curves in the (x_1,x_2)-plane. In doing this, we shall represent, not just one solution, but the entirety of solutions with various initial conditions. By doing this, one gets a qualitative understanding of the behaviour of solutions that is simply not achievable by looking at a closed-form solution or by looking at plots of $t\mapsto\xi_1(t)$ and $t\mapsto\xi_2(t)$ of *fixed* solutions with a single initial condition. The resulting collection of solutions, represented as parameterised curves, is called the ***phase portrait***.

We shall break down the analysis into various cases, based on the character of eigenvalues and eigenvectors.

5.5.1.1 Stable nodes We first consider the case where there are two negative real eigenvalues. In this case, there are a few cases to consider, but all fall into the general category of what we call a ***stable node***, since, as we shall see, all solutions tend to $(0,0)$ as $t \to \infty$.

Distinct eigenvalues

Here we suppose that we have eigenvalues $\lambda_1, \lambda_2 \in \mathbb{R}$ with $\lambda_1 < \lambda_2 < 0$. The behaviour in the case is then determined by the eigenvectors. Let us first look at the simple case where the eigenvectors are the standard basis vectors $e_1 = (1,0)$ and $e_2 = (0,1)$. In this case, A is given by

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

Then

$$\begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix} = \mathrm{e}^{At} \begin{bmatrix} \xi_1(0) \\ \xi_2(t) \end{bmatrix} = \begin{bmatrix} \xi_1(0)\mathrm{e}^{\lambda_1 t} \\ \xi_2(0)\mathrm{e}^{\lambda_2 t} \end{bmatrix}.$$

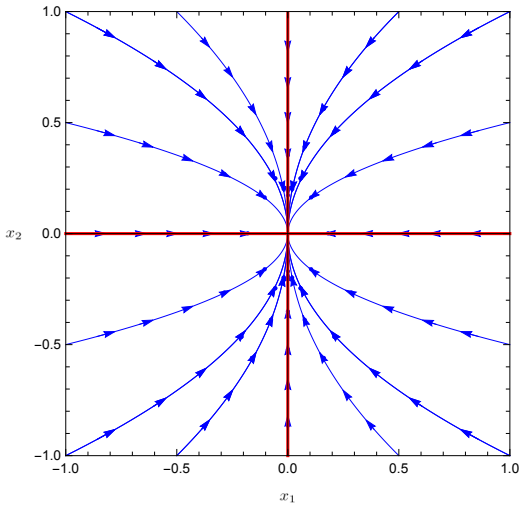
In Figure 5.1a we show the phase portrait, i.e., the family of solutions plotted as parameterised curves in the (x_1, x_2)-plane. Let us make a few comments about the nature of the phase portrait so as to explain the nature of its essential features.

1. The eigenvectors, which are e_1 and e_2 in this case, show up as lines through the origin with the property that solutions that start on these lines remain on these lines. These are, then, ***invariant subspaces*** for the dynamics. In Figure 5.1a these are indicated in red. In this case, because the eigenvalues are negative, the solutions along these lines approach $(0,0)$ as $t \to \infty$, as can be seen from the direction of the arrows.
2. Solutions corresponding to other initial conditions also approach $(0,0)$ as $t \to \infty$. From Figure 5.1a we can see that all of these other solutions approach $(0,0)$ tangent to the eigenvector e_2. The reason for this is that the eigenvalue λ_1 is the "more negative" eigenvalue, and so solutions decay to zero more quickly in the direction of the corresponding eigenvector e_1.

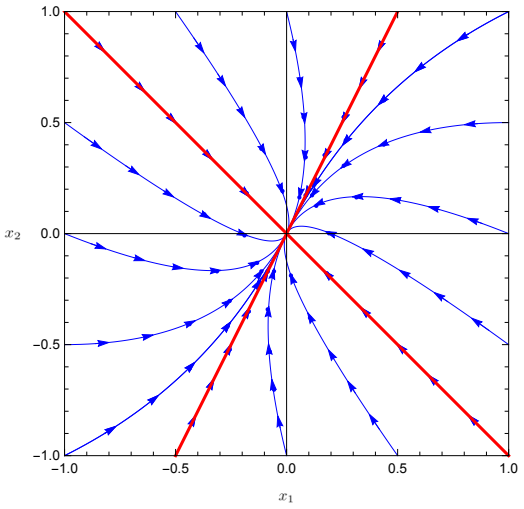
In the phase portrait of Figure 5.1a the eigenvectors are the standard basis vectors, and this was selected to make the process easier to visualise and explain. However, typically the eigenvectors are *not* the standard basis vectors, of course. However, the same ideas apply: (1) the eigenvectors represent invariant subspaces for the dynamics and (2) solutions approach $(0,0)$ more quickly in the direction of the "more negative" eigenvector. Let us illustrate this with an example, taking

$$A = \begin{bmatrix} -\frac{5}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{4}{3} \end{bmatrix}.$$

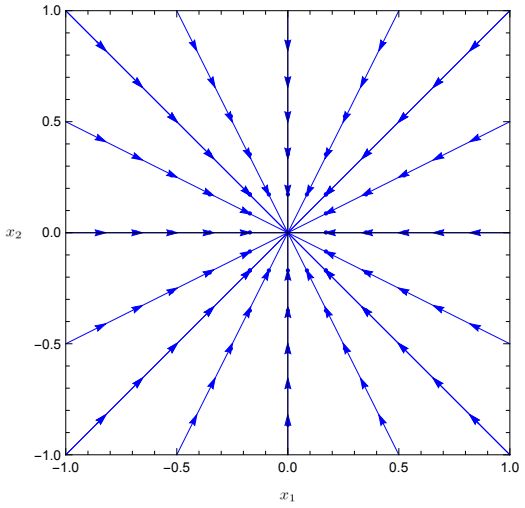
In this case we compute the eigenvalues of A to be $\lambda = -1$ and $\lambda_2 = -2$, i.e., the same eigenvalues as in the example illustrated in Figure 5.1a. Corresponding eigenvectors are $v_1 = (1,-1)$ and $v_2 = (1,2)$. In Figure 5.1b we show the phase portrait. In



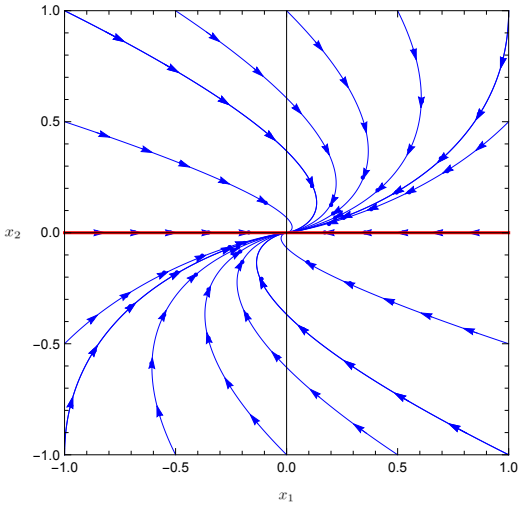
(a) Stable node with the distinct eigenvalues $\lambda_1 = -2$ and $\lambda_2 = -1$, and standard basis vectors as eigenvectors



(b) Stable node with distinct eigenvalues $\lambda_1 = -2$ and $\lambda_2 = -1$ and eigenvectors $v_1 = (1, -1)$, and $v_2 = (1, 2)$



(c) Stable node with repeated eigenvalue $\lambda = -1$ and geometric multiplicity 2



(d) Stable node with repeated eigenvalue $\lambda = -1$ and geometric multiplicity 1

Figure 5.1 Stable nodes

red we denote the invariant subspaces corresponding to the eigenvectors. Note that, essentially, once one understand the phase portrait in Figure 5.1a with the standard basis vectors as eigenvectors, it is a matter of "distortion" to produce the phase portrait of Figure 5.1b with its different eigenvectors.

Repeated eigenvalue with geometric multiplicity 2

Next we consider the case where A has a single eigenvalue $\lambda \in \mathbb{R}_{<0}$ with $m_a(\lambda, A) = m_g(\lambda, A) = 2$. In this case note that we simply have

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \implies e^{At} = \begin{bmatrix} e^{\lambda t} & 0 \\ 0 & e^{\lambda t} \end{bmatrix}.$$

That is to say, all vectors are eigenvectors. Thus the phase portrait of Figure 5.1c is perhaps not surprising.

Repeated eigenvalue with geometric multiplicity 1

Here we again consider the case where A has a single eigenvalue $\lambda \in \mathbb{R}_{<0}$ with $m_a(\lambda, A) = 2$. But in this case we assume that $m_g(\lambda, A) = 1$. A representative example is given by

$$A = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \implies e^{At} = \begin{bmatrix} e^{\lambda t} & t \\ 0 & e^{\lambda t} \end{bmatrix}.$$

The phase portrait is shown in Figure 5.1d, with the single invariant subspace indicated in red.

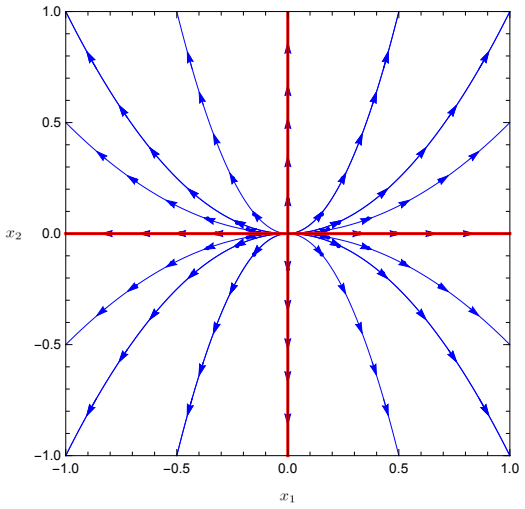
5.5.1.2 Unstable nodes The cases we consider in this section are rather like those in the previous section, except that here we will work with positive eigenvalues. In this case we have an ***unstable node*** since all solutions, except the one with initial condition $(0, 0)$, diverge to infinity as $t \to \infty$. The analysis is quite like that for stable nodes, so we will be briefer.

Distinct eigenvalues

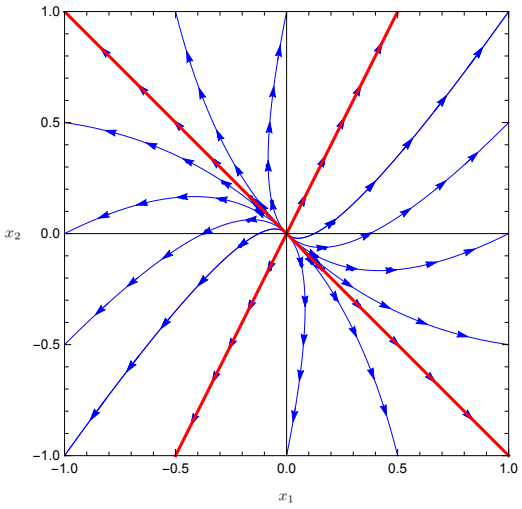
We first consider the case where A has distinct negative real eigenvalues. In this case, there will be two linearly independent eigenvectors that will each span a one-dimensional invariant subspace for the differential equation. Consider first the case where

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \implies e^{At} = \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix}$$

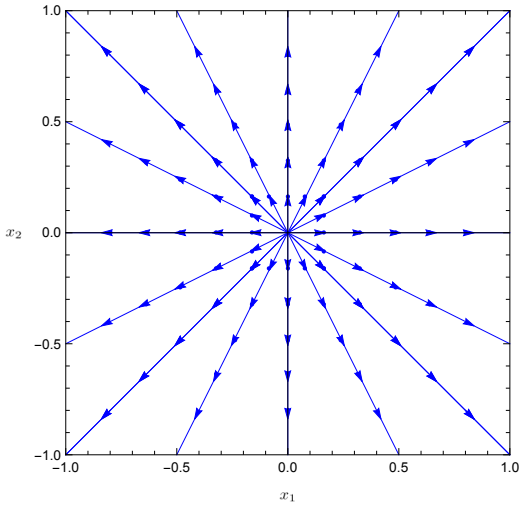
for $0 < \lambda_1 < \lambda_2$. In this case the eigenvectors are the standard basis vectors e_1 and e_2. The phase portrait is shown in Figure 5.2a for this case. We see that, the phase portrait is, in some sense, the "opposite" of that in Figure 5.1a for a stable node. One still has the invariant subspaces, but now the parameterised curves for solutions are diverging from the equilibrium at $(0, 0)$. Note that, since the divergence from $(0, 0)$ is faster in the direction of e_2, solution curves approach $(0, 0)$ faster going backwards in time. This is why solutions approach $(0, 0)$ tangent to the e_1-direction.



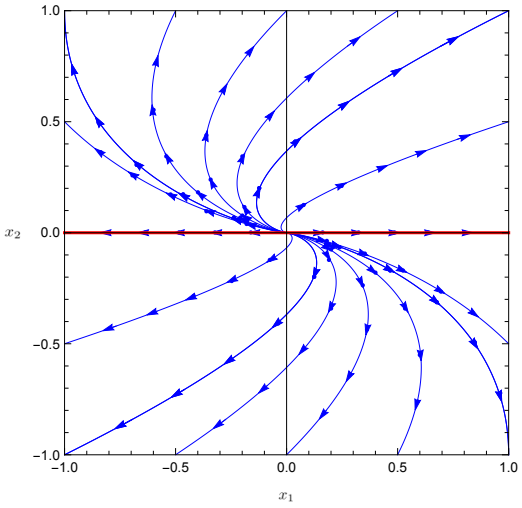
(a) Unstable node with the distinct eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 2$, and standard basis vectors as eigenvectors



(b) Unstable node with distinct eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 2$ and eigenvectors $v_1 = (1, -1)$, and $v_2 = (1, 2)$



(c) Unstable node with repeated eigenvalue $\lambda = 1$ and geometric multiplicity 2



(d) Unstable node with repeated eigenvalue $\lambda = 1$ and geometric multiplicity 1

Figure 5.2 Unstable nodes

Let us also consider a case where the eigenvectors are not the standard basis vectors. Here we take

$$\boldsymbol{A} = \begin{bmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{5}{3} \end{bmatrix},$$

which has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 2$. Associated eigenvectors are $v_1 = (1, -1)$

and $v_2 = (2, 1)$. As we see in Figure 5.2b, the phase portrait is the expected "distortion" of the phase portrait from Figure 5.2a.

Repeated eigenvalue with geometric multiplicity 2

Next we consider the case of a positive real eigenvalue λ with $m_a(\lambda, A) = m_g(\lambda, A) = 2$. In this case, A is necessarily given by

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \quad \Longrightarrow \quad e^{At} = \begin{bmatrix} e^{\lambda t} & 0 \\ 0 & e^{\lambda t} \end{bmatrix}.$$

In this case, every one-dimensional subspace is an invariant subspace along which solutions diverge to ∞. The phase portrait is shown in Figure 5.2c, and shows the expected features.

Repeated eigenvalue with geometric multiplicity 1

The final unstable node is associated to a positive eigenvalue λ with $m_a(\lambda, A) = 2$ and $m_g(\lambda, A) = 1$. In this case, we have only one one-dimensional invariant subspace associated to an eigenvector. In Figure 5.2d we show the phase portrait for this case associated with the typical example

$$A = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \quad \Longrightarrow \quad e^{At} = \begin{bmatrix} e^{\lambda t} & t \\ 0 & e^{\lambda t} \end{bmatrix}.$$

Again, we note that all solution curves, except for the one at the equilibrium $(0, 0)$, diverge to ∞ as $t \to \infty$.

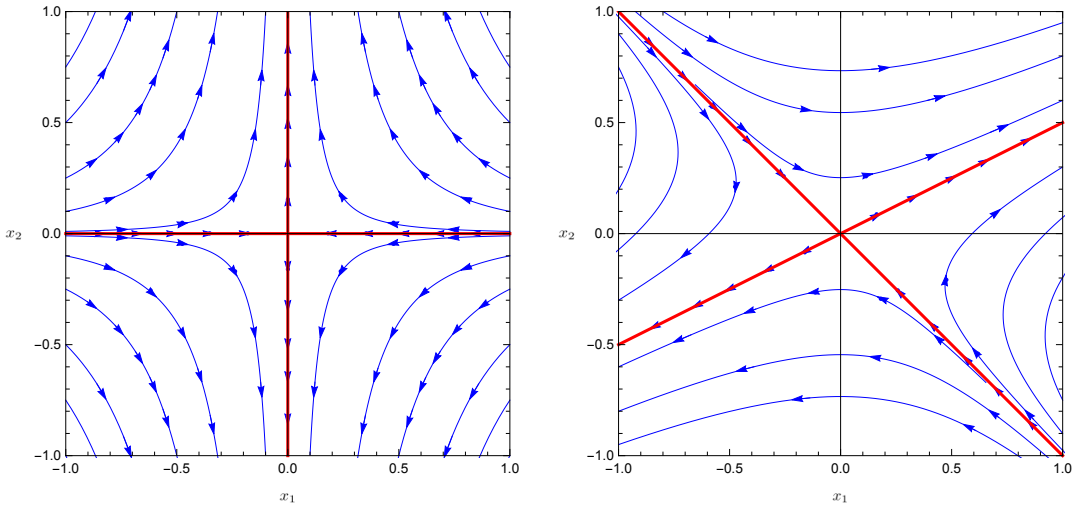
5.5.1.3 Saddle points The next case we consider is where the real eigenvalues λ_1 and λ_2 satisfy $\lambda_1 < 0 < \lambda_2$. In this case we have what is called a ***saddle point***, in reference to the setting of a function of two variables at a point where the derivative of the function vanishes and its Hessian has one positive and one negative eigenvalue.

In this case, eigenvectors for the distinct eigenvalues are necessarily linearly independent, so we do not have to carefully consider cases of differing algebraic and geometric multiplicities. Let us begin with the special case

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad \Longrightarrow \quad e^{At} = \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix},$$

where the eigenvectors are the standard basis vectors e_1 and e_2. In Figure 5.3a we show the phase portrait in this case. Let us make a few comments on what we see.

1. There are two invariant subspaces corresponding to the linearly independent eigenvectors. On the invariant subspace associated with the negative eigenvalue, the solutions converge to $(0, 0)$ as $t \to \infty$. On the invariant subspace associated with the positive eigenvalue, solutions diverge to ∞ as $t \to \infty$.



(a) Saddle point with eigenvalues $\lambda_1 = -1$ and $\lambda_2 = 2$, and standard basis vectors as eigenvectors

(b) Saddle point with eigenvalues $\lambda_1 = -1$ and $\lambda_2 = 2$, and eigenvectors $v_1 = (1, -1)$ and $v_2 = (2, 1)$

Figure 5.3 Saddle points

2. All other solutions, except for that at the equilibrium point $(0, 0)$, diverge to ∞ as $t \to \infty$, but do so after possibly falling temporarily under the influence of the negative eigenvalue.

We can, of course, adapt this to situations where the eigenvectors are not the standard basis vectors. To illustrate, let us take

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}.$$

Then the eigenvalues of A are $\lambda_1 = -1$ and $\lambda_2 = 2$, and the associated eigenvectors $v_1 = (1, -1)$ and $v_2 = (2, 1)$. The phase portrait here we depict in Figure 5.3b. It is, as expected, a "distortion" of the phase portrait in Figure 5.3a with the standard basis vectors as eigenvectors.

5.5.1.4 Centres We next consider cases where A has complex eigenvalues, first looking at the case where the eigenvalues of A are purely imaginary, say $\lambda_1 = \mathrm{i}\omega$ and $\lambda_2 = -\mathrm{i}\omega$, with $\omega \in \mathbb{R}_{>0}$. In this case we say we have a ***centre***. The prototypical case here is

$$A = \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix}.$$

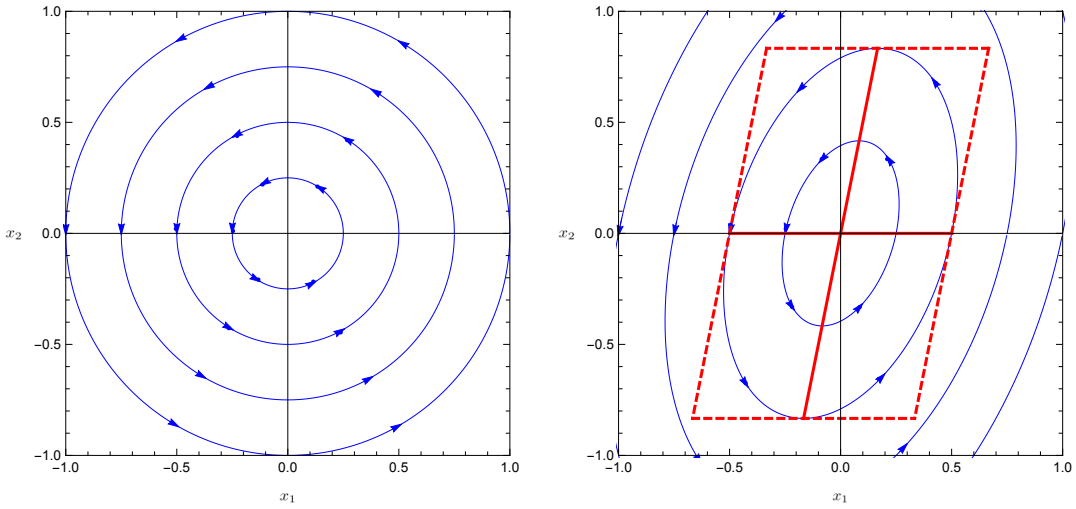
In this case we have, using Procedure 5.2.23,

$$e^{At} = \begin{bmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{bmatrix}.$$

Note that, if

$$\begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix} = e^{At} \begin{bmatrix} \xi_1(0) \\ \xi_2(0) \end{bmatrix},$$

then $\|\xi(t)\| = \|\xi(0)\|$. Thus the parameterised solution curves reside in circles centred at $(0,0)$, and this is illustrated in Figure 5.4a.



(a) "Canonical" centre (b) "Noncanonical" centre

Figure 5.4 Centres

For more generic cases, the solutions will still be periodic, and the solution curves will then live on ellipses. To describe the ellipses, we suppose that we have eigenvalues $\lambda_1 = i\omega$ and $\lambda_2 = -i\omega$. We suppose that the associated eigenvectors are $w_1 = u + iv$ and $w_2 = u - iv$ for $u, v \in \mathbb{R}^2$. To illustrate how u and v prescribe the ellipses traced out by solutions, we shall consider an example:

$$A = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{5}{3} & -\frac{1}{3} \end{bmatrix}.$$

The eigenvalues in this case are $\lambda_1 = i$ and $\lambda_2 = -i$. The eigenvectors are $w_1 = u + iv$ and $w_2 = u - iv$, where

$$u = (1, 5), \quad v = (3, 0).$$

In Figure 5.4b we illustrate the phase portrait in this case, and also show scaled eigenvectors in red, and a box centred at $(0,0)$ whose sides are parallel to the eigenvectors. As one can see, the ellipse along which solution curves evolve is the unique ellipse tangent to an appropriately scaled box.

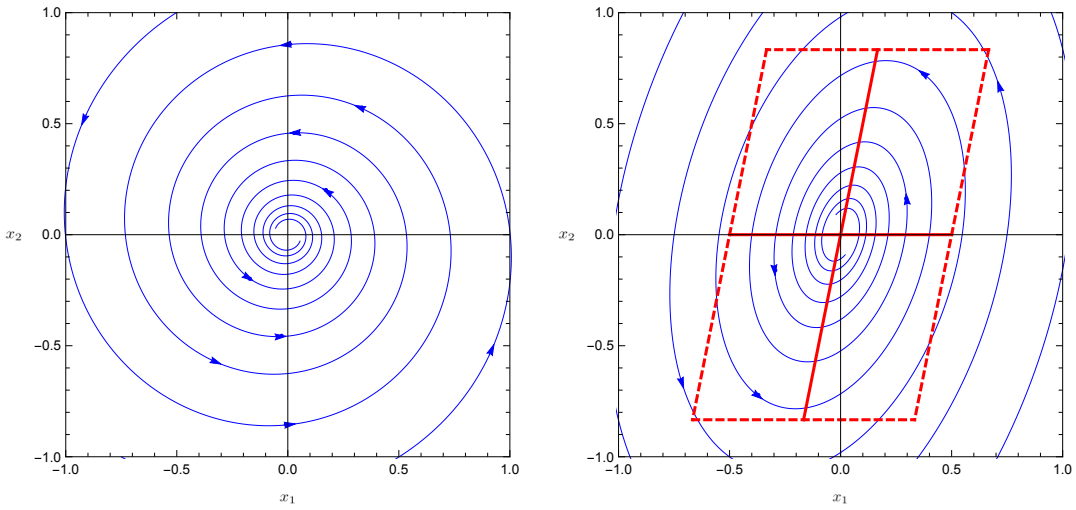
5.5.1.5 Stable spirals We continue thinking about cases with complex eigenvalues, but now we consider eigenvalues with nonzero real part. First we consider the situation where the real part is negative, this being called a ***stable spiral***. First let us consider the prototypical case where

$$A = \begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix},$$

with eigenvalues $\lambda_1 = \sigma + \mathrm{i}\omega$ and $\lambda_2 = \sigma - \mathrm{i}\omega$, where we take $\sigma \in \mathbb{R}_{<0}$. We have, using Procedure 5.2.23,

$$\mathrm{e}^{At} = \mathrm{e}^{\sigma t} \begin{bmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{bmatrix}.$$

The phase portrait in this case we depict in Figure 5.5a, and one can see why the



(a) "Canonical" stable spiral (b) "Noncanonical" stable spiral

Figure 5.5 Stable spirals

name "stable spiral" is applied in this case.

We can also consider a more generic case to illustrate, as in the case of centres, the rôle of the eigenvectors. We take

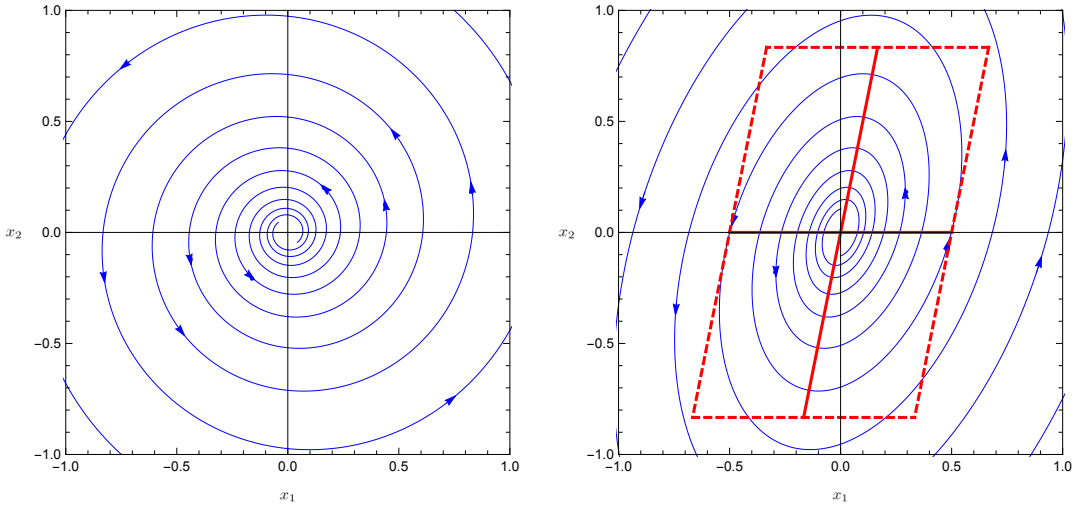
$$A = \begin{bmatrix} \frac{7}{30} & -\frac{2}{3} \\ \frac{5}{3} & -\frac{13}{30} \end{bmatrix},$$

and determine the eigenvalues to be $\lambda_1 = -\frac{1}{10} + \mathrm{i}$ and $\lambda_2 = -\frac{1}{10} - \mathrm{i}$. The eigenvectors are $w_1 = u + \mathrm{i}v$ and $w_2 = u - \mathrm{i}v$, where

$$u = (1, 5), \quad v = (3, 0),$$

i.e., the eigenvectors are the same as for the centre in the previous section. In Figure 5.5b we depict the phase plane in this case, and also overlay the box used to illustrate the rôle of the eigenvectors in the case of a centre.

5.5.1.6 Unstable spirals Next we consider the case where A has complex eigenvalues with positive real part, this being the case of an ***unstable spiral***. The "canonical" case is exactly like that for a stable spiral, except now $\sigma \in \mathbb{R}_{>0}$. The phase portrait in this case is depicted in Figure 5.6a. The situation is the "opposite" of



(a) "Canonical" unstable spiral (b) "Noncanonical" unstable spiral

Figure 5.6 Unstable spirals

that for the stable spiral in Figure 5.5a.

We can also give a more generic case by considering

$$A = \begin{bmatrix} \frac{13}{30} & -\frac{2}{3} \\ \frac{5}{3} & -\frac{7}{30} \end{bmatrix}.$$

In this case, the eigenvalues are $\lambda_1 = \frac{1}{10} + \mathrm{i}$ and $\lambda_2 = \frac{1}{10} - \mathrm{i}$ and the eigenvectors are $w_1 = u + \mathrm{i}v$ and $w_2 = u - \mathrm{i}v$, where

$$u = (1, 5), \quad v = (3, 0).$$

Note that these are the same eigenvalues as for the centre and the stable spiral considered above. In Figure 5.6b we show the phase portrait in this case, along with a box determined by the eigenvectors as in our discussion of the spiral above.

5.5.1.7 Nonisolated equilibrium points The remaining situations we consider are "degenerate" and do not arise as frequently as the preceding cases (although they *do* arise). All of these correspond to cases of a zero eigenvalue. Note that, if one has a zero eigenvalue and if v is any corresponding eigenvector, then any multiple of v is an equilibrium point for the differential equation. Thus, when one is considering cases with zero eigenvalues, the equilibrium point at $(0, 0)$ is not isolated.

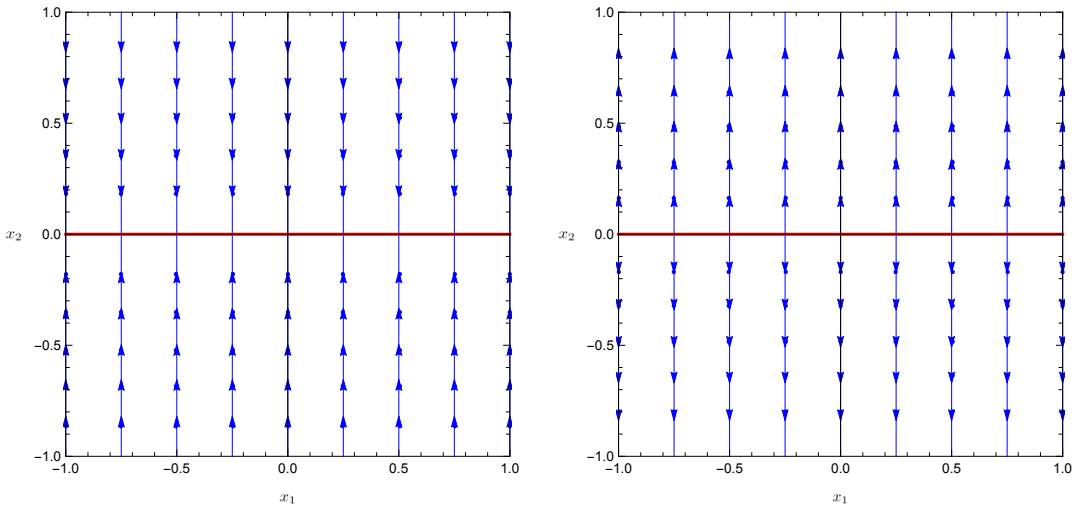
Let us consider the various cases.

Zero eigenvalue with algebraic multiplicity 1

We begin by supposing that A has eigenvalues $\lambda_1 = 0$ and $\lambda_2 = \lambda \neq 0$. In this case, we suppose that

$$A = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}.$$

The behaviour of the solution curves in the phase portrait depends on whether λ is positive or negative. In Figure 5.7a we depict the case when $\lambda \in \mathbb{R}_{<0}$. We see, in this



(a) One zero eigenvalue and one negative eigenvalue

(b) One zero eigenvalue and one positive eigenvalue

Figure 5.7 Zero eigenvalue with algebraic multiplicity 1

case, that the subspace (in red) generated by the eigenvector e_1 for the eigenvalue 0 is populated with equilibria, and that, because λ is negative, all solution curves approach one of these equilibria as $t \to \infty$.

The situation for $\lambda \in \mathbb{R}_{>0}$ is rather similar, and is depicted in Figure 5.7b.

Zero eigenvalue with algebraic multiplicity 2

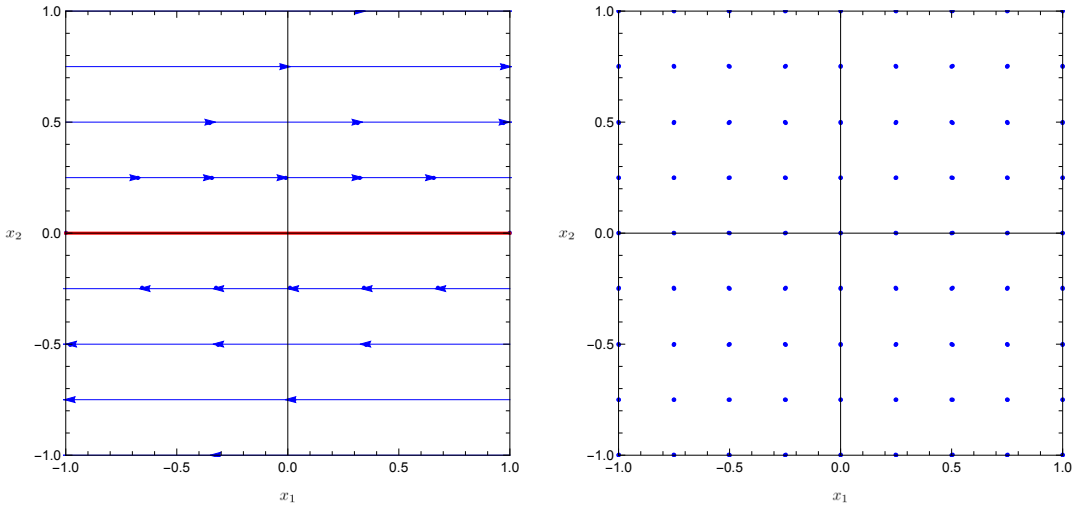
Finally we consider the case of a repeated zero eigenvalue. There are two situations to consider here, one when $m_{\mathrm{g}}(0, \boldsymbol{A}) = 1$ and another when $m_{\mathrm{g}}(0, \boldsymbol{A}) = 2$. In the former case, we consider

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

and in the latter case we must have $\boldsymbol{A} = \boldsymbol{0}$. In the former case we have

$$\mathrm{e}^{\boldsymbol{A}t} = \begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix}$$

and in the latter case we have $\mathrm{e}^{\boldsymbol{0}t} = \boldsymbol{I}_n$. In Figure 5.8a and Figure 5.8b we show



(a) Zero eigenvalue with algebraic multiplicity 2 and geometric multiplicity 1

(b) Zero eigenvalue with algebraic and geometric multiplicity 2

Figure 5.8 Zero eigenvalue with algebraic multiplicity 2

the phase portraits. Of course, the phase portrait in Figure 5.8b is spectacularly uninteresting, since it consists entirely of equilibria!

5.5.2 An introduction to phase portraits for nonlinear systems

The analysis of the preceding section for planar linear ordinary differential equations with constant coefficients was quite comprehensive, exactly because the setting was so simple. Extensions to either higher-dimensions than planar and/or to nonlinear ordinary differential equations are difficult, the former for reasons of difficulty of representation, the latter for reasons of plain ol' difficulty. In this section we consider some *ad hoc* techniques for understanding phase portraits for planar nonlinear ordinary differential equations.

5.5.2.1 Phase portraits near equilibrium points

5.5.2.2 Periodic orbits

5.5.2.3 Attractors

5.5.3 Extension to higher dimensions

5.5.3.1 Behaviour near equilibrium points

5.5.3.2 Attractors

Exercises

5.5.1 For the scalar linear homogeneous ordinary differential equations in $\mathbb{R}^2$ defined by the following 2×2 matrices, do the following:

1. determine what type of planar linear system this is, i.e., "stable node," "unstable node," "saddle point," etc.;
2. sketch the phase portrait, clearly indicating the essential features (knowing what these are is part of the question).

(a) $A = \begin{bmatrix} 2 & -5 \\ 0 & 3 \end{bmatrix}$;

(b) $A = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$;

(c) $A = \begin{bmatrix} -1 & -2 \\ 1 & -3 \end{bmatrix}$;

(d) $A = \begin{bmatrix} 4 & -1 \\ 4 & 0 \end{bmatrix}$;

(e) $A = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$;

(f) $A = \begin{bmatrix} -4 & 6 \\ -1 & 1 \end{bmatrix}$;

(g) $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$;

(h) $A = \begin{bmatrix} 2 & 4 \\ -2 & 6 \end{bmatrix}$;

(i) $A = \begin{bmatrix} -4 & 9 \\ -1 & 2 \end{bmatrix}$.

Section 5.6

Systems of linear homogeneous ordinary difference equations

In this section we undertake a development for difference equations of the treatment in Section 5.2 for differential equations. Thus, following our development of Section 3.3.3.3, we work with a system of linear homogeneous ordinary difference equations F in a finite-dimensional $\mathbb{R}$-vector space X, whose right-hand side, therefore, takes the form

$$\begin{aligned} \widehat{F}\colon \mathbb{T} \times \mathsf{X} &\to \mathsf{X} \\ (t, x) &\mapsto \mathsf{A}(t)(x) \end{aligned} \tag{5.22}$$

for a map $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$. Thus, if $\mathbb{T} \subseteq \mathbb{Z}(h)$, we are looking at difference equations whose solutions $t \mapsto \xi(t)$ satisfy

$$\xi(t + h) = \mathsf{A}(t)(\xi(t)).$$

Our treatment will be structured in the same way as was the treatment in Section 4.6 for scalar equations, to emphasise the similarities between the two theories.

Do I need to read this section? This material is fundamental to the study of linear system theory. •

5.6.1 Equations with time-varying coefficients

We begin by a consideration of general systems with time-varying coefficients, i.e., for which A is not a constant function of time.

5.6.1.1 Solutions and their properties

First let us verify that the basic existence and uniqueness result holds for the difference equations we are considering.

5.6.1 Proposition (Local existence and uniqueness of solutions for systems of linear homogeneous ordinary difference equations) *Consider the system of linear homogeneous ordinary difference equations* F *with right-hand side (5.22). Let* $(\mathrm{t}_0, \mathrm{x}_0) \in \mathbb{T} \times \mathsf{X}$. *Then there exists a unique* $\xi\colon \mathbb{T}_{\geq \mathrm{t}_0} \to \mathsf{X}$ *that is a solution for* F *and which satisfies* $\xi(\mathrm{t}_0) = \mathrm{x}_0$. *If* F *is invertible, then there exists a unique* $\xi\colon \mathbb{T} \to \mathsf{X}$ *that is a solution for* F *and which satisfies the initial conditions.*

Proof Since the state space is $U = \mathsf{X}$, it follows that F is complete and so the first assertion follows from Theorem 3.4.2. The second assertion follows from Theorem 3.4.6. ■

The same sort of comments as given following Proposition 4.6.1 are valid here, in terms of comparing the preceding result with Proposition 5.2.2. In particular, there is this notion of invertibility for difference equations that does not arise for differential equations. Let us clearly enunciate the character of invertibility in the current setting.

5.6.2 Proposition (Invertible systems of linear homogeneous ordinary difference equations) *A system of linear homogeneous ordinary difference equations* F *with right-hand side*

$$\widehat{\mathrm{F}}(\mathrm{t},\mathrm{x}) = \mathsf{A}(\mathrm{t})(\mathrm{x}),$$

for $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X};\mathsf{X})$ *is invertible if and only if* $\det \mathsf{A}(\mathrm{t}) \neq 0$ *for every* $\mathrm{t} \in \mathbb{T}_{\mathrm{F}}$.

Proof This follows immediately from the definition of invertibility in Definition 3.4.5. ■

Now we can discuss the set of all solutions of a system of linear homogeneous ordinary difference equation F with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{T} \times \mathsf{X} &\to \mathsf{X}\\ (t,x) &\mapsto \mathsf{A}(t)(x).\end{aligned}$$

For $t_0 \in \mathbb{T}$, we denote by

$$\mathrm{Sol}_{t_0}(F) = \left\{\xi \in \mathsf{X}^{\mathbb{T}_{\geq t_0}} \;\middle|\; \xi(t+h) = \mathsf{A}(t)(\xi(t)),\ t \in \mathbb{T}_{F,\geq t_0}\right\}$$

the set of solutions for F from t_0. If F is additionally invertible, then we denote by

$$\mathrm{Sol}(F) = \left\{\xi \in \mathsf{X}^{\mathbb{T}} \;\middle|\; \xi(t+h) = \mathsf{A}(t)(\xi(t)),\ t \in \mathbb{T}_{F}\right\}$$

the set of solutions for F. The following result is then the main structural result about the set of solutions to a system of linear homogeneous ordinary difference equations.

5.6.3 Theorem (Vector space structure of sets of solutions) *Consider the system of linear homogeneous ordinary difference equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.22). Then* $\mathrm{Sol}_{\mathrm{t}_0}(\mathrm{F})$ *is an* n*-dimensional subspace of* $\mathsf{X}^{\mathbb{T}_{\geq \mathrm{t}_0}}$. *If* F *is additionally invertible, then* Sol(F) *is an* n*-dimensional subspace of* $\mathsf{X}^{\mathbb{T}}$.

Proof The proof goes like that of Theorem 5.6.3, *mutatis mutandis*. ■

The following corollary, immediate from the proof of the theorem, gives an easy check on the linear independence of subsets of Sol(F).

5.6.4 Corollary (Linear independence in Sol(F)) *Consider the system of linear homogeneous ordinary difference equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.22). Let* $\mathrm{t}_0 \in \mathbb{T}$. *Then the following statements hold:*

(i) a subset $\{\xi_1,\dots,\xi_{\mathrm{k}}\} \subseteq \mathrm{Sol}_{\mathrm{t}_0}(\mathrm{F})$ *is linearly independent if and only if the subset* $\{\xi_1(\mathrm{t}_0),\dots,\xi_{\mathrm{k}}(\mathrm{t}_0)\} \subseteq \mathsf{X}$ *is linearly independent;*

(ii) a subset $\{\xi_1,\dots,\xi_{\mathrm{k}}\} \subseteq \mathrm{Sol}(\mathrm{F})$ *is linearly independent if and only if, for some* $\mathrm{t} \in \mathbb{T}$, *the subset* $\{\xi_1(\mathrm{t}),\dots,\xi_{\mathrm{k}}(\mathrm{t})\} \subseteq \mathsf{X}$ *is linearly independent.*

As with scalar linear homogeneous ordinary difference equations, the theorem allows us to give a special name to bases for $\mathrm{Sol}_{t_0}(F)$ and Sol(F).

5.6.5 Definition (Fundamental set of solutions) Consider the system of linear homogeneous ordinary difference equations F in the n-dimensional $\mathbb{R}$-vector space X with right-hand side (5.22).

(i) A set $\{\xi_1, \dots, \xi_n\}$ of linearly independent elements of $\mathrm{Sol}_{t_0}(F)$ is a ***fundamental set of solutions*** for F.

(ii) If F is invertible, then a set $\{\xi_1, \dots, \xi_n\}$ of linearly independent elements of $\mathrm{Sol}(F)$ is a ***fundamental set of solutions*** for F. •

5.6.1.2 The discrete-time state transition map We now present a particular way of organising a fundamental set of solutions into one object that, for all intents and purposes, completely characterises $\mathrm{Sol}(F)$. This we organise as the following theorem.

5.6.6 Theorem (Existence of, and properties of, the discrete-time state transition map) *Consider the system of linear homogeneous ordinary difference equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.22). Let* $\mathbb{T} \subseteq \mathbb{Z}(\mathrm{h})$ *be a discrete time-domain and let* $\mathrm{t}_0 \in \mathbb{T}$*. Then there exists a unique map* $\Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0} \colon \mathbb{T}_{\geq \mathrm{t}_0} \to \mathrm{L}(\mathsf{X};\mathsf{X})$ *with the following properties:*

(i) the mapping $\Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0}$ *satisfies the initial value problem*

$$\Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0}(\mathrm{t}+\mathrm{h}) = \mathsf{A}(\mathrm{t}) \circ \Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0}(\mathrm{t}), \quad \Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0}(\mathrm{t}_0) = \mathrm{id}_{\mathsf{X}};$$

(ii) the solution to the initial value problem

$$\xi(\mathrm{t}+\mathrm{h}) = \mathsf{A}(\mathrm{t})(\xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathrm{x}_0,$$

is $\mathrm{t} \mapsto \Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0}(\mathrm{t})(\mathrm{x}_0)$*;*

(iii) $\det(\Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0}(\mathrm{t})) = \prod_{\mathrm{j}=0}^{(\mathrm{t}-\mathrm{t}_0-\mathrm{h})/\mathrm{h}} \mathsf{A}(\mathrm{t}_0 + \mathrm{jh})$ *(the* ***Abel–Jacobi–Liouville formula****).*

Moreover, if F *is invertible, then there exists a unique map* $\Phi^{\mathrm{d}}_{\mathsf{A}} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{X};\mathsf{X})$ *with the following properties:*

(iv) for $\mathrm{t}_0 \in \mathbb{T}$*, the mapping* $\mathrm{t} \mapsto \Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t},\mathrm{t}_0)$ *satisfies the initial value problem*

$$\Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t}+\mathrm{h},\mathrm{t}_0) = \mathsf{A}(\mathrm{t})\Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t},\mathrm{t}_0), \quad \Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t}_0,\mathrm{t}_0) = \mathrm{id}_{\mathsf{X}};$$

(v) for $\mathrm{t}_0 \in \mathbb{T}$*, the solution to the initial value problem*

$$\xi(\mathrm{t}+\mathrm{h}) = \mathsf{A}(\mathrm{t})(\xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathrm{x}_0,$$

is $\mathrm{t} \mapsto \Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t},\mathrm{t}_0)(\mathrm{x}_0)$*;*

(vi) for $\mathrm{t},\mathrm{t}_0 \in \mathbb{T}$, $\det(\Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t},\mathrm{t}_0)) = \prod_{\mathrm{j}=0}^{(\mathrm{t}-\mathrm{t}_0-\mathrm{h})/\mathrm{h}} \mathsf{A}(\mathrm{t}_0 + \mathrm{jh})$ *(the* ***Abel–Jacobi–Liouville formula*** *again);*

(vii) for $\mathrm{t},\mathrm{t}_0,\mathrm{t}_1 \in \mathbb{T}$, $\Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t},\mathrm{t}_0) = \Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t},\mathrm{t}_1) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t}_1,\mathrm{t}_0)$*;*

(viii) for each $t, t_0 \in \mathbb{T}$, $\Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0)$ *is invertible and* $\Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0)^{-1} = \Phi^{\mathrm{d}}_{\mathsf{A}}(t_0, t)$.

Proof First of all, we define $\Phi^{\mathrm{d}}_{\mathsf{A},t_0}$ and $\Phi^{\mathrm{d}}_{\mathsf{A}}$ are defined by their satisfying the initial value problems in parts (i) and (iv), respectively. Note that these are initial value problems associated with the system of linear homogeneous ordinary difference equations in $\mathrm{L}(\mathsf{X};\mathsf{X})$, as in the proof of Theorem 5.2.6 for differential equations. This proves the existence and uniqueness and parts (i) and (iv).

(ii) and (v) We compute

$$\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t+h)(x_0) = \mathsf{A}(t) \circ \Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)(x_0)$$

and $\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t_0)(x_0) = x_0$, which shows that $t \mapsto \Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)(x_0)$ solves the initial value problem from part (ii). By uniqueness of such solutions, this gives part (ii) of the theorem. Part (v) follows entirely similarly.

(iii) and (vi) Note that we explicitly have

$$\Phi^{\mathrm{d}}_{\mathsf{L},t_0} = \prod_{j=0}^{(t-t_0-h)/h} \mathsf{A}(t_0 + jh),$$

and so (iii) follows from Proposition I-5.3.3(ii). Part (vi) follows in the same way.

(vii) We compute

$$\Phi^{\mathrm{d}}_{\mathsf{A}}(t+h, t_1) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t_1, t_0)) = \mathsf{A}(t) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t_1, t_0)$$

and

$$\Phi^{\mathrm{d}}_{\mathsf{A}}(t_1, t_1) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t_1, t_0) = \Phi^{\mathrm{d}}_{\mathsf{A}}(t_1, t_0).$$

We also have

$$\Phi^{\mathrm{d}}_{\mathsf{A}}(t+h, t_0) = \mathsf{A}(t) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0).$$

That is to say, both $t \mapsto \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0)$ and $t \mapsto \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_1) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t_1, t_0)$ satisfy the initial problem

$$\Phi(t+h) = \mathsf{A}(t) \circ \Phi(t), \quad \Phi(t_1) = \Phi^{\mathrm{d}}_{\mathsf{A}}(t_1, t_0).$$

By uniqueness of solutions for systems of linear homogeneous ordinary differential equations, we conclude that $\Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0) = \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_1) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t_1, t_0)$, as desired.

(viii) The invertibility of $\Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0)$ follows from part (vi) and Theorem I-5.3.10. The specific formula for the inverse follows from the formula

$$\mathrm{id}_{\mathsf{X}} = \Phi^{\mathrm{d}}_{\mathsf{A}}(t_0, t_0) = \Phi^{\mathrm{d}}_{\mathsf{A}}(t_0, t) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0),$$

which itself follows from part (vii). ■

Let us formally name the mappings defined in the theorem.

5.6.7 Definition (Discrete-time state transition map) Consider the system of linear homogeneous ordinary differential equations F in the n-dimensional $\mathbb{R}$-vector space X with right-hand side (5.22).

(i) For $t_0 \in \mathbb{T}$, the map $\Phi^{\mathrm{d}}_{\mathsf{A}} \colon \mathbb{T}_{\geq t_0} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ from Theorem 5.6.6(i) is the ***discrete-time state transition map from*** $\mathbf{t_0}$.

(ii) If F is invertible, then the map $\Phi^{\mathrm{d}}_{\mathsf{A}} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ from Theorem 5.6.6(iv) is the ***discrete-time state transition map***. •

Of course, if F is invertible, then $\Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0) = \Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)$.

One imagines that it is possible to compute the discrete-time state transition map if one is given a fundamental set of solutions. The following procedure gives an explicit means of doing this.

5.6.8 Procedure (Determining the discrete-time state transition map from a fundamental set of solutions) Given a system of linear homogeneous ordinary difference equations F in the n-dimensional $\mathbb{R}$-vector space X with right-hand side equation

$$\widehat{F}(t, x) = \mathsf{A}(t)(x),$$

given $t_0 \in \mathbb{T}$, and given a fundamental set of solutions $\{\xi_1, \dots, \xi_n\}$ from t_0, do the following.

1. Choose a basis $\{e_1, \dots, e_n\}$.
2. Let $\boldsymbol{\xi}_j \colon \mathbb{T} \to \mathbb{R}^n$ be the components of ξ_j, $j \in \{1, \dots, n\}$, i.e.,

$$\xi_j(t) = \xi_{1,j}(t)e_1 + \cdots + \xi_{j,n}(t)e_n.$$

   If $\mathsf{X} = \mathbb{R}^n$, one can just take the components of ξ_j, $j \in \{1, \dots, n\}$, in the standard basis, as usual.
3. Assemble the matrix function $\Xi \colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ by making the components of $\boldsymbol{\xi}_1(t), \dots, \boldsymbol{\xi}_j(t)$ the columns of $\Xi(t)$:

$$\Xi(t) = \begin{bmatrix} \xi_{1,1}(t) & \xi_{2,1}(t) & \cdots & \xi_{n,1}(t) \\ \xi_{1,2}(t) & \xi_{2,2}(t) & \cdots & \xi_{n,2}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{1,n}(t) & \xi_{2,n}(t) & \cdots & \xi_{n,n}(t) \end{bmatrix}.$$

   (Be sure you understand that $\xi_{j,k}(t)$ is the kth component of $\boldsymbol{\xi}_j(t)$.) We call the matrix-valued function $\Xi \colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ a ***fundamental matrix*** for F.
4. Define $\boldsymbol{\Phi}_{t_0}(t) = \Xi(t)\Xi(t_0)^{-1}$.
5. Then $\boldsymbol{\Phi}_{t_0}(t)$ is the matrix representative of $\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)$ in the basis $\{e_1, \dots, e_n\}$.

If, additionally, F is invertible, then, given a fundamental set of solutions $\{\xi_1, \dots, \xi_n\}$, we apply the above procedure to give

$$\boldsymbol{\Phi}(t, t_0) = \Xi(t)\Xi(t_0)^{-1}$$

as the matrix representative for $\Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0)$, $(t, t_0) \in \mathbb{T} \times \mathbb{T}$. •

Let us verify that the preceding procedure does indeed yield the discrete-time state transition map.

5.6.9 Proposition (Determining the discrete-time state transition map from a fundamental set of solutions) *Consider the system of linear homogeneous ordinary difference equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.22). Then Procedure 5.6.8 will produce the discrete-time state transition map from* t_0 *and will, in the case that* F *is invertible, produce the discrete-time state transition map.*

Proof By choosing a basis $\{e_1, \dots, e_n\}$ as in Procedure 5.6.8, we can assume that $\mathsf{X} = \mathbb{R}^n$. (This is legitimate by virtue of Exercises 5.6.1 and 5.6.2.) Let us denote by $\boldsymbol{A}(t)$ the matrix representative of $\mathsf{A}(t)$. Defining $\boldsymbol{\Phi}_{t_0}(t)$ as in the given procedure, we have

$$\boldsymbol{\Phi}_{t_0}(t+h) = \boldsymbol{\Xi}(t+h)\boldsymbol{\Xi}(t_0)^{-1}.$$

Noting that each of ξ_j, $j \in \{1, \dots, n\}$, is a solution for F, we have

$$\xi_{j,k}(t+h) = \sum_{l=1}^{n} A^k_l(t)\xi_{j,l}(t), \qquad j \in \{1, \dots, n\},\ t \in \mathbb{T}_{\geq t_0}.$$

Therefore, in matrix notation,

$$\begin{aligned} &\left[\ \xi_1(t+h) \ \middle|\ \cdots \ \middle|\ \xi(t+h)\ \right] = \boldsymbol{A}(t)\left[\ \xi_1(t) \ \middle|\ \cdots \ \middle|\ \xi(t)\ \right] \\ \implies\ &\boldsymbol{\Xi}(t+h) = \boldsymbol{A}(t)\boldsymbol{\Xi}(t), \quad t \in \mathbb{T}_{\geq t_0}. \end{aligned}$$

Therefore,

$$\boldsymbol{\Phi}_{t_0}(t) = \boldsymbol{A}(t)\boldsymbol{\Xi}(t)\boldsymbol{\Xi}(t_0)^{-1} = \boldsymbol{A}(t)\boldsymbol{\Phi}(t, t_0).$$

Moreover, $\boldsymbol{\Phi}(t_0, t_0) = \boldsymbol{I}_n$. Thus $t \mapsto \boldsymbol{\Phi}(t, t_0)$ satisfies the matrix representative of the initial value problem satisfied by $t \mapsto \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0)$, i.e., $\boldsymbol{\Phi}(t, t_0)$ is the matrix representative of $\Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0)$. ■

We observe that the computation of the discrete-time state transition map is at once trivial and difficult to get satisfactory form for. To the first of these attributes, we simply have

$$\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t) = \prod_{j=0}^{(t-t_0-h)/h} \mathsf{A}(t_0 + jh). \tag{5.23}$$

To the second of these attributes, one does not typically have a nice "closed form" expression for $\Phi^{\mathrm{d}}_{\mathsf{A},t_0}$. We shall see shortly that such a nice formula can be obtained in the constant coefficient case.

5.6.1.3 The adjoint equation In this section we consider a difference equation "dual" to a system of linear homogeneous ordinary difference equations, just as we did in Section 5.2.1.4 for differential equations.

5.6.10 Definition (Adjoint of a system of linear homogeneous ordinary difference equations) Consider the system of linear homogeneous ordinary difference equations F in the n-dimensional $\mathbb{R}$-vector space X with right-hand side (5.22). Suppose that F is invertible. The ***adjoint equation*** for F is the system F^* of linear homogeneous ordinary difference equations in X^* with right-hand side

$$\begin{aligned}\widehat{F^*}\colon \mathbb{T} \times \mathsf{X}^* &\to \mathsf{X}^*\\ (t, p) &\mapsto (\mathsf{A}^{-1})^*(t)(p).\end{aligned}$$

•

Thus solutions $t \mapsto p(t)$ for the adjoint equation satisfy

$$p(t + h) = (\mathsf{A}^{-1})^*(t)(p(t)).$$

Let us give the discrete-time state transition map for the adjoint equation.

5.6.11 Proposition (Discrete-time state transition map for the adjoint equation) *Consider the system of linear homogeneous ordinary difference equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.22). Suppose that* F *is invertible. Then the discrete-time state transition map for the adjoint equation is defined by* $\Phi^{\mathrm{d}}_{(\mathsf{A}^{-1})^*}(\mathrm{t}, \mathrm{t}_0) = \Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t}_0, \mathrm{t})^*$ *for* $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}$.

Proof This is obvious, given the general formula (5.23) for the discrete-time state transition map, along with the property Proposition I-5.7.20(ii) of the dual of a composition and the property Proposition I-4.1.6(iv) of the inverse of a composition. ■

We have not yet addressed the important question, "Why should one care about the adjoint equation?" We convert this question into another question with the following result.

5.6.12 Proposition (A property of the adjoint equation) *Consider the system of linear homogeneous ordinary difference equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.22). Suppose that* F *is invertible. Let* $\mathrm{t}_0 \in \mathbb{T}$, $\mathrm{x}_0 \in \mathsf{X}$, *and* $\mathrm{p}_0 \in \mathsf{X}^*$, *and denote* $\mathrm{x}(\mathrm{t}) = \Phi^{\mathrm{d}}_{\mathsf{A}}(\mathrm{t}, \mathrm{t}_0)(\mathrm{x}_0)$ *and* $\mathrm{p}(\mathrm{t}) = (\Phi^{\mathrm{d}}_{\mathsf{A}})^*(\mathrm{t}_0, \mathrm{t})(\mathrm{p}_0)$. *Then*

$$\langle \mathrm{p}(\mathrm{t}); \mathrm{x}(\mathrm{t})\rangle = \langle \mathrm{p}_0; \mathrm{x}_0\rangle.$$

Proof We have

$$\begin{aligned}\Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0) &= \mathsf{A}(t - h) \cdots \mathsf{A}(t_0 + h) \circ \mathsf{A}(t_0),\\ (\Phi^{\mathrm{d}}_{\mathsf{A}}(t_0, t))^* &= (\mathsf{A}(t - h)^{-1})^* \cdots (\mathsf{A}(t_0 + h)^{-1})^* \circ (\mathsf{A}(t_0)^{-1})^*.\end{aligned}$$

The result now follows by the definition of the dual of a linear map. ■

When $\alpha \in \mathsf{X}^*$ and $v \in \mathsf{X}$ satisfy $\alpha(v) = 0$, we say that α ***annihilates*** v. This is a sort of "orthogonality condition," although it most definitely is not an actual orthogonality condition, there being no inner product in sight. One of the upshots of the preceding result is the following corollary, saying that the adjoint equation preserves the annihilation condition.

5.6.13 Corollary (The geometric meaning of the adjoint equation) *Consider the system of linear homogeneous ordinary difference equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (5.22). Let* $t_0 \in \mathbb{T}$, $x_0 \in \mathsf{X}$, *and* $p_0 \in \mathsf{X}^*$, *and denote* $x(t) = \Phi_{\mathsf{A}}^{\mathrm{d}}(t, t_0)(x_0)$ *and* $p(t) = (\Phi_{\mathsf{A}}^{\mathrm{d}})^*(t_0, t)(p_0)$. *If* $\langle p_0; x_0 \rangle = 0$, *then* $\langle p(t); x(t) \rangle = 0$ *for all* $t \in \mathbb{T}$.

It is this property of the adjoint equation that makes it an important tool in optimal control theory, but this is not a subject into which we shall dwell deeply here.

5.6.2 Equations with constant coefficients

We now consider the special case of systems of linear homogeneous equations with constant coefficients, i.e., those systems of linear ordinary difference equations F in a vector space X with right-hand sides

$$\widehat{F}(t, x) = \mathsf{A}(x), \tag{5.24}$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$. As with the scalar version of such equations that we studied in Section 4.6.2, there is a great deal more that we can say about such equations, beyond the general assertions in the preceding section. Indeed, one can say that, in principle, one can "solve" such equations, and we shall present a procedure for doing so.

Before we do so, however, we reiterate that the ordinary difference equations we are considering in this section are special cases of the time-varying equations of the preceding section, so all of the general statements made there apply here as well. In particular, Proposition 5.6.1 and Theorem 5.6.3 hold for equations of the form (5.24).

As with linear ordinary differential equations with constant coefficients as treated in Section 5.2.2, linear algebra plays a rôle in the theory of systems of linear homogeneous ordinary difference equations. The background material required for this was developed comprehensively in Sections I-5.4.9, I-5.4.10, and I-5.8.10, among other places. We refer to the beginning of Section 5.2.2 for a summary of the facts from linear algebra to which we shall make reference, as these are the same for difference equations as for differential equations.

5.6.2.1 Complexification of systems of linear ordinary difference equations

In Section 4.6.2.1 we complexified a scalar linear homogeneous ordinary difference equation with constant coefficients. The reason we had to do so was that the characteristic polynomial for such an equation will generally have complex roots, and these complex roots lead naturally to complex solutions of the difference equation. It is only after taking real and imaginary parts of a complex solution that we recover the real solutions. The same sort of thing happens with systems of linear homogeneous ordinary difference equations with constant coefficients. In this case, the issue that arises is that one will generally have complex eigenvalues.

The process of complexification is an easy one, and requires no words like "everything we have done in the real case also works in the complex case," since we are working with systems defined on abstract $\mathbb{R}$-vector spaces, and $\mathsf{X}_{\mathbb{C}}$ is certainly a $\mathbb{R}$-vector space.

5.6.14 Definition (Complexification of a system of linear ordinary difference equations) Consider the system of linear homogeneous ordinary difference equations F with constant coefficients and with right-hand side (5.24). The ***complexification*** of F is the system of linear homogeneous ordinary difference equations $F_{\mathbb{C}}$ with constant coefficients given by

$$\begin{aligned} F_{\mathbb{C}} \colon \mathbb{T} \times \mathsf{X}_{\mathbb{C}} \times \mathsf{X}_{\mathbb{C}} &\to \mathsf{X}_{\mathbb{C}} \\ (t, z, w) &\mapsto w - \mathsf{A}_{\mathbb{C}}(z). \end{aligned}$$

•

A ***solution*** for $F_{\mathbb{C}}$ is a locally absolutely continuous map $\zeta \colon \mathbb{T} \to \mathsf{X}_{\mathbb{C}}$ that satisfies

$$\dot{\zeta}(t) = \mathsf{A}_{\mathbb{C}}(\zeta(t)).$$

Note that, as $\mathsf{X}_{\mathbb{C}} = \mathsf{X} \times \mathsf{X}$, we can write $\zeta(t) = (\xi(t), \eta(y))$ for locally absolutely continuous maps $\xi, \eta \colon \mathbb{T} \to \mathsf{X}$ that are the ***real part*** and ***imaginary part*** of ζ, respectively.

As in the scalar case, the real and imaginary parts of a solution separately satisfy the uncomplexified differential equation.

5.6.15 Lemma (Real and imaginary parts of complex solutions are solutions) *Consider the system of linear homogeneous ordinary difference equations* F *with constant coefficients, with right-hand side (5.24) and with complexification* $\mathrm{F}_{\mathbb{C}}$. *If* $\zeta \colon \mathbb{T} \to \mathsf{X}_{\mathbb{C}}$ *is a solution for* $\mathrm{F}_{\mathbb{C}}$, *then* $\mathrm{Re}(\zeta)$ *and* $\mathrm{Im}(\zeta)$ *are solutions for* F.

Proof Given $\zeta \colon \mathbb{T} \to \mathsf{X}_{\mathbb{C}}$ we write $\zeta(t) = (\xi(t), \eta(t))$ so that $\xi = \mathrm{Re}(\zeta)$ and $\eta = \mathrm{Im}(\zeta)$. Since ζ is a solution for $F_{\mathbb{C}}$, we have

$$\zeta(t + h) = (\xi(t + h), \eta(t + h)) = \mathsf{A}_{\mathbb{C}}(\zeta(t)) = (\mathsf{A}(\xi(t)), \mathsf{A}(\eta(t))$$

by definition of $\mathsf{A}_{\mathbb{C}}$. Equating the second and fourth terms in this string of equalities gives the lemma. ■

5.6.2.2 The operator power function

In this section we consider the constant coefficient version of the discrete-time state transition map.

5.6.16 Definition (Operator power function) Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, X be a finite-dimensional $\mathbb{F}$-vector space, and let $\mathsf{L} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$. The ***operator power function*** of L is the map $\mathsf{P}_{\mathsf{L}} \colon \mathbb{Z}_{\geq 0} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ defined by $\mathsf{P}_{\mathsf{L}}(j) = \mathsf{L}^j$. If L is invertible, then we can define $\mathsf{P}_{\mathsf{L}} \colon \mathbb{Z} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ by $\mathsf{P}_{\mathsf{L}}(j) = \mathsf{L}^j$. •

The operator power function we consider here is the analogue of the operator exponential from Definition 5.2.19, but for difference equations. We defined the

operator exponential as the solution of a differential equation, and one can similarly define the operator power function to be the solution to a difference equation. Indeed, if we consider the system of linear ordinary difference equations F with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{Z}\times \mathsf{X} &\to \mathsf{X}\\ (j,x) &\mapsto \mathsf{L}(x),\end{aligned}$$

then $\mathsf{P}_{\mathsf{L}}(j) = \Phi^{\mathrm{d}}_{\mathsf{L},0}(j)$, $j \in \mathbb{Z}_{\geq 0}$. When L is invertible, then we can write this as $\mathsf{P}_{\mathsf{L}}(j) = \Phi^{\mathrm{d}}_{\mathsf{L}}(j,0)$ for $j \in \mathbb{Z}$.

Let us give some alternative characterisations and properties of the operator exponential.

5.6.17 Theorem (Properties of the operator power function) *Let $\mathbb{F} \in \{\mathbb{R},\mathbb{C}\}$, let X be a finite-dimensional $\mathbb{F}$-vector space, and let $\mathsf{L},\mathsf{M} \in \mathrm{L}(\mathsf{X};\mathsf{X})$. Then the following statements hold:*

(i) *if $\mathbb{F} = \mathbb{C}$, then P_{L} is a $\mathbb{C}$-linear map;*

(ii) $\mathsf{P}_{\mathsf{L}}(\mathrm{j}+1) = \mathsf{L}\circ\mathsf{P}_{\mathsf{L}}(\mathrm{j}) = \mathsf{P}_{\mathsf{L}}(\mathrm{j})\circ\mathsf{L}$*;*

(iii) $\mathsf{P}_{\mathsf{L}}(0) = \mathrm{id}_{\mathsf{X}}$*;*

(iv) *for $\alpha \in \mathbb{F}$, $\alpha^{\mathrm{j}}\,\mathrm{id}_{\mathsf{X}} = \mathsf{P}_{\alpha\,\mathrm{id}_{\mathsf{X}}}(\mathrm{j})$;*

(v) $\mathsf{P}_{\mathsf{L}}(\mathrm{j})\circ\mathsf{P}_{\mathsf{M}}(\mathrm{k}) = \mathsf{P}_{\mathsf{M}}(\mathrm{j})\circ\mathsf{P}_{\mathsf{L}}(\mathrm{k})$ *for all $\mathrm{j},\mathrm{k} \in \mathbb{Z}_{\geq 0}$ if and only if $\mathsf{L}\circ\mathsf{M} = \mathsf{M}\circ\mathsf{L}$;*

(vi) *if L is invertible then $\mathsf{P}_{\mathsf{L}}(\mathrm{j})$ is invertible for every $\mathrm{j} \in \mathbb{Z}$, and $(\mathsf{P}_{\mathsf{L}})^{-1} = \mathsf{P}_{\mathsf{L}^{-1}}$;*

(vii) *if $\mathsf{U} \subseteq \mathsf{X}$ is L-invariant, then it is also $\mathsf{P}_{\mathsf{L}}(\mathrm{j})$-invariant for every $\mathrm{j} \in \mathbb{Z}$;*

(viii) *the solution to the initial value problem with time-domain $\mathbb{Z}(\mathrm{h})$ given by*

$$\xi(\mathrm{t}+\mathrm{h}) = \mathsf{L}(\xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathrm{x}_0,$$

is $\xi(\mathrm{t}) = \mathsf{P}_{\mathsf{L}}\left(\frac{\mathrm{t}-\mathrm{t}_0}{\mathrm{h}}\right)(\mathrm{x}_0)$, $\mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0}$ (or $\mathrm{t} \in \mathbb{T}$ if L is invertible).

Proof (i) This is obvious since the composition of $\mathbb{C}$-linear maps is a $\mathbb{C}$-linear map.

(ii) This is obvious.

(iii) This follows by the convention that we take $\mathsf{L}^0 = \mathrm{id}_{\mathsf{X}}$.

(iv) This is obvious.

(v) If $\mathsf{L}\circ\mathsf{M} = \mathsf{M}\circ\mathsf{L}$, then one can freely permute the order of the terms to give

$$\mathsf{P}_{\mathsf{L}}(j)\circ\mathsf{P}_{\mathsf{M}}(k) = \underbrace{\mathsf{L}\circ\cdots\circ\mathsf{L}}_{j \text{ times}}\circ\underbrace{\mathsf{M}\circ\cdots\circ\mathsf{M}}_{k \text{ times}} = \underbrace{\mathsf{M}\circ\cdots\circ\mathsf{M}}_{k \text{ times}}\circ\underbrace{\mathsf{L}\circ\cdots\circ\mathsf{L}}_{j \text{ times}} = \mathsf{P}_{\mathsf{M}}(k)\circ\mathsf{P}_{\mathsf{L}}(j)$$

for every $j \in \mathbb{Z}_{\geq 0}$. Conversely, if

$$\mathsf{P}_{\mathsf{L}}(j)\circ\mathsf{P}_{\mathsf{M}}(k) = \mathsf{P}_{\mathsf{M}}(k)\circ\mathsf{P}_{\mathsf{L}}(j), \qquad j,k \in \mathbb{Z}_{\geq 0},$$

then we have $\mathsf{L}\circ\mathsf{M} = \mathsf{M}\circ\mathsf{L}$, taking $j = k = 1$.

(vi) This follows from Proposition I-5.1.24(iii).

(vii) This was proved during the first part of the proof of Theorem 5.2.20(viii).

(viii) This follows immediately from (5.23) and Theorem 5.6.6(ii) (or from part (v) of the same theorem, if L is invertible). ■

Let us consider the representation of the operator exponential in a basis.

5.6.18 Proposition (The matrix representation of the operator power function is the operator power function of the matrix representation) *Let* $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$*, let* X *be an* n*-dimensional* $\mathbb{F}$*-vector space, let* $\mathsf{L} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$*, and let* $\mathscr{B} = \{e_1, \dots, e_n\}$ *be a basis for* X*. Then*

$$[\mathsf{P}_{\mathsf{L}}(j)]^{\mathscr{B}}_{\mathscr{B}} = \mathsf{P}_{[\mathsf{L}]^{\mathscr{B}}_{\mathscr{B}}}(j).$$

Proof This follows from the definition of the operator power function and Exercise 5.6.2. ■

5.6.2.3 Bases of solutions

Now, for equations with constant coefficients, we construct "explicitly" a basis for $\mathrm{Sol}_{t_0}(F)$.

5.6.19 Procedure (Basis of solutions for a system of linear homogeneous ordinary difference equations with constant coefficients) Given a system of linear homogeneous ordinary difference equations

$$F\colon \mathbb{T} \times \mathsf{X} \oplus \mathsf{X} \to \mathsf{X}$$

with time domain $\mathbb{T} \subseteq \mathbb{Z}(h)$ in an n-dimensional $\mathbb{R}$-vector space X, with right-hand side

$$\widehat{F}(t, x) = \mathsf{A}(x),$$

and with $t_0 \in \mathbb{T}$, do the following.

1. Choose a basis $\{e_1, \dots, e_n\}$ for X. Let $\boldsymbol{A}$ be the matrix representative of A with respect to this basis. If $\mathsf{X} = \mathbb{R}^n$, one can just take $\boldsymbol{A}$ to be the usual matrix associated with $\mathsf{A} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$.
2. Compute the characteristic polynomial $P_{\mathsf{A}} = \det(X\boldsymbol{I}_n - \boldsymbol{A})$.
3. Compute the roots of P_{A}, i.e., the eigenvalues of $\mathsf{A}_{\mathbb{C}}$, and organise them as follows. We have a zero eigenvalue, we have distinct nonzero real eigenvalues

   $$\ell_1, \dots, \ell_r,$$

   and distinct complex eigenvalues

   $$\lambda_1 = \rho_1 \mathrm{e}^{\mathrm{i}\theta_1}, \dots, \lambda_s = \rho_s \mathrm{e}^{\mathrm{i}\theta_s},$$

   $\rho_1, \dots, \rho_s \in \mathbb{R}_{>0}$, $\theta_1, \dots, \theta_s \in (0, \pi)$, along with their complex conjugates.
4. Let $m_0 = m_{\mathrm{a}}(0, \mathsf{A})$, $m_j = m_{\mathrm{a}}(\ell_j, \mathsf{A})$, $j \in \{1, \dots, r\}$, and $\mu_j = m_{\mathrm{a}}(\lambda_j, \mathsf{A})$, $j \in \{1, \dots, s\}$, be the algebraic multiplicities.
5. Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{m_0}$ be a basis for

   $$\overline{\mathsf{W}}(0, \boldsymbol{A}) = \ker(\boldsymbol{A}^{m_0}).$$

6. For $j \in \{1, \dots, r\}$, let $\{\boldsymbol{x}_{j,1}, \dots, \boldsymbol{x}_{j,m_j}\}$ be a basis for

   $$\overline{\mathsf{W}}(\ell_j, \boldsymbol{A}) = \ker((\ell_j \boldsymbol{I}_n - \boldsymbol{A})^{m_j}).$$

7. For $j \in \{1, \dots, s\}$, let $\{\boldsymbol{z}_{j,1}, \dots, \boldsymbol{z}_{j,\mu_j}\}$ be a basis for

$$\overline{\mathsf{W}}(\lambda_j, \boldsymbol{A}_{\mathbb{C}}) = \ker((\lambda_j \boldsymbol{I}_n - \boldsymbol{A}_{\mathbb{C}})^{\mu_j}).$$

   Write $\boldsymbol{z}_{j,k} = \boldsymbol{a}_{j,k} + \mathrm{i}\boldsymbol{b}_{j,k}$ for each $k \in \{1, \dots, \mu_j\}$. Then

$$\{\boldsymbol{a}_{j,1}, \boldsymbol{b}_{j,1}, \dots, \boldsymbol{a}_{j,\mu_j}, \boldsymbol{b}_{j,\mu_j}\}$$

   is a basis for $\overline{\mathsf{W}}(\lambda_j, \boldsymbol{A})$.

8. For $k \in \{1, \dots, m_0\}$, define

$$\boldsymbol{\eta}_k(t) = \boldsymbol{A}^{(t-t_0)/h} \boldsymbol{y}_k.$$

9. For $j \in \{1, \dots, r\}$ and $k \in \{1, \dots, m_j\}$, define

$$\boldsymbol{\xi}_{j,k}(t) = \ell_j^{t/h} \sum_{m=0}^{\min\{t/h, m_j - 1\}} \binom{t/h}{m} (\ell_j^{-1} \boldsymbol{A} - \boldsymbol{I}_n)^m \boldsymbol{x}_{j,k}.$$

10. For $j \in \{1, \dots, s\}$ and $k \in \{1, \dots, \mu_j\}$, define

$$\begin{aligned}
&\boldsymbol{\alpha}_{j,k}(t) \\
&= \rho_j^{t/h} \Bigg(\Bigg(\sum_{m=0}^{\min\{t/h, \mu_j - 1\}} \binom{t/h}{m} \rho_j^{-m} \cos(m\theta_j) \sum_{l=0}^{\lceil \frac{m-1}{2} \rceil} \binom{m}{2l} (-1)^l \rho_j^{2l} \sin(\theta_j)^{2l} (\boldsymbol{A} - \rho_j \cos(\theta_j) \boldsymbol{I}_n)^{m-2l} \\
&\quad + \sum_{m=0}^{\min\{t/h, \mu_j - 1\}} \binom{t/h}{m} \rho_j^{-m} \sin(m\theta_j) \sum_{l=0}^{\lfloor \frac{m-1}{2} \rfloor} \binom{m}{2l+1} (-1)^{l+1} \rho_j^{2l+1} \sin(\theta_j)^{2l+1} (\boldsymbol{A} - \rho_j \cos(\theta_j) \boldsymbol{I}_n)^{m-2l-1} \Bigg) \\
&\qquad \times (\cos(\theta_j \tfrac{t}{h}) \boldsymbol{a}_{j,k} - \sin(\theta_j \tfrac{t}{h}) \boldsymbol{b}_{j,k}) \\
&\quad - \Bigg(\sum_{m=0}^{\min\{t/h, \mu_j - 1\}} \binom{t/h}{m} \rho_j^{-m} \cos(m\theta_j) \sum_{l=0}^{\lfloor \frac{m-1}{2} \rfloor} \binom{m}{2l+1} (-1)^{l+1} \rho_j^{2l+1} \sin(\theta_j)^{2l+1} (\boldsymbol{A} - \rho_j \cos(\theta_j) \boldsymbol{I}_n)^{m-2l-1} \\
&\qquad + \sum_{m=0}^{\min\{t/h, \mu_j - 1\}} \binom{t/h}{m} \rho_j^{-m} \sin(m\theta_j) \sum_{l=0}^{\lceil \frac{m-1}{2} \rceil} \binom{m}{2l} (-1)^l \rho_j^{2l} \sin(\theta_j)^{2l} (\boldsymbol{A} - \rho_j \cos(\theta_j) \boldsymbol{I}_n)^{m-2l} \Bigg) \\
&\qquad \times (\cos(\theta_j \tfrac{t}{h}) \boldsymbol{b}_{j,k} + \sin(\theta_j \tfrac{t}{h}) \boldsymbol{a}_{j,k}) \Bigg) \qquad (5.25)
\end{aligned}$$

and

$$\begin{aligned}
&\boldsymbol{\beta}_{j,k}(t)\\
&= \rho_j^{t/h}\Bigg(\Bigg(\sum_{m=0}^{\min\{t/h,\mu_j-1\}}\binom{t/h}{m}\rho_j^{-m}\cos(m\theta_j)\sum_{l=0}^{\lceil\frac{m-1}{2}\rceil}\binom{m}{2l}(-1)^l\rho_j^{2l}\sin(\theta_j)^{2l}(\boldsymbol{A}-\rho_j\cos(\theta_j)\boldsymbol{I}_n)^{m-2l}\\
&+\sum_{m=0}^{\min\{t/h,\mu_j-1\}}\binom{t/h}{m}\rho_j^{-m}\sin(m\theta_j)\sum_{l=0}^{\lfloor\frac{m-1}{2}\rfloor}\binom{m}{2l+1}(-1)^{l+1}\rho_j^{2l+1}\sin(\theta_j)^{2l+1}(\boldsymbol{A}-\rho_j\cos(\theta_j)\boldsymbol{I}_n)^{m-2l-1}\Bigg)\\
&\qquad\times(\cos(\theta_j\tfrac{t}{h})\boldsymbol{b}_{j,k}+\sin(\theta_j\tfrac{t}{h})\boldsymbol{a}_{j,k})\\
&+\Bigg(\sum_{m=0}^{\min\{t/h,\mu_j-1\}}\binom{t/h}{m}\rho_j^{-m}\cos(m\theta_j)\sum_{l=0}^{\lfloor\frac{m-1}{2}\rfloor}\binom{m}{2l+1}(-1)^{l+1}\rho_j^{2l+1}\sin(\theta_j)^{2l+1}(\boldsymbol{A}-\rho_j\cos(\theta_j)\boldsymbol{I}_n)^{m-2l-1}\\
&-\sum_{m=0}^{\min\{t/h,\mu_j-1\}}\binom{t/h}{m}\rho_j^{-m}\sin(m\theta_j)\sum_{l=0}^{\lceil\frac{m-1}{2}\rceil}\binom{m}{2l}(-1)^l\rho_j^{2l}\sin(\theta_j)^{2l}(\boldsymbol{A}-\rho_j\cos(\theta_j)\boldsymbol{I}_n)^{m-2l}\Bigg)\\
&\qquad\times(\cos(\theta_j\tfrac{t}{h})\boldsymbol{a}_{j,k}-\sin(\theta_j\tfrac{t}{h})\boldsymbol{b}_{j,k})\Bigg), \quad (5.26)
\end{aligned}$$

where, for $x \in \mathbb{R}$, $\lfloor x \rfloor$ is greatest integer less than or equal to x and $\lceil x \rceil$ is smallest integer greater than or equal to x.

11. For $k \in \{1, \ldots, m_0\}$, let $\eta_k \colon \mathbb{T} \to \mathsf{X}$ be the function whose components with respect to the basis $\{e_1, \ldots, e_n\}$ are the components of $\boldsymbol{\eta}_k$.
12. For $j \in \{1, \ldots, r\}$ and $k \in \{1, \ldots, m_j\}$, let $\xi_{j,k} \colon \mathbb{T} \to \mathsf{X}$ be the function whose components with respect to the basis $\{e_1, \ldots, e_n\}$ are the components of $\boldsymbol{\xi}_{j,k}$.
13. For $j \in \{1, \ldots, s\}$ and $k \in \{1, \ldots, \mu_j\}$, let $\alpha_{j,k}, \beta_{j,k} \colon \mathbb{T} \to \mathsf{X}$ be the functions whose components with respect to the basis $\{e_1, \ldots, e_n\}$ are the components of $\boldsymbol{\alpha}_{j,k}$ and $\boldsymbol{\beta}_{j,k}$, respectively.
14. Then the n functions

$$\begin{aligned}
\eta_k, &\qquad k \in \{1, \ldots, m_0\},\\
\xi_{j,k}, &\qquad j \in \{1, \ldots, r\},\ k \in \{1, \ldots, m_j\},\\
\alpha_{j,k}, \beta_{j,k}, &\qquad j \in \{1, \ldots, s\},\ k \in \{1, \ldots, \mu_j\},
\end{aligned}$$

are a basis for $\mathrm{Sol}_{t_0}(F)$ and, if A is invertible, then the n functions

$$\begin{aligned}
\xi_{j,k}, &\qquad j \in \{1, \ldots, r\},\ k \in \{1, \ldots, m_j\},\\
\alpha_{j,k}, \beta_{j,k}, &\qquad j \in \{1, \ldots, s\},\ k \in \{1, \ldots, \mu_j\},
\end{aligned}$$

are a basis for $\mathrm{Sol}(F)$. •

Of course, we should verify that the procedure does, indeed, produce a basis for $\mathrm{Sol}_{t_0}(F)$ and $\mathrm{Sol}(F)$.

5.6.20 Theorem (Basis of solutions for a system of linear homogeneous ordinary difference equations with constant coefficients) *Given a system of linear homogeneous ordinary difference equations*

$$\mathrm{F}\colon \mathbb{T} \times \mathsf{X} \oplus \mathsf{X} \to \mathsf{X}$$

in an n*-dimensional* $\mathbb{R}$*-vector space* X*, with right-hand side*

$$\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}) = \mathsf{A}(\mathrm{x}),$$

and with $\mathrm{t}_0 \in \mathbb{T}$*, define* n *functions as in Procedure 5.6.19. Then these functions for a basis for* $\mathrm{Sol}_{\mathrm{t}_0}(\mathrm{F})$ *and* $\mathrm{Sol}(\mathrm{F})$*, as asserted.*

Proof By virtue of Exercise 5.6.1, we can choose a basis $\{e_1, \dots, e_n\}$ for X and so assume that $\mathsf{X} = \mathbb{R}^n$.

Let us first work with the zero eigenvalue and show that the functions $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_{m_0}$ are solutions. This assertion, however, is clear since, just by definition, $\boldsymbol{\eta}_k$ is the solution corresponding to the initial condition $\boldsymbol{\eta}_k(t_0) = \boldsymbol{y}_k$. These m_0 solutions are, moreover, linearly independent since, when evaluated at t_0, they are linearly independent, cf. Corollary 5.6.4.

Let us next fix $j \in \{1, \dots, r\}$ and show that $\boldsymbol{\xi}_{j,k}$, $k \in \{1, \dots, m_j\}$, are solutions for $\boldsymbol{F}$. Let $t_0 \in \mathbb{T}$. Let us also fix $k \in \{1, \dots, m_j\}$. By Theorem 5.6.17(viii), the unique solution to the initial value problem

$$\boldsymbol{\xi}(t + h) = \boldsymbol{A}\boldsymbol{\xi}(t), \quad \boldsymbol{\xi}(t_0) = \mathsf{P}_{\boldsymbol{A}}\left(\tfrac{t_0}{h}\right)\boldsymbol{x}_{j,k},$$

is

$$t \mapsto \mathsf{P}_{\boldsymbol{A}}\left(\tfrac{t-t_0}{h}\right)\mathsf{P}_{\boldsymbol{A}}\left(\tfrac{t_0}{h}\right)\boldsymbol{x}_{j,k} = \mathsf{P}_{\boldsymbol{A}}\left(\tfrac{t}{h}\right)\boldsymbol{x}_{j,k},$$

using Theorem 5.6.17(v) and the obvious fact that the matrices $\frac{t}{h}\boldsymbol{A}$ and $\frac{t_0}{h}\boldsymbol{A}$ commute. Now we have

$$\begin{aligned}
\mathsf{P}_{\boldsymbol{A}}\left(\tfrac{t}{h}\right)\boldsymbol{x}_{j,k} &= (\ell_j \boldsymbol{I}_n + (\boldsymbol{A} - \ell_j \boldsymbol{I}_n))^{t/h}\boldsymbol{x}_{j,k} \\
&= \sum_{m=0}^{t/h} \binom{t/h}{m} \ell_j^{t/h-m}(\boldsymbol{A} - \ell_j \boldsymbol{I}_n)^m \boldsymbol{x}_{j,k} \\
&= \ell_j^{t/h} \sum_{m=0}^{t/h} \binom{t/h}{m} (\ell_j^{-1}\boldsymbol{A} - \boldsymbol{I}_n)^m \boldsymbol{x}_{j,k}
\end{aligned}$$

using parts (iv) and (v) of Theorem 5.6.17. Since $\boldsymbol{x}_{j,k} \in \overline{\mathsf{W}}(\ell_j, \boldsymbol{A})$, we in fact have

$$\mathsf{P}_{\boldsymbol{A}}\left(\tfrac{t}{h}\right)\boldsymbol{x}_{j,k} = \ell_j^{t/h} \sum_{m=0}^{\min\{t/h, m_j - 1\}} \binom{t/h}{m} (\ell_j^{-1}\boldsymbol{A} - \boldsymbol{I}_n)^m \boldsymbol{x}_{j,k}.$$

However, this last expression is exactly $\boldsymbol{\xi}_{j,k}(t)$, showing that this is indeed a solution for $\boldsymbol{F}$.

Next we show that, still keeping $j \in \{1, \dots, r\}$ fixed, the m_j solutions $\xi_{j,k}$, $k \in \{1, \dots, m_j\}$, are linearly independent. As we have seen,

$$\xi_{j,k}(t_0) = \mathsf{P}_A\left(\tfrac{t_0}{h}\right)\boldsymbol{x}_{j,k}, \qquad k \in \{1, \dots, m_j\}.$$

Thus, for $c_1, \dots, c_{m_j} \in \mathbb{R}$, we have

$$\begin{aligned}
& c_1\xi_{j,k}(t_0) + \cdots + c_{m_j}\xi_{j,m_j}(t_0) = \mathbf{0} \\
\implies \quad & c_1\mathsf{P}_A\left(\tfrac{t_0}{h}\right)\boldsymbol{x}_{j,1} + \cdots + c_{m_j}\mathsf{P}_A\left(\tfrac{t_0}{h}\right)\boldsymbol{x}_{j,m_j} = \mathbf{0} \\
\implies \quad & \mathsf{P}_A\left(\tfrac{t_0}{h}\right)(c_1\boldsymbol{x}_{j,1} + \cdots + c_{m_j}\boldsymbol{x}_{j,m_j}) = \mathbf{0} \\
\implies \quad & c_1\boldsymbol{x}_{j,1} + \cdots + c_{m_j}\boldsymbol{x}_{j,m_j} = \mathbf{0} \\
\implies \quad & c_1 = \cdots = c_{m_j} = 0,
\end{aligned}$$

since $\boldsymbol{x}_{j,1}, \dots, x_{j,m_j}$ are constructed as being linearly independent. Note that we have also use the fact that $\mathsf{A}|\overline{\mathsf{W}}(\ell_j, \mathsf{A})$ is invertible (its determinant is $\ell_j^{m_j}$). By Corollary 5.6.4 we conclude that $\xi_{j,1}, \dots, \xi_{j,m_j}$ are indeed linearly independent.

Now we fix $j \in \{1, \dots, s\}$ and work with the complex eigenvalue $\lambda_j = \rho_j e^{i\theta_j}$. First of all, let us define $\zeta_{j,k}\colon \mathbb{T} \to \mathbb{C}^n$, $k \in \{1, \dots, \mu_j\}$, by

$$\zeta_{j,k} = \mathsf{P}_{A_{\mathbb{C}}}\left(\tfrac{t}{h}\right)\boldsymbol{z}_{j,k}.$$

Then, exactly as above for the real eigenvalues, we have

$$\zeta_{j,k}(t) = \lambda_j^{t/h} \sum_{m=0}^{\min\{t/h, \mu_j - 1\}} \binom{t/h}{m} \lambda_j^{-m} (\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m \boldsymbol{z}_{j,k}.$$

Moreover, $\zeta_{j,k}$, $k \in \{1, \dots, \mu_j\}$, are solutions for $\boldsymbol{F}_{\mathbb{C}}$. Therefore, by Lemma 5.6.15, the real and imaginary parts of $\zeta_{j,k}$ are solutions for $\boldsymbol{F}$. Using Lemma 1 from the proof of Theorem 5.2.24, we have

$$\operatorname{Re}((\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m) = \sum_{l=0}^{\lceil \frac{m-1}{2} \rceil} \binom{m}{2l} (-1)^l \rho_j^{2l} \sin(\theta_j)^{2l} (\boldsymbol{A} - \rho_j \cos(\theta_j)\boldsymbol{I}_n)^{m-2l}$$

and

$$\operatorname{Im}((\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m) = \sum_{l=0}^{\lfloor \frac{m-1}{2} \rfloor} \binom{m}{2l+1} (-1)^{l+1} \rho_j^{2l+1} \sin(\theta_j)^{2l+1} (\boldsymbol{A} - \rho_j \cos(\theta_j)\boldsymbol{I}_n)^{m-2l-1}$$

With $\lambda_j = \rho_j(\cos(\theta_j) + i\sin(\theta_j))$, we then compute

$$\begin{aligned}
& \operatorname{Re}\left(\binom{t/h}{m} \lambda_j^{-m} (\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m\right) \\
& \quad = \binom{t/h}{m} \rho_j^{-m} \cos(m\theta_j) \sum_{l=0}^{\lceil \frac{m-1}{2} \rceil} \binom{m}{2l} (-1)^l \rho_j^{2l} \sin(\theta_j)^{2l} (\boldsymbol{A} - \rho_j \cos(\theta_j)\boldsymbol{I}_n)^{m-2l} \\
& \quad + \binom{t/h}{m} \rho_j^{-m} \sin(m\theta_j) \sum_{l=0}^{\lfloor \frac{m-1}{2} \rfloor} \binom{m}{2l+1} (-1)^{l+1} \rho_j^{2l+1} \sin(\theta_j)^{2l+1} (\boldsymbol{A} - \rho_j \cos(\theta_j)\boldsymbol{I}_n)^{m-2l-1}
\end{aligned}$$

and

$$\begin{aligned}
&\operatorname{Im}\left(\binom{t/h}{m}\lambda_j^{-m}(\boldsymbol{A}_{\mathbb{C}}-\lambda_j\boldsymbol{I}_n)^m\right)\\
&\quad=\binom{t/h}{m}\rho_j^{-m}\cos(m\theta_j)\sum_{l=0}^{\lfloor\frac{m-1}{2}\rfloor}\binom{m}{2l+1}(-1)^{l+1}\rho_j^{2l+1}\sin(\theta_j)^{2l+1}(\boldsymbol{A}-\rho_j\cos(\theta_j)\boldsymbol{I}_n)^{m-2l-1}\\
&\qquad\qquad-\binom{t/h}{m}\rho_j^{-m}\sin(m\theta_j)\sum_{l=0}^{\lceil\frac{m-1}{2}\rceil}\binom{m}{2l}(-1)^{l}\rho_j^{2l}\sin(\theta_j)^{2l}(\boldsymbol{A}-\rho_j\cos(\theta_j)\boldsymbol{I}_n)^{m-2l}
\end{aligned}$$

With

$$\operatorname{Re}(\lambda_j^{t/h}\boldsymbol{z}_{j,k})=\rho_j^{t/h}\cos\left(\theta_j\tfrac{t}{h}\right)\boldsymbol{a}_{j,k}-\rho_j^{t/h}\sin\left(\theta_j\tfrac{t}{h}\right)\boldsymbol{b}_{j,k}$$

and

$$\operatorname{Im}(\lambda_j^{t/h}\boldsymbol{z}_{j,k})=\rho_j^{t/h}\cos\left(\theta_j\tfrac{t}{h}\right)\boldsymbol{b}_{j,k}+\rho_j^{t/h}\sin\left(\theta_j\tfrac{t}{h}\right)\boldsymbol{a}_{j,k},$$

one can then verify that

$$\begin{aligned}
\boldsymbol{\alpha}_{j,k}(t)&=\operatorname{Re}\left(\lambda_j^{t/h}\sum_{m=0}^{\min\{t/h,\mu_j-1\}}\binom{t/h}{m}\lambda_j^{-m}(\boldsymbol{A}_{\mathbb{C}}-\lambda_j\boldsymbol{I}_n)^m\boldsymbol{z}_{j,k}\right),\\
\boldsymbol{\beta}_{j,k}(t)&=\operatorname{Im}\left(\lambda_j^{t/h}\sum_{m=0}^{\min\{t/h,\mu_j-1\}}\binom{t/h}{m}\lambda_j^{-m}(\boldsymbol{A}_{\mathbb{C}}-\lambda_j\boldsymbol{I}_n)^m\boldsymbol{z}_{j,k}\right),
\end{aligned}$$

for $k\in\{1,\dots,\mu_j\}$. This shows that $\boldsymbol{\alpha}_{j,k}$ and $\boldsymbol{\beta}_{j,k}$ are solutions for $\boldsymbol{F}$ for $k\in\{1,\dots,\mu_j\}$.

Now we verify that

$$\boldsymbol{\alpha}_{j,1},\dots,\boldsymbol{\alpha}_{j,\mu_j},\boldsymbol{\beta}_{j,1},\dots,\boldsymbol{\beta}_{j,\mu_j}$$

are linearly independent. As above in the real case, the complex solutions $\boldsymbol{\zeta}_{j,1},\dots,\boldsymbol{\zeta}_{j,\mu_j}$ for $\boldsymbol{F}_{\mathbb{C}}$ are linearly independent. Now let $t_0\in\mathbb{T}$ and $c_1,\dots,c_{\mu_j},d_1,\dots,d_{\mu_j}\in\mathbb{R}$, and note that

$$\begin{aligned}
&\sum_{k=1}^{\mu_j}(c_k\boldsymbol{\alpha}_{j,k}(t_0)+d_k\boldsymbol{\beta}_{j,k}(t_0))=\boldsymbol{0}\\
\implies\quad&\sum_{k=1}^{\mu_j}(c_k\operatorname{Re}(\boldsymbol{\zeta}_{j,k})(t_0)+d_k\operatorname{Im}(\boldsymbol{\zeta}_{j,k})(t_0))=\boldsymbol{0}\\
\implies\quad&\sum_{k=1}^{\mu_j}(c_k\operatorname{Re}(\mathsf{P}_{\boldsymbol{A}_{\mathbb{C}}}\left(\tfrac{t_0}{h}\right)\boldsymbol{z}_{j,k})+d_k\operatorname{Im}(\mathsf{P}_{\boldsymbol{A}_{\mathbb{C}}}\left(\tfrac{t_0}{h}\right)\boldsymbol{z}_{j,k}))\\
\implies\quad&\sum_{k=1}^{\mu_j}(c_k\mathsf{P}_{\boldsymbol{A}_{\mathbb{C}}}\left(\tfrac{t_0}{h}\right)\boldsymbol{a}_{j,k}+d_k\mathsf{P}_{\boldsymbol{A}_{\mathbb{C}}}\left(\tfrac{t_0}{h}\right)\boldsymbol{b}_{j,k})=\boldsymbol{0}\\
\implies\quad&\sum_{k=1}^{\mu_j}(c_k\boldsymbol{a}_{j,k}+d_k\boldsymbol{b}_{j,k})=\boldsymbol{0}\\
\implies\quad&c_1=\cdots=c_{\mu_j}=d_1=\cdots=d_{\mu_j}=0,
\end{aligned}$$

using the fact that, since $\boldsymbol{A}$ is real, $\mathsf{P}_{\boldsymbol{A}_{\mathbb{C}}}\left(\frac{t_0}{h}\right)$ is also real and, using Theorem I-5.4.68(i), this gives the linear independence of

$$\boldsymbol{\alpha}_{j,1},\dots,\boldsymbol{\alpha}_{j,\mu_j},\boldsymbol{\beta}_{j,1},\dots,\boldsymbol{\beta}_{j,\mu_j},$$

as claimed.

Now we have $m_0+m_1+\cdots+m_r+2(\mu_1+\cdots+\mu_s)=n$ solutions for $\boldsymbol{F}$. It remains to show that the collection of all of these solutions are linearly independent. Let us suppose that

$$\begin{aligned}
&\underbrace{c_1\boldsymbol{\eta}_1(t)+\cdots+c_{m_0}\boldsymbol{\eta}_{m_0}(t)}_{\in\overline{\mathsf{W}}(0,\boldsymbol{A})}\\
&\quad+\underbrace{d_{1,1}\boldsymbol{\xi}_{1,1}(t)+\cdots+d_{1,m_1}\boldsymbol{\xi}_{1,m_1}(t)}_{\in\overline{\mathsf{W}}(\ell_1,\boldsymbol{A})}+\cdots+\underbrace{d_{r,1}\boldsymbol{\xi}_{r,1}(t)+\cdots+d_{r,m_r}\boldsymbol{\xi}_{r,m_r}(t)}_{\in\overline{\mathsf{W}}(\ell_r,\boldsymbol{A})}\\
&\quad+\underbrace{e_{1,1}\boldsymbol{a}_{1,1}(t)+\cdots+e_{1,\mu_1}\boldsymbol{a}_{1,\mu_1}(t)}_{\in\overline{\mathsf{W}}(\lambda_1,\boldsymbol{A})}+\cdots+\underbrace{e_{s,1}\boldsymbol{a}_{s,1}(t)+\cdots+e_{s,\mu_s}\boldsymbol{a}_{s,\mu_s}(t)}_{\in\overline{\mathsf{W}}(\lambda_s,\boldsymbol{A})}\\
&\quad+\underbrace{f_{1,1}\boldsymbol{b}_{1,1}(t)+\cdots+f_{1,\mu_1}\boldsymbol{b}_{1,\mu_1}(t)}_{\in\overline{\mathsf{W}}(\lambda_1,\boldsymbol{A})}+\cdots+\underbrace{f_{s,1}\boldsymbol{b}_{s,1}(t)+\cdots+f_{s,\mu_s}\boldsymbol{b}_{s,\mu_s}(t)}_{\in\overline{\mathsf{W}}(\lambda_s,\boldsymbol{A})}=\boldsymbol{0},
\end{aligned}$$

for suitable scalar coefficients. Since the generalised eigenspaces intersect in $\{0\}$ by Proposition I-5.4.60, and since the generalised eigenspaces are invariant under $\mathsf{P}_{\boldsymbol{A}}\left(\frac{t}{h}\right)$ for all $t\in\mathbb{T}$ by Theorem 5.6.17(vii), for the preceding equation to hold, each of its components in each of the generalised eigenspaces must be zero, i.e.,

$$c_1\boldsymbol{\eta}_1(t)+\cdots+c_{m_0}\boldsymbol{\eta}_{m_0}(t),$$

and

$$c_{j,1}\boldsymbol{\xi}_{j,1}(t)+\cdots+c_{j,m_j}\boldsymbol{\xi}_{j,m_j}(t)=\boldsymbol{0},\qquad j\in\{1,\dots,r\},$$

and

$$d_{j,1}\boldsymbol{a}_{j,1}(t)+\cdots+d_{j,\mu_j}\boldsymbol{a}_{j,\mu_j}(t)+e_{j,1}\boldsymbol{b}_{j,1}(t)+\cdots+e_{j,\mu_j}\boldsymbol{b}_{j,\mu_j}(t)=\boldsymbol{0},\qquad j\in\{1,\dots,s\}.$$

This implies that all coefficients must be zero, since we have already shown the linear independence of the solutions with initial conditions in each of the subspaces $\overline{\mathsf{W}}(o,\boldsymbol{A})$, $\overline{\mathsf{W}}(\ell_l,\boldsymbol{A})$, $j\in\{1,\dots,r\}$, and $\overline{\mathsf{W}}(\lambda_j,\boldsymbol{A})$, $j\in\{1,\dots,s\}$. Thus we have the desired linear independence, and thus the theorem follows, as concerns $\mathrm{Sol}_{t_0}(F)$. It is clear that, when A is invertible, the solutions associated with the nonzero eigenvalues form a basis for $\mathrm{Sol}(F)$. ■

From the proof of the theorem, we provide the following comment on how one might deal with complex eigenvalues in practice.

5.6.21 Remark (Computing solutions associated with complex eigenvalues) The formulae (5.25) and (5.26) of Procedure 5.6.19, while fun to look at, are typically not the best ways to work out solutions associated with complex eigenvalues. However, the proof of the preceding theorem tells us an alternative that is easier in easy examples (although using a computer algebra package is even easier). Indeed, in the proof we saw that

$$\boldsymbol{\alpha}_{j,k}(t) = \operatorname{Re}\left(\lambda_j^{t/h} \sum_{m=0}^{\min\{t/h,\mu_j-1\}} \binom{t/h}{m} \lambda_j^{-m} (\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m \boldsymbol{z}_{j,k}\right),$$
$$\boldsymbol{\beta}_{j,k}(t) = \operatorname{Im}\left(\lambda_j^{t/h} \sum_{m=0}^{\min\{t/h,\mu_j-1\}} \binom{t/h}{m} \lambda_j^{-m} (\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m \boldsymbol{z}_{j,k}\right)$$

for $k \in \{1, \dots, \mu_j\}$. Thus, in practice, one might simply compute

$$\boldsymbol{\zeta}_{j,k}(t) = \lambda_j^{t/h} \sum_{m=0}^{\min\{t/h,\mu_j-1\}} \binom{t/h}{m} \lambda_j^{-m} (\boldsymbol{A}_{\mathbb{C}} - \lambda_j \boldsymbol{I}_n)^m \boldsymbol{z}_{j,k},$$

$k \in \{1, \dots, s\}$, and simply takes its real and imaginary parts as linearly independent solutions. •

We can now give an algorithm for computing, in principle, the operator power function. The following procedure, while given for computing P_{A}, obviously may be used as well to compute the discrete-time state transition matrix $\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t) = \mathsf{P}_{\mathsf{A}}\left(\frac{t-t_0}{h}\right)$ for a system of linear homogeneous ordinary difference equations with constant coefficients.

5.6.22 Procedure (Operator power function) Given an n-dimensional $\mathbb{R}$-vector space X and $\mathsf{A} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$, do the following.

1. Choose a basis $\{e_1, \dots, e_n\}$ and let $\boldsymbol{A}$ be the matrix representative of A. If $\mathsf{X} = \mathbb{R}^n$, one can just take $\boldsymbol{A}$ to be the usual matrix associated with $\mathsf{A} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$.
2. Using Procedure 5.6.19, determine a fundamental set of solutions $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n$ from $t_0 = 0$, defined on the time-domain $\mathbb{Z}_{\geq 0}$, for the system of linear homogeneous ordinary difference equations $\boldsymbol{F}$ in $\mathbb{R}^n$ with right-hand side
$$\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}.$$
3. Define
$$\boldsymbol{\Xi}(t) = \begin{bmatrix} \xi_{1,1}(t) & \xi_{2,1}(t) & \cdots & \xi_{n,1}(t) \\ \xi_{1,2}(t) & \xi_{2,2}(t) & \cdots & \xi_{n,2}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{1,n}(t) & \xi_{2,n}(t) & \cdots & \xi_{n,n}(t) \end{bmatrix},$$
where $\xi_{j,k}$ is the kth component of $\boldsymbol{\xi}_j$.

4. Using Procedure 5.6.8 calculate

$$\mathsf{P}_A(t) = \Phi^{\mathrm{d}}_{A,0}(t) = \Xi(t)\Xi(0)^{-1}, \qquad t \in \mathbb{Z}_{\geq 0}.$$ •

5.6.2.4 Some examples

Exercises

5.6.1 Let X be an n-dimensional $\mathbb{R}$-vector space and let F be a system of linear ordinary difference equations in X with right-hand side

$$\widehat{F}(t,x) = \mathsf{A}(t)(x) + b(t)$$

for $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X};\mathsf{X})$ and $b\colon \mathbb{T} \to \mathsf{X}$. Let $\{e_1,\dots,e_n\}$ be a basis for X and write

$$b(t) = \sum_{j=1}^{n} b_j(t)e_j, \quad \mathsf{A}(t)(e_j) = \sum_{k=1}^{n} A^k_j(t)e_k, \qquad j \in \{1,\dots,n\},$$

for functions $b_j\colon \mathbb{T} \to \mathbb{R}$, $j \in \{1,\dots,n\}$, and $A^k_j\colon \mathbb{T} \to \mathbb{R}$, $j,k \in \{1,\dots,n\}$. This defines $\boldsymbol{b}\colon \mathbb{T} \to \mathbb{R}^n$ and $\boldsymbol{A}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$. Denote by $\boldsymbol{F}$ the system of linear ordinary difference equations in $\mathbb{R}^n$ given by

$$\boldsymbol{F}(t,\boldsymbol{x},\boldsymbol{x}^{(1)}) = \boldsymbol{x}^{(1)} - \boldsymbol{A}(t)\boldsymbol{x} - \boldsymbol{b}(t).$$

Answer the following questions.

(a) Show that $\xi\colon \mathbb{T}' \to \mathsf{X}$ is a solution for F if and only if the function $\boldsymbol{\xi}\colon \mathbb{T}' \to \mathbb{R}^n$, defined by

$$\xi(t) = \sum_{j=1}^{n} \xi_j(t)e_j,$$

is a solution for $\boldsymbol{F}$.

Now let $\{\tilde{e}_1,\dots,\tilde{e}_n\}$ be another basis for X and let $\boldsymbol{P}$ be the change of basis matrix defined by

$$\tilde{e}_j = \sum_{k=1}^{n} P_{kj}e_k, \qquad j \in \{1,\dots,n\}.$$

Define $\tilde{\boldsymbol{b}}\colon \mathbb{T} \to \mathbb{R}^n$, $\tilde{\boldsymbol{A}}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$, and $\tilde{\boldsymbol{F}}$ as above, for this new basis.

(b) Show that $\tilde{\boldsymbol{b}}(t) = \boldsymbol{P}\boldsymbol{b}(t)$ and $\tilde{\boldsymbol{A}}(t) = \boldsymbol{P}^{-1}\boldsymbol{A}(t)\boldsymbol{P}$ for every $t \in \mathbb{T}$.
Hint: *Use the change of basis formulae from Proposition I-5.4.26 and from Theorem I-5.4.32.*

(c) Show that, if $\boldsymbol{\xi}\colon \mathbb{T}' \to \mathbb{R}^n$ is a solution for $\boldsymbol{F}$, then $\tilde{\boldsymbol{\xi}}\colon \mathbb{T}' \to \mathbb{R}^n$ is a solution for $\tilde{\boldsymbol{F}}$ if and only if $\tilde{\boldsymbol{\xi}}(t) = \boldsymbol{P}^{-1}\boldsymbol{\xi}(t)$ for every $t \in \mathbb{T}$.

5.6.2 Let X be an n-dimensional $\mathbb{R}$-vector space and let F be a system of linear homogeneous ordinary difference equations with right-hand side

$$\widehat{F}(t,x) = \mathsf{A}(t)(x)$$

for $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X};\mathsf{X})$. Let $\{e_1,\dots,e_n\}$ be a basis for X and let $\boldsymbol{A}(t)$ be the matrix representative for $\mathsf{A}(t)$, $t \in \mathbb{T}$, and let $\boldsymbol{F}$ be the corresponding system of linear homogeneous ordinary difference equations in $\mathbb{R}^n$ with right-hand side

$$\widehat{\boldsymbol{F}}(t,\boldsymbol{x}) = \boldsymbol{A}(t)\boldsymbol{x}.$$

cf. Exercise 5.6.1.

(a) Show that, for every $t_0 \in \mathbb{T}$ and $t \in \mathbb{T}_{\geq t_0}$, the matrix representative of $\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)$ is $\Phi^{\mathrm{d}}_{\boldsymbol{A},t_0}(t)$.

(b) Show that, for every $t, t_0 \in \mathbb{T}$, if F is invertible then the matrix representative of $\Phi^{\mathrm{d}}_{\mathsf{A}}(t,t_0)$ is $\Phi^{\mathrm{d}}_{\boldsymbol{A}}(t,t_0)$.

Now let $\{\tilde{e}_1,\dots,\tilde{e}_n\}$ be another basis for X and let $\boldsymbol{P}$ be the change of basis matrix defined by

$$\tilde{e}_j = \sum_{k=1}^{n} P_{kj} e_k, \qquad j \in \{1,\dots,n\}.$$

Define $\tilde{\boldsymbol{A}}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$ and $\tilde{\boldsymbol{F}}$ as above, for this new basis.

(c) Show that, for every $t_0 \in \mathbb{T}$ and $t \in \mathbb{T}_{\geq t_0}$,

$$\Phi^{\mathrm{d}}_{\tilde{\boldsymbol{A}},t_0}(t) = \boldsymbol{P}^{-1}\Phi^{\mathrm{d}}_{\boldsymbol{A},t_0}(t)\boldsymbol{P}.$$

(d) Show that, for every $t, t_0 \in \mathbb{T}$, if F is invertible then

$$\Phi^{\mathrm{d}}_{\tilde{\boldsymbol{A}}}(t,t_0) = \boldsymbol{P}^{-1}\Phi^{\mathrm{d}}_{\boldsymbol{A}}(t,t_0)\boldsymbol{P}.$$

Section 5.7

Systems of linear inhomogeneous ordinary difference equations

In this section we extend our discussion of homogeneous equations in Section 5.6 to inhomogeneous equations. Thus we are talking about systems of linear ordinary difference equations F in a finite-dimensional $\mathbb{R}$-vector space X with right-hand side

$$\begin{aligned} \widehat{F}\colon \mathbb{T} \times \mathsf{X} &\to \mathsf{X} \\ (t, x) &\mapsto \mathsf{A}(t)(x) + b(t) \end{aligned} \tag{5.27}$$

for maps $b\colon \mathbb{T} \to \mathsf{X}$ and $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$. In our treatment of scalar equations in Section 4.7, we gave no fewer than three methods for working with inhomogeneous equations, two general methods (using Casoratians in Section 4.7.1.2 and the theory of discrete-time Green's function in Section 4.7.1.3) and one method that only works for inhomogeneous terms that are pretty uninteresting (the "method of undetermined coefficients in Section 4.7.2.1). We shall not be so expansive for systems of linear inhomogeneous equations, and shall really only consider "the" method for working with such equations, since this method is as tractable as any other method in practice (which is to say, not very tractable at all, barring the use of a computer algebra package), and is exceptionally powerful in developing the theory of systems of linear ordinary difference equations.

As we have done in all preceding developments of linear ordinary difference equations, we work first in the general time-varying case, and then in the case of constant coefficients.

Do I need to read this section? The material in this section is fundamental to the theory of linear systems. •

5.7.1 Equations with time-varying coefficients

We state the, by now, more or less obvious results concerning existence and uniqueness, now for systems of linear inhomogeneous ordinary difference equations.

5.7.1 Proposition (Local existence and uniqueness of solutions for systems of linear inhomogeneous ordinary difference equations) *Consider the system of linear inhomogeneous ordinary difference equations* F *with right-hand side (5.27). Let* $(\mathrm{t}_0, \mathrm{x}_0) \in \mathbb{T} \times \mathsf{X}$. *Then there exists a unique* $\xi\colon \mathbb{T}_{\geq \mathrm{t}_0} \to \mathsf{X}$ *that is a solution for* F *and which satisfies* $\xi(\mathrm{t}_0) = \mathrm{x}_0$. *If* F *is invertible, then there exists a unique* $\xi\colon \mathbb{T} \to \mathsf{X}$ *that is a solution for* F *and which satisfies the initial conditions.*

Proof By Proposition 5.6.1, there exists a solution $\xi_{\mathrm{h}}\colon \mathbb{T}_{\geq t_0} \to \mathsf{X}$ for F_{h} satisfying $\xi_{\mathrm{h}}(t_0) = x_0$. Moreover, $\xi_{\mathrm{h}}(t) = \Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)(x_0)$. Now define

$$\begin{aligned}\xi\colon \mathbb{T}_{\geq t_0} &\to \mathsf{X}\\ t &\mapsto \Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)(x_0) + \sum_{j=0}^{(t-t_0-h)/h} \Phi^{\mathrm{d}}_{\mathsf{A},t_0+(j+1)h}(t)(b(t_0+jh)).\end{aligned}$$

Let us verify that this is a solution satisfying the initial conditions. We calculate

$$\begin{aligned}\xi(t+h) &= \Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t+h)(x_0) + \sum_{j=0}^{(t-t_0)/h} \Phi^{\mathrm{d}}_{\mathsf{A},t_0+(j+1)h}(t+h)b(t_0+jh)\\ &= \mathsf{A}(t)\circ\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)(x_0) + \sum_{j=0}^{(t-t_0-h)/h} \mathsf{A}(t)\circ\Phi^{\mathrm{d}}_{\mathsf{A},t_0+(j+1)h}(t)(b(t_0+jh)) + \Phi^{\mathrm{d}}_{\mathsf{A},t+h}(t+h)(b(t))\\ &= \mathsf{A}(t)\circ\left(\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)(x_0) + \sum_{j=0}^{(t-t_0-h)/h} \Phi^{\mathrm{d}}_{\mathsf{A},t_0+(j+1)h}(t)(b(t_0+jh))\right) + b(t)\\ &= \mathsf{A}(t)(\xi(t)) + b(t),\end{aligned}$$

i.e., ξ is a solution of F. Moreover, we also clearly have $\xi(t_0) = x_0$, by conventions with summations.

To conclude uniqueness, suppose that we have two solutions ξ_1 and ξ_2 defined on the same interval $\mathbb{T}'$. Then

$$\xi_1(t+h) = \mathsf{A}(t)(\xi_1(t)) + b(t), \quad \xi_2(t+h) = \mathsf{A}(t)(\xi_2(t)) + b(t),$$

and $\xi_1(t_0) = \xi_2(t_0) = x_0$. Therefore,

$$(\xi_1 - \xi_2)(t+h) = \mathsf{A}(t)(\xi_1(t) - \xi_2(t)), \quad (\xi_1 - \xi_2)(t_0) = 0.$$

By the uniqueness assertion of Proposition 5.6.1, we conclude that $\xi_1 - \xi_2 = 0$, i.e., $\xi_1 = \xi_2$. ∎

Since, in the proof of Proposition 5.7.1, we gave an explicit formula for solutions to initial value problems, it is worth extracting this explicit formula.

5.7.2 Corollary (An explicit solution for systems of linear inhomogeneous ordinary difference equations) *Consider the system of linear inhomogeneous ordinary difference equations* F *with right-hand side (5.27). Given* $\mathrm{t}_0 \in \mathbb{T}$ *and* $\mathrm{x}_0 \in \mathsf{X}$*, the unique solution* $\xi\colon \mathbb{T} \to \mathsf{X}$ *to the initial value problem*

$$\xi(\mathrm{t}+\mathrm{h}) = \mathsf{A}(\mathrm{t})(\xi(\mathrm{t})) + \mathrm{b}(\mathrm{t}), \quad \xi(\mathrm{t}_0) = \mathrm{x}_0,$$

is

$$\xi(\mathrm{t}) = \Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0}(\mathrm{t})(\mathrm{x}_0) + \sum_{\mathrm{j}=0}^{(\mathrm{t}-\mathrm{t}_0-\mathrm{h})/\mathrm{h}} \Phi^{\mathrm{d}}_{\mathsf{A},\mathrm{t}_0+(\mathrm{j}+1)\mathrm{h}}(\mathrm{t})(\mathrm{b}(\mathrm{t}_0+\mathrm{jh})), \qquad \mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0}. \tag{5.28}$$

The formula (5.28) for solutions to systems of linear inhomogeneous ordinary differential equations is often called the ***variation of constants formula***.

We note that this solution bears a strong resemblance in form to the discrete-time Green's function solution for scalar systems given in Theorem 4.7.6; indeed, one can think of the discrete-time state transition map as playing the rôle of a discrete-time Green's function in this case. In particular, given $b \in \mathsf{X}$ (a constant vector, note) the physical interpretation of Remark 4.7.8–2 applies to the map $t \mapsto \Phi^{\mathrm{d}}_{\mathsf{A},\tau}(t)(b)$, and leads us to think of this as being the result of applying an impulse at time τ with (vector) magnitude b. This leads to the important notion in system theory of the impulse response.

The same sort of comments as given following Proposition 4.6.1 are valid here, in terms of comparing the preceding result with Proposition 5.2.2. In particular, there is this notion of invertibility for difference equations that does not arise for differential equations. Let us clearly enunciate the character of invertibility in the current setting.

5.7.3 Proposition (Invertible systems of linear inhomogeneous ordinary difference equations) *A system of linear inhomogeneous ordinary difference equations* F *with right-hand side*

$$\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}) = \mathsf{A}(\mathrm{t})(\mathrm{x}) + \mathrm{b}(\mathrm{t}),$$

for $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$ *and* $\mathrm{b}\colon \mathbb{T} \to \mathsf{X}$ *is invertible if and only if* $\det \mathsf{A}(\mathrm{t}) \neq 0$ *for every* $\mathrm{t} \in \mathbb{T}_{\mathrm{F}}$.

Proof This follows immediately from the definition of invertibility in Definition 3.4.5, here because the affine mapping

$$x \mapsto \mathsf{A}(t)(x) + b(t)$$

is invertible if and only if $\mathsf{A}(t)$ is invertible. ■

Now we can discuss the set of all solutions of a system of linear inhomogeneous ordinary difference equation F with right-hand side

$$\begin{aligned} \widehat{F}\colon \mathbb{T} \times \mathsf{X} &\to \mathsf{X} \\ (t, x) &\mapsto \mathsf{A}(t)(x). \end{aligned}$$

To this end, we denote by

$$\mathrm{Sol}_{t_0}(F) = \left\{ \xi \in \mathsf{X}^{\mathbb{T}_{\geq t_0}} \;\middle|\; \xi(t + h) = \mathsf{A}(t)(\xi(t)),\ t \in \mathbb{T}_{F, \geq t_0} \right\}$$

the set of solutions for F from t_0. When F is invertible, we denote

$$\mathrm{Sol}(F) = \left\{ \xi \in \mathsf{X}^{\mathbb{T}} \;\middle|\; \xi(t + h) = \mathsf{A}(t)(\xi(t)),\ t \in \mathbb{T}_{F} \right\}$$

While $\mathrm{Sol}_{t_0}(F)$ and $\mathrm{Sol}(F)$ are vectors space in the homogeneous case, in the inhomogeneous case this is no longer the case. However, the set of all solutions for

the homogeneous case plays an important rôle, even in the inhomogeneous case. To organise this discussion, we let F_{h} be the "homogeneous part" of F. Thus the right-hand side of F_{h} is

$$\widehat{F_{\mathrm{h}}}(t, x) = \mathsf{A}(t)(x).$$

As in Theorem 4.7.2, $\mathrm{Sol}_{t_0}(F)$ and $\mathrm{Sol}(F_{\mathrm{h}})$ are $\mathbb{R}$-vector spaces of dimension $\dim_{\mathbb{R}}(\mathsf{X})$. The following result is then the main structural result about the set of solutions to a system of linear inhomogeneous ordinary differential equations, mirroring Theorem 4.7.2 for scalar systems.

5.7.4 Theorem (Affine space structure of sets of solutions) *Consider the system of linear inhomogeneous ordinary difference equations* F *in the* n*-dimensional* $\mathbb{R}$*-vector space* X *with right-hand side (4.41). Let* $\xi_{\mathrm{p}} \in \mathrm{Sol}_{t_0}(\mathrm{F})$. *Then*

$$\mathrm{Sol}_{t_0}(\mathrm{F}) = \{\xi + \xi_{\mathrm{p}} \mid \xi \in \mathrm{Sol}_{t_0}(\mathrm{F}_{\mathrm{h}})\}.$$

Similarly, if $\xi_{\mathrm{p}} \in \mathrm{Sol}_{t_0}(\mathrm{F})$, *then*

$$\mathrm{Sol}(\mathrm{F}) = \{\xi + \xi_{\mathrm{p}} \mid \xi \in \mathrm{Sol}(\mathrm{F}_{\mathrm{h}})\}.$$

Proof This follows in the manner of the proof of Theorem 5.3.4, *mutatis mutandis*. ■

As with scalar linear inhomogeneous ordinary difference equations, there is an insightful correspondence to be made between the situation described in Theorem 5.7.4 and that of systems of linear algebraic equations described in Proposition I-5.4.48.

5.7.5 Remark (Comparison of Theorem 5.7.4 with systems of linear algebraic equations) Let us compare here the result of Theorem 5.7.4 with the situation in Proposition I-5.4.48 concerning linear algebraic equations of the form $L(u) = v_0$, for vector spaces U and W, a linear map $L \in \mathrm{L}(\mathsf{U}; \mathsf{W})$, and a fixed $w_0 \in \mathsf{W}$. In the setting of systems of linear inhomogeneous ordinary difference equations in a $\mathbb{R}$-vector space X, we have

$$\begin{aligned} \mathsf{U} &= \mathsf{X}^{\mathbb{T}_{\geq t_0}}, \\ \mathsf{W} &= \mathsf{X}^{\mathbb{T}_{\geq t_0}}, \\ L(f)(t) &= f(t + h) - \mathsf{A}(t)(f(t)), \\ w_0 &= b. \end{aligned}$$

Then Proposition 5.7.1 tells us that L is surjective, and so $w_0 \in \mathrm{image}(L)$. Thus we are in case (I-ii) of Proposition I-5.4.48, which exactly the statement of Theorem 5.7.4. Note that L is not injective, since Theorem 5.6.3 tells us that $\dim_{\mathbb{R}}(\ker(L)) = \dim_{\mathbb{R}}(\mathsf{X})$. Similar constructions hold, of course, in the particular case that F is invertible and solutions are defined on the entire time-domain. •

5.7.6 Remark (What happened to the Casoratian?) In Section 4.7.1.2 we described how the Casoratian can be used for scalar linear inhomogeneous ordinary difference equations to generate a particular solution. A similar development is possible for systems of equations, but we shall not pursue it here. It is worth recording the reasons for not doing so.

1. In Corollary 5.7.2 we produce a specific and natural "particular solution" for a system of linear inhomogeneous ordinary difference equations, namely the function that assigns to the inhomogeneous term "b," the solution

$$\xi_{\mathrm{p}}(t) = \sum_{j=0}^{(t-t_0-h)/h} \Phi^{\mathrm{d}}_{\mathsf{A},t+0+(j+1)h}(t)(b(t_0 + jh)).$$

   Then the form of the solution of Corollary 5.7.2 is $\xi = \xi_{\mathrm{h}} + \xi_{\mathrm{p}}$, where $\xi_{\mathrm{h}} \in \mathrm{Sol}(F_{\mathrm{h}})$ satisfies the initial conditions. This is just so cool. . . why would you want to do more?

2. In Section 4.6.1 we discussed the notion of a fundamental set of solutions for scalar linear homogeneous ordinary difference equations. There is no really distinguished fundamental set of solutions, and the Casoratian-related constructions were developed for an *arbitrary* fundamental set of solutions. This has its benefits in this setting, as the results are general in this respect.

   However, in Section 5.6.1.2 we saw that there was *one* object that naturally describes the solutions for a system of linear homogeneous ordinary difference equations, the discrete-time state transition map. Note that in Procedure 5.6.8 we indicate how to build the discrete-time state transition map from a fundamental set of solutions for a system of equations, through the fundamental matrix-function Ξ that we build after choosing a basis. It is the fundamental matrix, and its determinant, that would be involved in Casoratian-type constructions for systems of equations. However, these are only arrived at after choosing a basis, and so seem quite unnatural in our setting of general vector spaces. •

5.7.2 Equations with constant coefficients

We now specialise the discussion in the preceding section to systems of linear inhomogeneous ordinary difference equations with constant coefficients. Thus we are looking at a system of linear inhomogeneous ordinary difference equations F in a finite-dimensional $\mathbb{R}$-vector space X and with right-hand side given by

$$\widehat{F}(t, x) = \mathsf{A}(x) + b(t) \tag{5.29}$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$ and $b \colon \mathbb{T} \to \mathsf{X}$. Of course, all general results concerning the existence and uniqueness of solutions (i.e., Proposition 5.7.1) and of the structure of the set of solutions (i.e., Theorem 5.7.4) apply in the constant coefficient case. Here,

however, we can refine a little the explicit solution of Corollary 5.7.2 because, as per Theorem 5.6.17(viii), $\Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t) = \mathsf{P}_{\mathsf{A}}\left(\frac{t-t_0}{h}\right)$ in this case. We can thus summarise the situation in the following theorem.

5.7.7 Theorem (An explicit solution for systems of linear inhomogeneous ordinary difference equations with constant coefficients) *Consider the system of linear inhomogeneous ordinary difference equations* F *with constant coefficients and right-hand side (5.16). Given* $\mathrm{t}_0 \in \mathbb{T}$ *and* $\mathrm{x}_0 \in \mathsf{X}$*, the unique solution* $\xi\colon \mathbb{T}_{\geq \mathrm{t}_0} \to \mathsf{X}$ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \mathsf{A}(\xi(\mathrm{t})) + \mathrm{b}(\mathrm{t}), \quad \xi(\mathrm{t}_0) = \mathrm{x}_0,$$

is

$$\xi(\mathrm{t}) = \mathsf{P}_{\mathsf{A}}\left(\tfrac{\mathrm{t}-\mathrm{t}_0}{\mathrm{h}}\right)(\mathrm{x}_0) + \sum_{\mathrm{j}=0}^{(\mathrm{t}-\mathrm{t}_0-\mathrm{h})/\mathrm{h}} \mathsf{P}_{\mathsf{A}}\left(\tfrac{\mathrm{t}-\mathrm{t}_0-(\mathrm{j}+1)\mathrm{h}}{\mathrm{h}}\right)(\mathrm{b}(\mathrm{t}_0 + \mathrm{jh})), \qquad \mathrm{t} \in \mathbb{T}.$$

We comment that our observations Remark 4.7.10 about the particular solution

$$\xi_{\mathrm{p},b} = \sum_{j=0}^{(t-t_0-h)/h} \mathsf{P}_{\mathsf{A}}\left(\tfrac{t-t_0-(j+1)h}{h}\right)(b(t_0 + jh))$$

for constant coefficient systems and its relation to convolution integrals is also valid here.

5.7.8 Remark (What happened to the "method of undetermined coefficients"?) In Section 4.7.2.1 we spent some time describing a rather *ad hoc* method, the "method of undetermined coefficients," for finding particular solutions for scalar linear inhomogeneous ordinary difference equations with constant coefficients. A similar strategy is possible for systems of linear inhomogeneous ordinary difference equations with constant coefficients, but we shall not pursue it here. Here is why.

1. The rationale of Remark 5.7.6–1 is equally valid here: we have such a nice characterisation in Corollary 5.7.2 of a particular solution that to mess this up with an *ad hoc* procedure that only works for pretty uninteresting functions is simply not a worthwhile undertaking.
2. While for scalar equations it might be argued that there is some reason for being able to quickly bang out particular solutions for specific pretty uninteresting functions—see, particular, the notion of "step response" in Example 4.7.17 and the notion of "frequency response" in Example 4.7.18—for systems of equations the benefit of this is not so clear, given the complexity of doing computation in any example. •

Exercises

5.7.1

Section 5.8

Laplace transform methods for systems of ordinary difference equations

In this section we consider an application of the causal DLT to systems of ordinary difference equations. As with our consideration of scalar equations in Section 4.8, we work with linear constant coefficient equations, both homogeneous and inhomogeneous.

Do I need to read this section? Like Section 4.8, one might skip this chapter at a first reading, until one is confronted with the transfer function methods of Chapter 7, and the use of the tool of the causal DLT makes more sense. •

5.8.1 Systems of homogeneous equations

Now we turn to studying systems of equations using the causal DLT, starting with the homogeneous case. As we did in Section 5.6, we shall work with systems whose state space is a finite-dimensional $\mathbb{R}$-vector space V. We refer to Section IV-9.2.7 for a discussion of how the causal DLT work in this setting.

We consider a system of linear ordinary difference equations F with constant coefficients in an n-dimensional $\mathbb{R}$-vector space V, and with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{Z}_{\geq 0}(h) \times \mathsf{V} &\to \mathsf{V}\\ x &\mapsto \mathsf{A}(x)\end{aligned}$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$. The associated initial value problem we study is then

$$\xi(t+h) = \mathsf{A}(\xi(t)), \quad \xi(0) = x_0. \tag{5.30}$$

Let us take the causal DLT of this initial value problem.

5.8.1 Proposition (Causal DLT of system of homogeneous equations) *The causal DLT of the solution of the initial value problem (5.30) is*

$$\mathscr{L}^1_{\mathrm{D}}(\xi)(\mathrm{z}) = \mathrm{hz}(\mathrm{z}\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}\mathrm{x}_0,$$

and $\mathscr{L}^1_{\mathrm{D}}(\xi)$ is defined on

$$\{\mathrm{z} \in \mathbb{C} \mid |\mathrm{z}| > |\lambda| \text{ for all } \lambda \in \mathrm{spec}(\mathsf{A})\}.$$

Proof This is a direct computation using Proposition IV-9.2.16:

$$z\mathscr{L}^1_{\mathrm{D}}(\xi)(z) - hz\xi(0) = \mathsf{A}\mathscr{L}^1_{\mathrm{D}}(\xi)(z),$$

from which the result follows immediately after noting that $z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A}$ is invertible if $|z|$ exceeds the magnitude of any eigenvalue of A. ■

As with scalar equations, the application of the causal DLT permits a solution for systems of linear homogeneous equations with constant coefficients using just algebraic computations in the transformed variables. In order to understand the inverse $(z\,\mathrm{id}_V - A)^{-1}$, first note that the comments preceding the statement of Proposition 5.4.2 are valid here for computing this inverse. Moreover, the inverse causal DLT of $(z\,\mathrm{id}_V - A)^{-1}$ is known to us already.

5.8.2 Proposition (Causal DLT of operator power function) *For an* n*-dimensional* $\mathbb{R}$*-vector space* V *and for* $A \in \mathrm{L}(V;V)$*, denote*

$$\begin{aligned} P_A \colon \mathbb{Z}_{\geq 0}(h) &\to \mathrm{L}(V;V) \\ kh &\mapsto A^k t^k. \end{aligned}$$

Then $\mathscr{L}^1_{\mathrm{D}}(P_A)(z) = hz(z\,\mathrm{id}_V - A)^{-1}$.

Proof By Theorem 5.6.6(i) and since $P_A(kh) = \Phi^{\mathrm{d}}_A(kh, 0)$, we note that P_A satisfies the initial value problem

$$P_A(kh + h) = A \circ P_A(kh), \quad P_A(0) = \mathrm{id}_V\,.$$

Taking the causal DLT of this initial value problem gives

$$z\,\mathscr{L}^1_{\mathrm{D}}(P_A)(z) - hz\,\mathrm{id}_V = A \circ \mathscr{L}^1_{\mathrm{D}}(P_A)(z) \quad \Longrightarrow \quad \mathscr{L}^1_{\mathrm{D}}(P_A)(z) = hz(z\,\mathrm{id}_V - A)^{-1},$$

as claimed. ■

Let's illustrate this in a simple example.

5.8.3 Example (Operator power function via the causal DLT)

It is a matter of taste whether one thinks that using the causal DLT to compute the operator exponential is preferable to Procedure 5.6.22. It is, however, not such an important matter to resolve in favour of one method or the other; actually computing the operator power function is seldom of interest *per se*. What is certainly true is that with the causal DLT one loses the insight offered by invariant subspaces in Procedure 5.6.22. The benefits of the causal DLT in this context arises in system theory, where complex function techniques offer some genuine insights.

5.8.2 Systems of inhomogeneous equations

Next we consider systems of homogeneous equations. Thus we have an ordinary difference equation with state space V and with right-hand side

$$\begin{aligned} \widehat{F} \colon \mathbb{Z}_{\geq 0}(h) \times V &\to V \\ x &\mapsto A(x) + b(t), \end{aligned} \tag{5.31}$$

for $A \in \mathrm{L}(V;V)$ and for $b \colon \mathbb{Z}_{\geq 0}(h) \to V$. The associated initial value problem we consider is

$$\xi(t + h) = A(\xi(t)) + b(t), \quad \xi(0) = x_0. \tag{5.32}$$

We can, of course, easily take the causal DLT of this initial value problem to get the following.

5.8.4 Proposition (Causal DLT of system of inhomogeneous equations) *Consider the system of scalar ordinary difference equations with right-hand side (5.31), and suppose that* b *is continuous and satisfies* $\mathsf{b} \in \mathsf{LT}^{1,+}(\mathbb{Z}_{\geq 0}(\mathrm{h}); \mathsf{V})$. *The causal DLT of the solution of the initial value problem (5.32) satisfies*

$$\mathscr{L}^1_{\mathrm{D}}(\xi)(\mathrm{z}) = (\mathrm{z}\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}(\mathrm{hz}x_0 + \mathscr{L}^1_{\mathrm{D}}(\mathrm{b})(\mathrm{z})).$$

Proof The proof is an easy adaptation of that of Proposition 5.8.1. ■

As was the case with our discussion of scalar inhomogeneous equations in Section 4.8.2, the preceding result can be interpreted in two ways, one having theoretical value and the other as a means of computing solutions. We shall explore both.

The first result makes a connection with the formula given in Corollary 5.7.2 for solutions to systems of linear inhomogeneous equations, in the general setting of time-varying systems.

5.8.5 Proposition (Causal DLT and convolution for solutions of linear inhomogeneous equations) *Consider the system of scalar ordinary difference equations with right-hand side (5.31), and suppose that* $\mathsf{b} \in \mathsf{LT}^{1,+}(\mathbb{Z}_{\geq 0}(\mathrm{h}); \mathsf{V})$. *Then the solution to the initial value problem (5.32) is*

$$\xi(\mathrm{kh}) = \mathsf{A}^{\mathrm{k}}(x_0) + \mathsf{P}_{\mathsf{A}} * \mathrm{b}((\mathrm{k}-1)\mathrm{h}).$$

Proof From Corollary 5.7.2 we have

$$\xi(kh) = \mathsf{P}_{\mathsf{A}}(kh)(x_0) + \sum_{j=0}^{(k-1)h} \mathsf{P}_{\mathsf{A}}((k-j-1)h)(b(t_0 + jh)), \qquad k \in \mathbb{Z}_{\geq 0},$$

and the result follows from this formula since

$$\mathsf{P}_{\mathsf{A}} * b(kh) = h \sum_{j=0}^{k} \mathsf{A}^{(k-j)h}(b(jh)).$$

However, here we shall give a proof using the causal DLT, valid when $b \in \mathsf{LT}^{1,+}(\mathbb{Z}_{\geq 0}(h); \mathsf{V})$.

From Proposition 5.8.4 we have

$$\mathscr{L}^1_{\mathrm{D}}(\xi)(z) = hz(z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}(x_0) + (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}\mathscr{L}^1_{\mathrm{D}}(b)(z).$$

By Proposition 5.8.2 we have

$$hz(z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1} = \mathscr{L}^1_{\mathrm{D}}(\mathsf{P}_{\mathsf{A}})(z).$$

For $x \in \mathsf{V}$, let us denote

$$\begin{aligned} \mathrm{ev}_x \colon \mathrm{L}(\mathsf{V}; \mathsf{V}) &\to \mathsf{V} \\ \mathsf{A} &\mapsto \mathsf{A}(x). \end{aligned}$$

We then have, noting that ev_{x_0} is a linear map,

$$\mathscr{L}^1_{\mathrm{D}}(\mathrm{ev}_{x_0} \circ \mathsf{P}_{\mathsf{A}})(z) = \mathrm{ev}_{x_0} \circ \mathscr{L}^1_{\mathrm{D}}(\mathsf{P}_{\mathsf{A}})(z) = hz(z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})(x_0).$$

Also, by Exercise IV-9.2.4 and Propositions IV-9.2.9 and IV-9.2.10,

$$\begin{aligned}\mathscr{L}^1_{\mathrm{D}}(\tau^*_h(\mathsf{P}_{\mathsf{A}} * b))(z) &= z^{-1}\mathscr{L}^1_{\mathrm{D}}(\mathsf{P}_{\mathsf{A}} * b)(z)\\ &= z^{-1}\mathscr{L}^1_{\mathrm{D}}(\mathsf{P}_{\mathsf{A}})(z)\mathscr{L}^1_{\mathrm{D}}(b)(z)\\ &= (z\,\mathrm{id}_{\mathsf{V}} - \mathsf{A})^{-1}\mathscr{L}^1_{\mathrm{D}}(b)(z).\end{aligned}$$

Therefore,

$$\mathscr{L}^1_{\mathrm{D}}(\xi)(z) = \mathrm{ev}_{x_0} \circ \mathscr{L}^1_{\mathrm{D}}(\mathsf{P}_{\mathsf{A}}) + \mathscr{L}^1_{\mathrm{D}}(\tau^*_h(\mathsf{P}_{\mathsf{A}} * b))(z).$$

Taking the inverse causal DLT gives

$$\xi(k\Delta) = \mathrm{ev}_{x_0} \circ \mathsf{P}_{\mathsf{A}}(k) + \mathsf{P}_{\mathsf{A}} * b((k-1)h) = \mathsf{A}^k(x_0) + \mathsf{P}_{\mathsf{A}} * b((k-1)h),$$

as claimed. ■

Finally, in the case when b is an also pretty interesting function (meaning that, in a basis for V, the components of b are also pretty uninteresting functions), we can use Proposition 5.8.4, and partial fraction expansions, to compute solutions. We only validate this by a simple example since, in reality, this is not something one ever does.

5.8.6 Example (Solving systems of inhomogeneous equations using the causal DLT)

As with systems of homogeneous equations, the use of the causal DLT to solve inhomogeneous equations does not have a lot to recommend it from a computational point of view. The advantages it has come more from exploiting the algebraic structure of the difference equation as a function of the transformed independent variable z.

Exercises

5.8.1

Section 5.9

Phase-plane analysis for difference equations

Section 5.10

The relationship between differential and difference equations

It will not take an overly astute reader to see connections between differential and difference equations. It is possible to develop these connections in a fairly general setting. However, to do so, one must either (1) be a little vague and imprecise or (2) develop some complicated notation to state the connections clearly. We shall adopt a compromise, developing the connections clearly, but only in the limited setting of systems of homogeneous linear differential and difference equations. We then mention how these precise ideas can be adapted to more general settings.

Do I need to read this section? The material in this section is, we hope, interesting, but is not really required to understand any results that follow. •

5.10.1 From systems to linear homogeneous ordinary differential equations to systems of linear homogeneous ordinary difference equations

We start by establishing the connection from differential to difference equations in the special setting of systems of linear homogeneous ordinary equations. Thus let F be a system of homogeneous linear ordinary differential equations with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V}\\ (t, x) &\mapsto \mathsf{A}(t)(x).\end{aligned}$$

Suppose that $h \in \mathbb{R}_{>0}$ and let

$$\mathbb{T}_{\text{disc}} = \{jh \mid \ jh \in \mathbb{T}\}.$$

We define the ***discretisation*** of F to be the system of homogeneous linear ordinary difference equations F_{disc} with right-hand side

$$\begin{aligned}\widehat{F}_{\text{disc}}\colon \mathbb{T}_{\text{disc}} \times \mathsf{V} &\to \mathsf{V}\\ (t, x) &\mapsto \mathsf{A}_{\text{disc}}(t)(x),\end{aligned}$$

where

$$\mathsf{A}_{\text{disc}}(t) = \Phi^{\text{c}}_{\mathsf{A}}(t + h, t).$$

Then, for $t = jh$ and $t_0 = j_0 h$, we have

$$\begin{aligned}\Phi^{\text{c}}_{\mathsf{A}}(t, t_0) &= \Phi^{\text{c}}_{\mathsf{A}}(t, t - h) \circ \cdots \Phi^{\text{c}}_{\mathsf{A}}(t_0 + h, t_0)\\ &= \mathsf{A}_{\text{disc}}(t - h) \circ \cdots \circ \mathsf{A}_{\text{disc}}(t_0) = \Phi^{\text{d}}_{\mathsf{A}_{\text{disc}}}(t, t_0).\end{aligned}$$

This, then, establishes a precise link from differential to difference equations. We see that the rôle of "A" is quite different in the case of differential and difference equations. For difference equations, A_{disc} behaves like a continuous-time state transition map; indeed, the discrete-time state transition map is merely the composition of the A_{disc}'s.

The discussion becomes particularly interesting and easy to understand when F is a system of linear homogeneous ordinary differential equations with constant coefficients with right-hand side

$$\widehat{F}(t, x) = \mathsf{A}(x)$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$. In this case,

$$\Phi^{\mathrm{c}}_{\mathsf{A}}(t + h, t) = \mathrm{e}^{\mathsf{A}((t+h)-t)} = \mathrm{e}^{\mathsf{A}h}.$$

Thus $\mathsf{A}_{\text{disc}} = \mathrm{e}^{\mathsf{A}h}$.

5.10.2 From systems to linear homogeneous ordinary difference equations to systems of linear homogeneous ordinary differential equations

We next consider reversing the direction of the construction in the preceding section. Thus we start now with a system of linear homogeneous ordinary difference equations F with right-hand side

$$\begin{aligned} \widehat{F}\colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V} \\ (t, x) &\mapsto \mathsf{A}(t)(x), \end{aligned}$$

where $\mathbb{T} \subseteq \mathbb{Z}(h)$ is a discrete time-domain. We then ask the question, "Does there exist a system of linear homogeneous ordinary differential equations F_{cont} with right-hand side

$$\begin{aligned} \widehat{F}_{\text{cont}}\colon \mathbb{T}_{\text{cont}} \times \mathsf{V} &\to \mathsf{V} \\ (t, x) &\mapsto \mathsf{A}_{\text{cont}}(t)(x), \end{aligned}$$

where $\mathbb{T}$ is an interval, and for which

$$\Phi^{\mathrm{c}}_{\mathsf{A}_{\text{cont}}}(t, t_0) = \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0), \qquad t, t_0 \in \mathbb{T}?"$$

Unlike the situation in the preceding section, where we could always associate a difference equation to a differential equation, the answer to the question we ask here is, "Generally, no." Let us understand some of the difficulties of making this transition.

1. First of all, one should think a little about what the continuous time-domain $\mathbb{T}_{\text{cont}}$ should be. The natural choice is to take $\mathbb{T}_{\text{cont}}$ to be the smallest interval containing the discrete time-domain $\mathbb{T}$.

2. By Theorem 5.2.6(v), the discretisation of a system of linear homogeneous ordinary differential equations is an *invertible* system of linear homogeneous ordinary difference equations. Thus a partial answer to the question asked is, "No, if F is not invertible."
3. Even if F is invertible, it is possible that the answer to the question is, "No." We have seen an instance of why this must be so when we considered first-order scalar linear ordinary differential and difference equations in Examples 4.2.21 and 4.6.20. There we saw that the difference equation version of these two sorts of equations exhibited oscillatory behaviour that is not possible for the differential equation version, even when the difference equation was invertible. The reason for this is that the continuous-time state transition map has more properties than merely being invertible. Indeed, we have the following result.

1 Lemma *For a system of linear homogeneous ordinary differential equations* F *with right-hand side* $\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}) = \mathsf{A}(\mathrm{t})(\mathrm{x})$, $\det \Phi^{\mathrm{c}}_{\mathsf{A}}(\mathrm{t}, \mathrm{t}_0) > 0$ *for all* $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}$.

Proof This follows from Theorem 5.2.6(iii). ▼

Thus we can refine our answer to the question we are asking to be, "No, if F is such that $\det \mathsf{A}(t) \leq 0$ for some $t \in \mathbb{T}$."

4. Even when $\det \mathsf{A}(t) > 0$ for every $t \in \mathbb{T}$, it may be the case that the answer to the question is, "No." Let us illustrate with an example.
Take $\mathsf{V} = \mathbb{R}^2$ and let A be defined by the 2×2 matrix

$$\begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}.$$

Note that $\det \mathsf{A} = 2 > 0$. However, $\boldsymbol{A}$ is not given by $\mathsf{A} = \mathrm{e}^{\mathsf{L}}$ for any $\mathsf{L} \in \mathrm{L}(\mathbb{R}^2; \mathbb{R}^2)$. To see this, we consider the three possible Jordan normal forms for L:

$$\begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}, \quad \begin{bmatrix} \alpha & 1 \\ 0 & \alpha \end{bmatrix}, \quad \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}.$$

In the case that $\mathrm{e}^{\mathsf{L}} = \mathsf{A}$, the corresponding operator exponentials are

$$\begin{bmatrix} \mathrm{e}^{\alpha} & 0 \\ 0 & \mathrm{e}^{\beta} \end{bmatrix}, \quad \begin{bmatrix} \mathrm{e}^{\alpha} & t \\ 0 & \mathrm{e}^{\alpha} \end{bmatrix}, \quad \mathrm{e}^{\sigma} \begin{bmatrix} \cos(\omega) & \sin(\omega) \\ -\sin(\omega) & \cos(\omega) \end{bmatrix}.$$

We claim that none of these can be similar to A. Indeed, in the first two cases, the trace of the operator exponential is positive, while the trace of A is negative. In the third case, the eigenvalues of the operator exponential are $\mathrm{e}^{\sigma \pm \mathrm{i}\omega}$ while the eigenvalues for A are -1 and -2. Thus A is not in the image of the operator exponential. Note, as a side treat, that the curve

$$[0, 1] \ni t \mapsto \begin{bmatrix} \cos(\pi t) & \sin(\pi t) \\ -\sin(\pi t) & (1 + t)\cos(\pi t) \end{bmatrix}$$

is a continuous curve connecting id_V to A, so A is in the connected component of the identity.

The preceding example shows that the operator exponential is not surjective onto the space of invertible linear mappings with positive determinant. Interestingly, this characterisation changes when we work with $\mathbb{C}$-vector spaces, and let us record this in the following lemma.

2 Lemma *If V is a finite-dimensional $\mathbb{C}$-vector space and if $E \in \mathrm{L}(V;V)$ is invertible, then there exists $A \in \mathrm{L}(V;V)$ such that $\mathrm{e}^A = E$.*

Proof Let us decompose V into a direct sum of the generalised eigenspaces $\overline{W}(\lambda, E)$ of E. Each eigenspace corresponds to an eigenvalue $\lambda \in \mathbb{C}$ that is necessarily nonzero. It suffices to show that, for each eigenvalue λ, there exists

$$A_\lambda \in \mathrm{L}(\overline{W}(\lambda, E); \overline{W}(\lambda, E))$$

such that $\mathrm{e}^{A_\lambda} = E|\overline{W}(\lambda, E)$. That is to say, we can suppose that E has one generalised eigenspace for a nonzero eigenvalue λ, which means that $N = E - \lambda \mathrm{id}_V$ is nilpotent by Theorems I-5.8.69 and I-5.8.56. Thus we can write

$$E = \lambda \mathrm{id}_V + N = \lambda \mathrm{id}_V(\mathrm{id}_V - (-\lambda^{-1}N)) = \lambda \mathrm{id}_V(\mathrm{id}_V - \hat{N}),$$

where $\hat{N} = -\lambda^{-1}N$. Note that, if $\mathrm{e}^\ell = \lambda$, then, taking $A_0 = \ell \mathrm{id}_V$, we have

$$\mathrm{e}^{A_0} = \mathrm{e}^\ell \mathrm{id}_V = \lambda \mathrm{id}_V,$$

by Theorem 5.2.20(v). Now note that the Taylor series for the logarithm is

$$\log(1 - a) = -\sum_{j=1}^{\infty} \frac{a^j}{j}$$

for $a \in \mathbb{C}$. Thus $\exp \circ \log(1 - a) = 1 - a$. One can use this to see that, if $\hat{N}^k = 0$,

$$\exp\left(-\sum_{j=1}^{k-1} \frac{\hat{N}^j}{j}\right) = \exp\left(-\sum_{j=1}^{\infty} \frac{\hat{N}^j}{j}\right) = \\ = \exp \log(\mathrm{id}_V - \hat{N}) = \mathrm{id}_V - \hat{N}.$$

Therefore, if we take

$$A_1 = -\sum_{j=1}^{k-1} \frac{\hat{N}^j}{j},$$

then we have $\mathrm{e}^{A_1} = \mathrm{id}_V - \hat{N}$. Finally, since A_0 and A_1 commute and by Theorem 5.2.20(vi),

$$\mathrm{e}^{A_0 + A_1} = \mathrm{e}^{A_0} \circ \mathrm{e}^{A_1} = \lambda \mathrm{id}_V(\mathrm{id}_V - \hat{N}) = E,$$

as desired. ▼

Even if the answer to the question is, "Yes," for a given F, there are still some caveats of which one must be aware.

5. There may be many differential equations with the same discretisation. For example, if we take $\mathsf{V} = \mathbb{R}^2$, $\mathsf{A} = \mathrm{id}_{\mathbb{R}^2}$, and $h = 1$, then the constant coefficient differential equations with

$$\mathsf{A}_{\text{cont}} \in \left\{ \begin{bmatrix} 0 & 2n\pi \\ -2n\pi & 0 \end{bmatrix} \;\middle|\; n \in \mathbb{Z} \right\}$$

all give A as their discretisation. Essentially, because the difference equation samples at a certain interval, anything happening between the sampling times is not captured by the difference equation. In the example given, there is high-frequency behaviour in the continuous-time system that cannot be represented by the discrete-time system.

5.10.3 Generalisation to not necessarily linear ordinary differential equations

We close this section by saying a few words about extending the ideas in the preceding two sections to general ordinary differential and difference equations.

First let us consider the situation where we are given an ordinary differential equation $\boldsymbol{F}$ with right-hand side

$$\widehat{\boldsymbol{F}} \colon \mathbb{T} \times U \to \mathbb{R}^n.$$

This will have a flow $\Phi^F \colon D_F \to U$, and one might want to try to define its discretisation, with sampling interval $h \in \mathbb{R}_{>0}$, by "sampling" the flow:

$$\widehat{\boldsymbol{F}}_{\text{disc}}(t, \boldsymbol{x}) = \Phi^F(t + h, t, \boldsymbol{x}),$$

adapting what we did in the linear case. One then gets, just as in the linear case,

$$\Phi^{F_{\text{disc}}}(t, t_0, \boldsymbol{x}) = \Phi^F(t, t_0, \boldsymbol{x}),$$

but now with a caveat: $\Phi^F(t + h, t, \boldsymbol{x})$ may not exist for all $(t, \boldsymbol{x}) \in \mathbb{T}_{\text{disc}} \times U$, cf. Example 3.2.5. In this sense, the discretisation may fail to exist, in general.

The matter of going from a difference equation to a differential equation has the same perils as have already been encountered for linear equations, of course.

Section 5.11

Using a computer to work with systems of ordinary differential equations

We thank Jack Horn for putting together the MATHEMATICA® *and* MATLAB® *results in this section.*

In this section we illustrate how to use computer packages to obtain analytical and numerical solutions for systems of ordinary differential equations. We restrict our attention to attention to linear equations with constant coefficients, since these are really the only significant class of equations that one can work with analytically. For numerical solutions, the techniques here are extended in the obvious way to nonlinear or time-varying systems. As in Section 4.9, we restrict our attention to illustrating the use of MATHEMATICA® and MATLAB®.

5.11.1 Using MATHEMATICA® to obtain analytical and/or numerical solutions

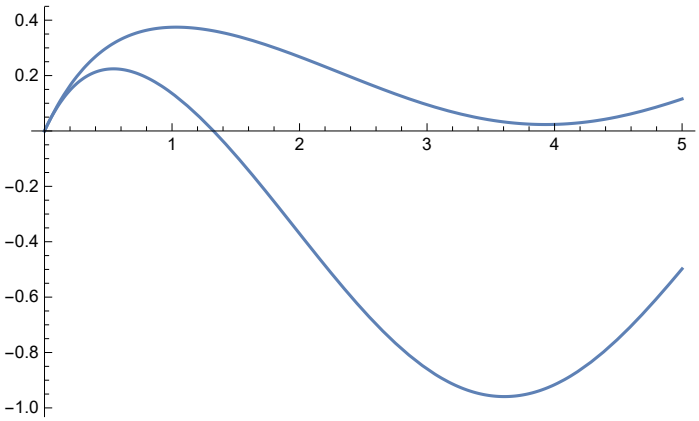
Solving systems of differential equations in MATHEMATICA® requires a similar procedure as solving a single ordinary differential equation. You must use the `DSolve` command, while keeping your system in the form $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{f}(t)$, for a given matrix A and vector function $\boldsymbol{f}$.

5.11.1 Example (Using `DSolve` to solve systems of ordinary differential equations) The first system we will consider is:

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t}(t) = \begin{bmatrix} -1 & -2 \\ 1 & -3 \end{bmatrix} \boldsymbol{y}(t) + \begin{bmatrix} \cos(t) \\ 1 \end{bmatrix}$$

The following script will find and plot the solutions to this system.

```
A = {{-1, -2}, {1, -3}};

Y[t_] = {y1[t], y2[t]};

solution = DSolve[{Y'[t] == A.Y[t] + {Cos[t], 1}, Y[0] == {0, 0}}, {y1, y2}, t];

Plot[{y1[t], y2[t]}/.solution, {t, 0, 5}]
```



Note that the "." in MATHEMATICA® means matrix-vector multiplication in the above code. •

5.11.2 Example (Matrix exponential in MATHEMATICA®) MATHEMATICA® is also an incredibly handy software for various aspects of linear algebra. In this example we will work with the matrix

$$A = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

and will compute matrix exponentials, first using the `MatrixExp` command, then by following the process in Procedure 5.2.26.

A = {{−1, 1, 0}, {−1, −1, 0}, {0, 0, 2}};

MatrixExp[t * A]//MatrixForm

$$\begin{pmatrix} e^{-t}\text{Cos}[t] & e^{-t}\text{Sin}[t] & 0 \\ -e^{-t}\text{Sin}[t] & e^{-t}\text{Cos}[t] & 0 \\ 0 & 0 & e^{2t} \end{pmatrix}$$

Now we will follow the steps from class, and compare the results.

Eigenvals = Eigenvalues[A];

Eigenvect = Eigenvectors[A];

F1 = Exp[t * Eigenvals[[1]]] * Eigenvect[[1]];

F2 = Exp[t * Eigenvals[[2]]] * Eigenvect[[2]];

F3 = Exp[t * Eigenvals[[3]]] * Eigenvect[[3]];

Fund = Transpose[{F1, F2, F3}];

FundInv = Inverse[Fund];

B = FundInv/.$t \to 0$;

Indirect = Fund.B//MatrixForm

This "indirect" method gives us the ugly looking matrix shown below:

$$\begin{pmatrix} \frac{1}{2}e^{(-1-i)t} + \frac{1}{2}e^{(-1+i)t} & \frac{1}{2}ie^{(-1-i)t} - \frac{1}{2}ie^{(-1+i)t} & 0 \\ -\frac{1}{2}ie^{(-1-i)t} + \frac{1}{2}ie^{(-1+i)t} & \frac{1}{2}e^{(-1-i)t} + \frac{1}{2}e^{(-1+i)t} & 0 \\ 0 & 0 & e^{2t} \end{pmatrix}$$

However, this is equivalent to the matrix found by using the `MatrixExp` command, which can be seen by applying the `ComplexExpand` command.

ComplexExpand[Indirect]//MatrixForm

$$\begin{pmatrix} e^{-t}\text{Cos}[t] & e^{-t}\text{Sin}[t] & 0 \\ -e^{-t}\text{Sin}[t] & e^{-t}\text{Cos}[t] & 0 \\ 0 & 0 & e^{2t} \end{pmatrix}$$

Sometimes it is not so easy to see that identical symbolic expressions in MATHEMATICA® are, in fact, identical. For things that are not excessively disgusting to look at, sometimes the `Simplify` command is useful. For complex things, `ComplexExpand` is sometimes useful. •

Next we consider inhomogeneous equations, using Corollary 5.3.3.

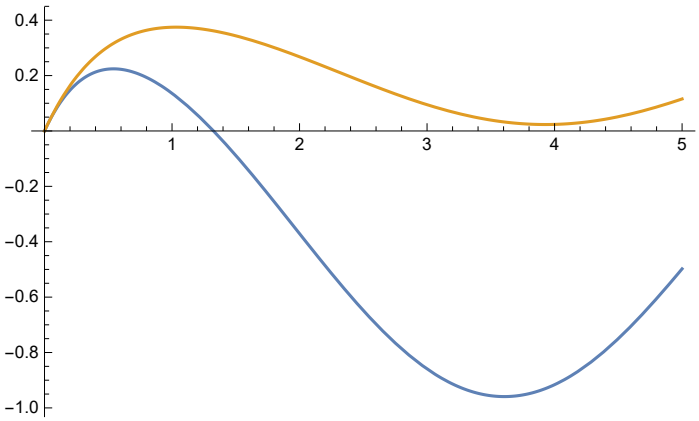
5.11.3 Example (Inhomogeneous linear systems of equations using MATHEMATICA®) Now that we are comfortable with commands such as `MatrixExp`, we will see how it is also possible to solve systems of ordinary differential equations using the formula

$$x(t) = \mathrm{e}^{At}x_0 + \int_0^{t_0} \mathrm{e}^{A(t_0-\tau)}f(\tau)\,\mathrm{d}\tau.$$

We will show this by solving the same system given Exercise 5.11.1.

x[t] = MatrixExp[t * A].{0, 0} + Integrate[MatrixExp[A * (t − T)].{Cos[T], 1}, {T, 0, t}];

Plot[x[t], {t, 0, 5}]



As you can see, the plots are identical to the direct results in Exercise 5.11.1. •

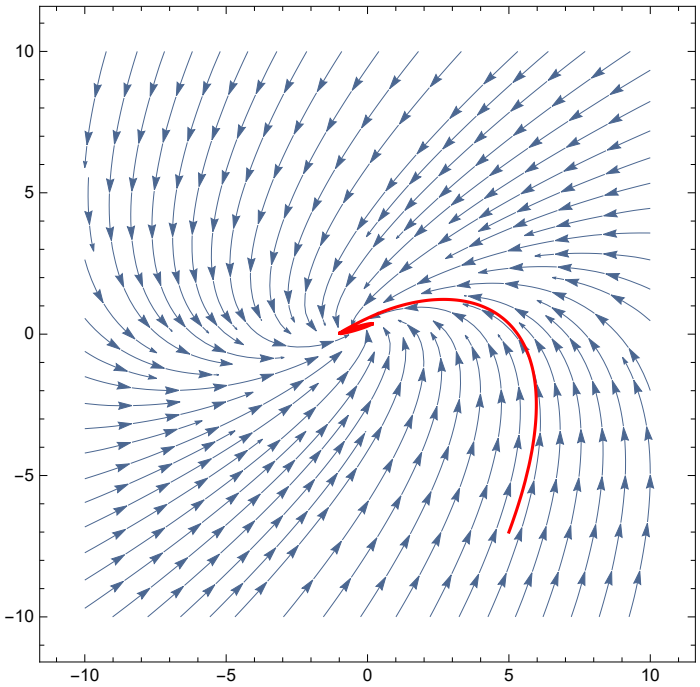
One can also use MATHEMATICA® to produce phase portraits. There are sophisticated MATHEMATICA® packages for doing this (we used DynPac for the plots from Section 5.5), and here we shall indicate how to do this with standard MATHEMATICA® commands.

5.11.4 Example (Phase plane using MATHEMATICA®) We consider the planar system of linear equations

$$\begin{bmatrix}\dot{x}(t)\\ \dot{y}(t)\end{bmatrix} = \begin{bmatrix}-1 & -2\\ 1 & -3\end{bmatrix}\begin{bmatrix}x(t)\\ y(t)\end{bmatrix} + \begin{bmatrix}\cos(t)\\ 1\end{bmatrix}.$$

We use the commands `StreamPlot` and `ParametricPlot`.

```
splot = StreamPlot[{−x − 2y, x − 3y}, {x, −10, 10}, {y, −10, 10}];

Show[splot,

ParametricPlot[

Evaluate[

First[

{x[t], y[t]}/.DSolve[{x'[t] == −x[t] − 2 * y[t] + Cos[t], y'[t] == x[t] − 3 * y[t] + 1,

{x[0], y[0]} == {5, −7}}, {x[t], y[t]}, t]]], {t, 0, 10}, PlotStyle → Red]]
```



We have plotted, using `StreamPlot`, the phase plane for the homogeneous system, and superimposed in red one solution for the inhomogeneous system. •

5.11.2 Using MATLAB® to obtain numerical solutions

In MATLAB®, solving systems of differential equations is not much different than solving a single ordinary differential equation. You must create a function for your system, which must then be passed into a script that will use the `ode45` solver.

5.11.5 Example (Using `ode45` to solve systems of ordinary differential equations) We will once again be considering the same examples as we did in MATHEMATICA®, this time using MATLAB®. First we will solve the following system:

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t}(t) = \begin{bmatrix} -1 & -2 \\ 1 & -3 \end{bmatrix} \boldsymbol{y}(t) + \begin{bmatrix} \cos(t) \\ 1 \end{bmatrix}.$$

```
function [ dydt ] = Example2( t,y )

A = [-1 -2;1 -3];

dydt = A*y + [cos(t); 1];

end
```

Below is the main script that will plot the solution to this system. See Figure 5.9 for the MATLAB® generated plots.

```
clc
clear all
close all
%% Solving Numerically

t  = linspace(0,5);
y0 = [0 0];

y = ode45(@(t,y)Example2(t,y),t,y0);

plot(y.x,y.y)
xlabel('Time [s]');
ylabel('y(t)');
legend('y1(t)','y2(t)');
```

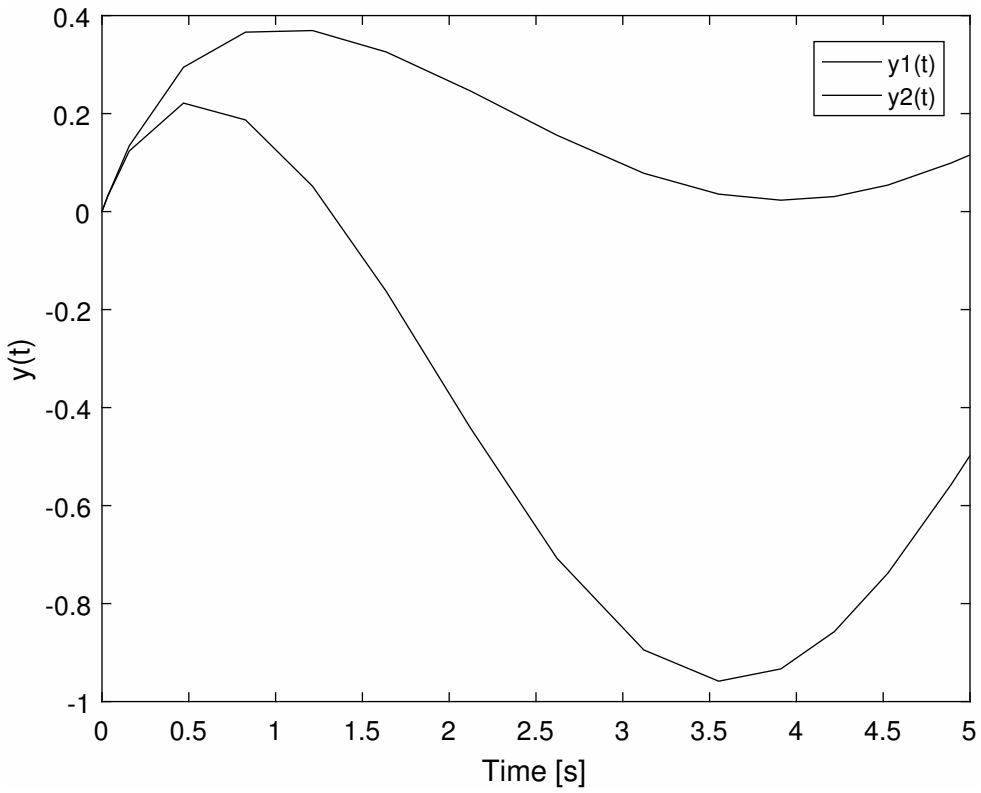


Figure 5.9 Plots generated by MATLAB® for Exercise 5.11.5

One can see that the solutions are quite similar to those from Exercise 5.11.1 using MATHEMATICA®. The jagged character of the plots is indicative of the fact that the time step for `ode45` can be decreased. This can be done by specifying

```
t_int = tinit:tstep:tfinal
```

where the meaning of `tinit`, `tfinal`, and `tstep` is just what you think they are. •

MATLAB® is also very useful for linear algebra.

5.11.6 Example (Matrix exponential in MATLAB®) We will consider the same matrix exponential example

$$A = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

as in Example 5.11.2. Again, it is possible to calculate the matrix exponential both directly (using the `expm` command), or you can follow the steps from Procedure 5.2.26.

```
clc
clear all
close all

%% Calculating Matrix Exponential Directly
A = [-1 1 0; -1 -1 0; 0 0 -2];
syms t
MatrixExpDirect = expm(t*A)
%% Calculating Matrix Exponential Using Procedure from Class

[EigenVectors,EigenValues] = eig(t*A);

F1 = exp(EigenValues(1,1)).*EigenVectors(:,1);
F2 = exp(EigenValues(2,2)).*EigenVectors(:,2);
F3 = exp(EigenValues(3,3)).*EigenVectors(:,3);

Fund = [F1 F2 F3];
FundInv = inv(Fund);
B = subs(FundInv,0); %Here we are evaluating the fundamental
    matrix at t = 0

MatrixExponential = Fund*B
```

Here is the output from the MATLAB® code

```
MatrixExponential =
[exp(t*(-1-1i))/2+exp(t*(-1+1i))/2,
(exp(t*(-1-1i))*1i)/2-(exp(t*(-1+1i))*1i)/2, 0]
[-(exp(t*(-1-1i))*1i)/2+(exp(t*(-1+1i))*1i)/2,
exp(t*(-1-1i))/2+exp(t*(-1+1i))/2, 0]
[0, 0, exp(-2*t)]
```

Of course, the result here is the same as we saw using MATHEMATICA®. •

Finally, let us see how MATLAB® can be used to create phase portraits.

5.11.7 Example To create phase portraits in MATLAB®, you must use the `meshgrid` command, and evaluate the first derivatives of y_1 and y_2 at each point for $t = 0$. Once you have done this, use the `quiver` command to plot the vector field. To plot a specific solution, simply use the `ode45` command, and plot the first and second columns of the outputted matrix. See Figure 5.10 for the result.

```
clc
clear all
close all

[x,y] = meshgrid(-10:1:10,-10:1:10);

u = zeros(size(x));
v = zeros(size(x));

t = 0;
for i = 1:numel(x)
    dydt = Example2(t,[x(i);y(i)]);
    u(i) = dydt(1);
    v(i) = dydt(2);
end

quiver(x,y,u,v,'b');
xlabel('y1(t)');
ylabel('y2(t)');

t = linspace(0,5);
y0 = [5,-7];

hold on
y = ode45(@(t,y)Example2(t,y),t,y0);
plot(y.y(1,:),y.y(2,:))
hold off

print -deps PhasePortrait
```

It is possible to customise MATLAB® output to look prettier, but this is something we leave to the reader as they progress through their professional lives. •

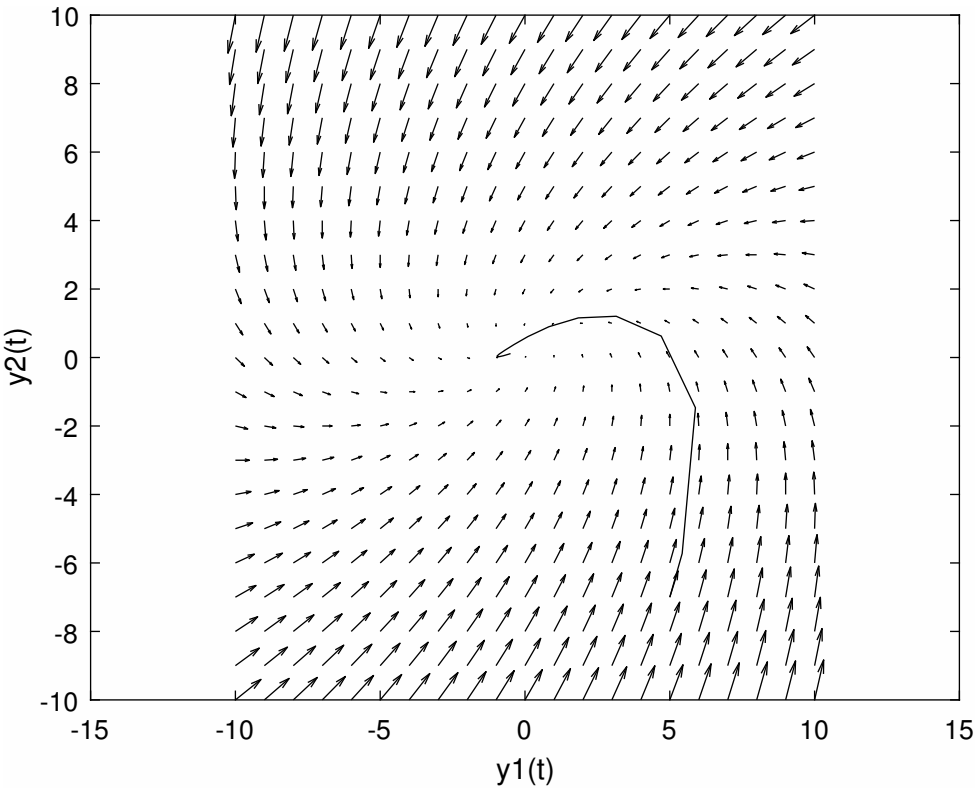


Figure 5.10 Phase portrait generated by MATLAB®

Chapter 6

Classes of continuous- and discrete-time systems

Now that we have carefully understood ordinary differential and difference equations, we turn to systems described by such equations. We shall begin with a presentation of a class of general, not necessarily linear,[1] systems. However, the study of such systems in any substantial way is a specialised subject that itself deserves a volume of material that duplicates in scope, and exceeds in depth, the material presented in these volumes. Instead, we concentrate on—but do not consider exclusively—a detailed presentation of linear system theory. For the purposes of this chapter we mainly present definitions for the classes of systems we shall subsequently consider, and characterise them in terms of the general system descriptors introduced in Chapter 2.

In Sections 6.1–6.4 we consider definitions and general results for not necessarily linear systems. While the class of systems we consider here is general enough to include the linear systems that are treated as part of the classical subject of "signals and systems," it is far from a presentation of a completely general class of systems. Let us list the ways in which the theory we present here is not general.

1. One can consider linear systems with infinite-dimensional state spaces. Examples of such systems include the heat, wave, and potential equations of Sections 1.1.11, 1.1.12, and 1.1.13, all of which are described by partial differential equations. While some of the techniques for linear systems we consider can be adapted to these infinite-dimensional systems, generally speaking the treatment of these systems is substantially more technical and requires a self-contained treatment.
2. One can consider systems with state spaces that are finite-dimensional, but are not subsets of Euclidean space, such as we consider here. Examples of such systems include mechanical systems involving the motion of rigid bodies, as these systems typically include "angle variables," and angles live in circles, not intervals. A treatment of systems such as these involves a detailed study of the

[1] We have elected to banish the commonly used terminology "nonlinear." The reason for doing this is that "nonlinear systems" are simply systems, while "linear systems" warrant the adjective "linear" since they are indeed special among the class of all systems.

state spaces involved, and this is something that will take one on a journey of independent interest.

3. One can consider systems that combine continuous- and discrete-time domains in some way. For example, one might have various continuous-time systems between which one switches according to some discrete-time prescription. These sorts of models are often referred to as "hybrid systems." One must develop techniques for dealing with hybrid systems, and this is largely something that has not been done in a unified and comprehensive way.
4. There are many problems that can be characterised by automata models such as were considered in various places in Chapter 2. For these systems, the techniques and problems typically have a different flavour than what we work with in this volume.

Note that it is not realistic to develop a meaningful system theory that captures the points of interest to all of these sorts of systems, along with the ones we *do* consider. Indeed, the theory of general systems we present in Chapter 2 is an attempt to perform a unification of this sort, and it suffers from the lack of depth one often sees with work that is "too general."

The theory we begin to investigate in this chapter weaves together much of the mathematics we have developed in previous volumes, namely convolution (Chapter IV-4), transform theory (Chapters IV-6, IV-7, and IV-9), and distribution theory (Chapter IV-3). Indeed, the theory of linear systems we present here is the *raison d'être* for our presentation of this background material (although it is certainly the case that the existence of all of this material has justification beyond what we do here).

Do I need to read this chapter? The material in this chapter is the beginning of the core of the material in this volume. •

Contents

Section 6.1

Continuous-time state space systems

Much of the system theoretic methodology in this volume relates exclusively to what we shall call "linear systems." These techniques are not generally applicable to systems that do not have the property of linearity. However, even if one is only ultimately interested in linear systems, it is valuable for context to begin the discussion with a more general classes of systems, of which linear systems are a particular instance. In this section we introduce the first of these larger classes. Consistent with our discussion of background material in differential and difference equations, we shall give equal footing to both continuous-time and discrete-time systems. Here we consider continuous-time systems.

Do I need to read this section? If you have any interest at all in the existence of systems that are not linear, or in understanding the context for linear systems, then this section is essential. •

6.1.1 Definitions and system theoretic properties

Let us introduce the basic object of study, recalling from Section 2.2.2 the notation concerning partially defined functions on time-domains.

6.1.1 Definition (Continuous-time state space system) A *continuous-time state space system* is a sextuple $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$, where

(i) $X \subseteq \mathbb{R}^n$ is an open set (the *state space*),

(ii) $U \subseteq \mathbb{R}^m$ (the *control set*),

(iii) $\mathbb{T} \subseteq \mathbb{R}$ is an interval (the *time-domain*),

(iv) $\mathscr{U} \subseteq U^{(\mathbb{T})}$ (the *control functions* or *controls*),

(v) $f\colon \mathbb{T} \times X \times U \to \mathbb{R}^n$ (the *dynamics*), and

(vi) $h\colon \mathbb{T} \times X \times U \to \mathbb{R}^k$ (the *output map*).

Associated with a continuous-time state space system Σ we have the following notions:

(vii) a *controlled trajectory* for Σ is a pair (ξ, μ), where $\mu \in \mathscr{U}$ and where $\xi \in \mathsf{AC}_{\mathrm{loc}}(\mathrm{dom}(\mu); X)$ are such that

$$\dot{\xi}(t) = f(t, \xi(t), \mu(t)), \qquad \text{a.e. } t \in \mathrm{dom}(\mu); \tag{6.1}$$

(viii) a *controlled output* for Σ is a pair (η, μ), where $\mu \in \mathscr{U}$ and where $\eta\colon \mathrm{dom}(\mu) \to \mathbb{R}^k$ satisfies

$$\eta(t) = h(t, \xi(t), \mu(t)), \qquad t \in \mathrm{dom}(\mu),$$

for some controlled trajectory (ξ, μ).

We denote by $\mathrm{Ctraj}(\Sigma)$ the set of controlled trajectories and by $\mathrm{Cout}(\Sigma)$ the set of controlled outputs. •

Of course, since controlled trajectories are defined by solutions to a differential equation, there are conditions one must give to the dynamics f to ensure that $\mathrm{Ctraj}(\Sigma) \neq \emptyset$. Such conditions will be given in the next section. Here we shall consider the system theoretic attributes of continuous-time state space systems; that is, we make reference to the general system theory of Chapter 2, and see which attributes apply to the systems of this section. In doing this, we will not consider the logical interrelations between the various notions, since part of the point of the discussion here is to see how one applies the definitions of Chapter 2.

Let us consider a few attributes of continuous-time state space systems that often arise in practice.

6.1.2 Definition (Autonomous, proper continuous-time state space systems) A continuous-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ is

(i) ***autonomous*** if there exists

$$f_0 \colon X \times U \to \mathbb{R}^n, \quad h_0 \colon X \times U \to \mathbb{R}^k$$

such that

$$f(t, x, u) = f_0(x, u), \quad h(t, x, u) = h_0(x, u)$$

for every $(t, x, u) \in \mathbb{T} \times X \times U$, and is

(ii) ***proper*** if there exists $h_0 \colon X \times U \to \mathbb{R}^k$ such that

$$h(t, x, u) = h_0(t, x)$$

for every $(t, x, u) \in \mathbb{T} \times X \times U$. •

If only f (resp. h) satisfies the conditions for the system to be autonomous, we shall say that the system is ***dynamically autonomous*** (resp. ***output autonomous***).

We shall see the system theoretic significance of these notions shortly.

Indeed, we next indicate whether/how a continuous-time state space system is a system of the various types introduced in Chapter 2.

6.1.3 Remarks (Continuous-time state space systems as general systems) We let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a continuous-time state space system.

1. A continuous-time state space system is a general input/output system as per Definition 2.1.3. To see this, take
   (a) "$\mathscr{U} = \mathscr{U}$," i.e., the inputs for the general input/output system are the same as the controls for the continuous-time state space system,
   (b) $\mathscr{Y} = (\mathbb{R}^k)^{(\mathbb{T})}$, i.e., the outputs for the general input/output system are the partial $\mathbb{R}^k$-valued functions on $\mathbb{T}$, and

   (c) $\mathscr{B} = \mathrm{Cout}(\Sigma)$, i.e., the behaviours for the general input/output system are exactly the controlled outputs for the continuous-time state space system.

2. A continuous-time state space system is, more specifically, a general time system as per Definition 2.2.9. To see this, take
   (a) "$U = U$," i.e., the input set for the general time system is the same as the control set for the continuous-time state space system,
   (b) $Y = \mathbb{R}^k$, i.e., the output set for the general time system is $\mathbb{R}^k$,
   (c) "$\mathscr{U} = \mathscr{U}$," i.e., the admissible input signals for the general input/output system are the same as the controls for the continuous-time state space system,
   (d) $\mathscr{Y} = (\mathbb{R}^k)^{(\mathbb{T})}$, i.e., the admissible output signals for the general input/output system are the partial $\mathbb{R}^k$-valued functions on $\mathbb{T}$, and
   (e) $\mathscr{B} = \mathrm{Cout}(\Sigma)$, i.e., the behaviours for the general time system are exactly the controlled outputs for the continuous-time state space system. •

Next we consider the issue of various forms of completeness for continuous-time systems, as introduced in a general setting in Section 2.2.4.

6.1.4 Remarks (Completeness for continuous-time state space systems) We let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a continuous-time state space system. In our constructions here, we make use of the notation for flows of continuous-time state space systems introduced in Definition 6.1.12.

1. *Continuous-time state space systems are output complete:* Let $\boldsymbol{\mu} \in \mathscr{U}$ and let $(I, \preceq)$ be a totally ordered set, and let $(\boldsymbol{\eta}_i)_{i\in I}$ be a family of outputs satisfying conditions (a)–(f) of Definition 2.2.12. Note that

$$\boldsymbol{\eta}_i(t) = \boldsymbol{h}(t, \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(t)), \qquad t \in \mathrm{dom}(\boldsymbol{\eta}_i).$$

   Now let $\mathbb{S} = \cup_{i\in I} \mathrm{dom}(\boldsymbol{\eta}_i)$ and let $\boldsymbol{\eta} \colon \mathbb{S} \to \mathbb{R}^k$ be such that $\boldsymbol{\eta}_{\mathrm{dom}(\boldsymbol{\eta}_i)} = \boldsymbol{\eta}_i$, $i \in I$. Then, if $t \in \mathbb{S}$, we must have $t \in \mathrm{dom}(\boldsymbol{\eta}_i)$ for some $i \in I$. Therefore,

$$\boldsymbol{\eta}(t) = \boldsymbol{\eta}_i(t) = \boldsymbol{h}(t, \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(t)).$$

   As this holds for every $t \in \mathrm{dom}(\boldsymbol{\eta})$, we conclude output completeness.

2. *Generally, a continuous-time state space system is not complete:* This is exhibited in Example 2.2.21. •

Next let us give the form for the general time system representations of Section 2.2 for continuous-time state space systems.

6.1.5 Remarks (General time system representations for continuous-time state space systems) We let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a continuous-time state space system. In our constructions here, we make use of the notation for flows of continuous-time state space systems introduced in Definition 6.1.12. In the following, we suppose that Σ is complete.

1. Σ *has an initial response function:* Let $t_0 \in \mathbb{T}$ and let $(\eta, \mu) \in \mathrm{Cout}(\Sigma)$ with

$$\eta(t) = h(t, \xi(t), \mu(t)), \qquad t \in \mathrm{dom}(\mu),$$

for $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma)$. Suppose that $t_0 \in \mathrm{dom}(\mu)$. Then

$$\xi(t) = \Phi^{\Sigma}(t, t_0, x_0, \mu)$$

for some $x_0 \in \mathrm{dom}(\mu) \times X$. We can then denote

$$\rho^{\Sigma}_{t_0}(x_0, \mu)(t) = h(t, \Phi^{\Sigma}(t, t_0, x, \mu), \mu(t)),$$

and this defines the initial response function $\rho^{\Sigma}_{t_0}$ from t_0 with initial state object X. One has to verify the conditions of Definition 2.2.14, and this is straightforward. Note that we require completeness in order to ensure the existence of Φ^{Σ} for all arguments.

2. Σ *has a family of state transition maps:* Let $t_0 \in \mathbb{T}$ and, given $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, we take $X_{t_1} = X_{t_2} = X$ and define

$$\Phi_{t_2,t_1}(\mu, x_1) = \Phi^{\Sigma}(t_2, t_1, x_1, \mu),$$

defining the family of state transition maps. The properties of flows enunciated in Proposition 3.2.12 ensure that the conditions of Definition 2.2.15 are satisfied, and we leave the elementary verification of this to the reader. (Indeed, it is the conclusions of Proposition 3.2.12 that explain the conditions of Definition 2.2.15.)

Again, we see that completeness is required.

3. Σ *has a dynamical system representation:* The response function and the family of state transition maps above combine to give a dynamical systems representation at $t_0 \in \mathbb{T}$, as per Definition 2.2.19. One can readily verify the conditions of Definition 2.2.19.

One can show that this dynamical system representation is full. This is a consequence of the invertibility of the flow, as in Proposition 3.2.12(iii).

As in the preceding two items, completeness is obviously required.

4. Σ *has a state space representation:* As output function at $t_0 \in \mathbb{T}$, as per Definition 2.2.24, is simply given by

$$\gamma^{\Sigma}_{t,t_0}(x, u) = h(t, x, u).$$

One readily verifies that the conditions of Definition 2.2.24 are satisfied. •

Now let us see which of the general time system theoretic attributes of Section 2.2 are held by a continuous-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$. In order to make the connections to the general time system notions of Section 2.2 to the specific case here, we give the translation of these notions that into language applicable to the class of system we consider here. In our definitions, we make use of the notation for flows of continuous-time state space systems introduced in Definition 6.1.12. The proofs of the following results are mere applications of the definitions of the symbols involved.

We begin with causality.

6.1.6 Proposition (Causality for continuous-time state space systems) *Let* $\Sigma = (\mathbf{X}, \mathbf{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ *be a continuous-time state space system and let* $t_0 \in \mathbb{T}$.

(i) The system Σ *is* ***causal*** *from* t_0 *if, for every* $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathscr{U}$ *and every* $t \in \mathbb{T}_{\geq t_0} \cap \operatorname{dom}(\boldsymbol{\mu}_1) \cap \operatorname{dom}(\boldsymbol{\mu}_2)$,

$$\mu_1|[t_0, t] = \mu_2|[t_0, t] \implies \mathbf{h}(t, \Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_1)) = \mathbf{h}(t, \Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_2))$$

for every $\mathbf{x}_0 \in \mathbf{X}$.

(ii) The system Σ *is* ***strongly causal*** *from* t_0 *if, for every* $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathscr{U}$ *and every* $t \in \mathbb{T}_{\geq t_0} \cap \operatorname{dom}(\boldsymbol{\mu}_1) \cap \operatorname{dom}(\boldsymbol{\mu}_2)$,

$$\mu_1|[t_0, t) = \mu_2|[t_0, t) \implies \mathbf{h}(t, \Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_1)) = \mathbf{h}(t, \Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_2))$$

for every $\mathbf{x}_0 \in \mathbf{X}$. •

Now we consider stationarity.

6.1.7 Proposition (Stationarity for continuous-time state space systems) *Let* $\Sigma = (\mathbf{X}, \mathbf{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ *be a continuous-time state space system and let* $t_0 \in \mathbb{T}$.

(i) The system Σ *is* ***stationary*** *from* t_0 *if* $\tau^*_{t_0, t_0+a}(\mathscr{U}) \subseteq \mathscr{U}$ *for every* $a \in \mathbb{R}_{>0}$ *and if, for every* $\boldsymbol{\mu} \in \mathscr{U}$ *and every* $t \in \mathbb{T}_{\geq t_0} \cap \operatorname{dom}(\boldsymbol{\mu})$,

$$\mathbf{h}(t + a, \Phi^\Sigma(t + a, t_0 + a, \mathbf{x}_0, \tau^*_{t_0, t_0+a}\boldsymbol{\mu}), \tau^*_{t_0, t_0+a}\boldsymbol{\mu}(t)) = \mathbf{h}(t, \Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(t))$$

for every $a \in \mathbb{R}_{>0}$ *and every* $\mathbf{x}_0 \in \mathbf{X}$.

(ii) The system Σ *is* ***strongly stationary*** *from* t_0 *if it is stationary from* t_0 *and if, for every* $a \in \mathbb{R}_{>0}$, *every* $\mathbf{x}_0 \in \mathbf{X}$, *and every* $\boldsymbol{\mu} \in \mathscr{U}$, *there exists* $\mathbf{x}'_0 \in \mathbf{X}$ *such that*

$$\mathbf{h}(t, \Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(t))\mathbf{h}(t + a, \Phi^\Sigma(t + a, t_0 + a, \mathbf{x}'_0, \tau^*_{t_0, t_0+a}\boldsymbol{\mu}(t)), \tau^*_{t_0, t_0+a}\boldsymbol{\mu}(t)).$$ •

Note that a consequence of this definition of stationarity is that $\sup \mathbb{T} = \infty$.

With these definitions, we have the following statements.

6.1.8 Remarks (System theoretic attributes of continuous-time state space systems) We let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a continuous-time state space system. In our constructions here, we make use of the notation for flows of continuous-time state space systems introduced in Definition 6.1.12.

1. Σ *is causal and sometimes strongly causal:* Let $t_0 \in \mathbb{T}$. It follows from the formula (6.1) (assuming the conditions of Theorem 6.1.10 below) for controlled trajectories that a controlled trajectory (ξ, μ) satisfies

$$\xi(t) = \xi(t_0) + \int_{t_0}^{t} \boldsymbol{f}(\tau, \xi(\tau), \mu(\tau))\, \mathrm{d}\tau, \qquad t \in \mathbb{R}_{\geq t_0} \cap \operatorname{dom}(\mu).$$

   Therefore, if controls μ_1 and μ_2 agree on $[t_0, t]$, then the controlled trajectories for Σ on $[t_0, t]$ agree. Thus Σ is causal from t_0. If, additionally, $\boldsymbol{h}$ is independent of U, i.e., Σ is proper, then we claim that Σ is also strongly causal from t_0. Indeed, from the construction in Theorem 6.1.10 of a controlled trajectory as a solution to an ordinary differential equation, we see that altering a control on a set of measure zero does not change the trajectory. Note, however, that if $\boldsymbol{h}$ *does* depend on control, then we generally have causality, but not strong causality.
2. Σ *is sometimes past determined:* First of all, the definition of being past-determined requires completeness, so one needs to assume completeness to make any statements about past-determinacy. Thus we do this. This ensures that the the first part of the definitions of past-determined and strong past-determined holds.

   Now let $t_0 \in \mathbb{T}$. For the second parts of these definitions, considerations such as those for causality above allow us to conclude that Σ is past-determined from any $\tau \in \mathbb{T}_{>t_0}$, and is strongly past-determined if $\boldsymbol{h}$ is independent of control.
3. Σ *is finitely observable:* Let $t_0 \in \mathbb{T}$ and let $\tau \in \mathbb{T}_{>t_0}$. We suppose that Σ satisfies the conditions of Theorem 6.1.10 below. Then we see that Σ is finitely observable from τ. Indeed, a controlled trajectory on $[t_0, \tau)$, for a fixed control $\mu \in \mathscr{U}$, is uniquely determined by the initial state. This uniqueness then applies to uniqueness for times greater than τ.
4. *Conditions for* Σ *to be stationary:* Generally, a continuous-time state space system is not stationary. However, it is most common to consider systems that *are* stationary, so we consider such systems here. First of all, the definition of stationarity from $t_0 \in \mathbb{T}$ requires that $\tau^*_{t-t_0}(\mathscr{U}_{\geq t_0}) = \mathscr{U}_{\geq t_0}$, i.e., the set of controls is shift-invariant. Then one sees that Σ is strongly stationary if it is autonomous. The argument for this follows along the lines of that for doing Exercise 3.1.19, and we leave the working out of this to the reader. Note that a continuous-time state space system is stationary if and only if it is strongly stationary, and this is a result of the fact that flows are "reversible," cf. Proposition 6.1.13(iii).
5. Σ *is not generally linear:* Presumably, since in Section 6.6 we shall specifically consider linear continuous-time state space systems, it is not the case that all continuous-time state space systems are linear. To see this, one need only

produce a counterexample, and such an example can be seen in Example 2.2.21, and the lack of linearity here is a reflection of the fact that, in the formula (2.4) for controlled trajectories, the input μ does not appear linearly. •

6.1.2 Existence and uniqueness of controlled trajectories, and flows for continuous-time state space systems

As one should expect given the discussion in Section 3.2.1.1, there will be conditions on the dynamics f for a continuous-time state space system that ensure the existence of controlled trajectories. Since we are working, not just with differential equations but with systems, one must also take into account the manner in which f depends on control. One must also think about the most general class of controls that are allowable that ensure the existence of trajectories. In this section, we present all such conditions.

First let us consider the classes of controls we consider, and topologies for these. The theory for this was developed in Section III-6.5.4 and was overviewed in Section IV-1.3.5. Here we make use of this theory for signals taking values in $\mathbb{R}^m$, as explained in Section IV-1.4.3. Thus the following definitions have already appeared in these volumes, but we repeat them here in close proximity to their use.

6.1.9 Definition (Controls for continuous-time state space systems) Let $\mathbb{T} \subseteq \mathbb{R}$ be a continuous time-domain and let $U \subseteq \mathbb{R}^m$.

(i) We let $\mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}; U)$ be the set of U-valued locally essentially bounded functions, i.e., $f \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}; U)$ if, for every compact interval $\mathbb{S} \subseteq \mathbb{S}$ and for every $a \in \{1, \dots, m\}$,

$$\operatorname{ess\,sup}\{|f_a(t)| \mid t \in \mathbb{S}\} < \infty.$$

We equip $\mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}; U)$ with the topology defined by the family of seminorms $\|\cdot\|_{\infty,\mathbb{S}}$ defined by

$$\|f\|_{\infty,\mathbb{S}} = \max\{\operatorname{ess\,sup}\{|f_a(t)| \mid t \in \mathbb{S}\} \mid a \in \{1, \dots, m\}\}.$$

(ii) For $p \in [1, \infty)$, we let $\mathsf{L}^p_{\mathrm{loc}}(\mathbb{T}; U)$ be the set of U-valued locally integrable functions, i.e., $f \in \mathsf{L}^p_{\mathrm{loc}}(\mathbb{T}; U)$ if, for every compact interval $\mathbb{S} \subseteq \mathbb{S}$ and for every $a \in \{1, \dots, m\}$,

$$\int_{\mathbb{S}} |f_a(t)|^p \, \mathrm{d}t < \infty.$$

We equip $\mathsf{L}^p_{\mathrm{loc}}(\mathbb{T}; U)$ with the topology defined by the family of seminorms $\|\cdot\|_{p,\mathbb{S}}$ defined by

$$\|f\|_{p,\mathbb{S}} = \max\left\{\left(\int_{\mathbb{S}} |f_a(t)|^p \, \mathrm{d}t\right)^{1/p} \;\middle|\; a \in \{1, \dots, m\}\right\}.$$ •

We refer to Section III-6.2.3 for a description of how topological concepts such as convergence and continuity work in these spaces with their topology described by seminorms.

With an understanding of the spaces of inputs that we shall consider, let us now characterise the conditions on the dynamics f of a continuous-time state space system that ensure existence and uniqueness of controlled trajectories.

6.1.10 Theorem (Existence and uniqueness of controlled trajectories for continuous-time state space systems) *Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a continuous-time state space system and assume the following:*

(i) the map $t \mapsto \mathbf{f}(t, \mathbf{x}, \mathbf{u})$ is measurable for each $(\mathbf{x}, \mathbf{u}) \in X \times U$;

(ii) the map $\mathbf{x} \mapsto \mathbf{f}(t, \mathbf{x}, \mathbf{u})$ is locally Lipschitz for each $(t, \mathbf{u}) \in \mathbb{T} \times U$;

(iii) $(\mathbf{x}, \mathbf{u}) \mapsto \mathbf{f}(t, \mathbf{x}, \mathbf{u})$ is continuous for every $t \in \mathbb{T}$;

(iv) for each $(t, \mathbf{x}, \mathbf{u}) \in \mathbb{T} \times X \times U$, there exist $r_1, r_2, \alpha \in \mathbb{R}_{>0}$ and

$$g, L \in L^1([t - \alpha, t + \alpha]; \mathbb{R}_{\geq 0})$$

such that

$$\|\mathbf{f}(s, \mathbf{x}', \mathbf{u}')\| \leq g(s), \qquad (s, \mathbf{x}', \mathbf{u}') \in ([t-\alpha, t+\alpha] \cap \mathbb{T}) \times B^n(r_1, \mathbf{x}) \times (B^m(r_2, \mathbf{u}) \cap U), \tag{6.2}$$

and

$$\|\mathbf{f}(s, \mathbf{x}_1', \mathbf{u}') - \mathbf{f}(s, \mathbf{x}_2', \mathbf{u}')\| \leq L(s)\|\mathbf{x}_1' - \mathbf{x}_2'\|, \quad s \in [t - \alpha, t + \alpha] \cap \mathbb{T},\ \mathbf{x}_1', \mathbf{x}_2' \in B^n(r_1, \mathbf{x}),\ \mathbf{u}' \in (B^m(r_2, \mathbf{u}) \cap U). \tag{6.3}$$

(v) $\mathscr{U} \subseteq L^\infty_{\mathrm{loc}}((\mathbb{T}); U)$.

Then, for $\boldsymbol{\mu} \in \mathscr{U}$ and $(t_0, \mathbf{x}_0) \in \mathbb{T} \times X$, there exists a subinterval $\mathbb{T}' \subseteq \mathbb{T}$, relatively open in $\mathbb{T}$ and with $t_0 \in \mathrm{int}_{\mathbb{T}}(\mathbb{T}')$, and a locally absolutely continuous curve $\boldsymbol{\xi} \colon \mathbb{T}' \to X$ such that $\boldsymbol{\xi}(t_0) = \mathbf{x}_0$ and such that $(\boldsymbol{\xi}, \boldsymbol{\mu}|\mathbb{T}') \in \mathrm{Ctraj}(\Sigma)$. Moreover, if $\mathbb{T}''$ is another such interval and $\boldsymbol{\eta} \colon \mathbb{T}'' \to X$ is another such curve, then $\boldsymbol{\eta}(t) = \boldsymbol{\xi}(t)$ for all $t \in \mathbb{T}'' \cap \mathbb{T}'$.

Proof We define an ordinary differential equation $\boldsymbol{F}$ with right-hand side

$$\begin{aligned} \widehat{\boldsymbol{F}} \colon \mathbb{T} \times X &\to \mathbb{R}^n \\ (t, \boldsymbol{x}) &\mapsto (t, \boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{\mu}(t))). \end{aligned}$$

We claim that $\boldsymbol{F}$ satisfies the conditions of part (ii) of Theorem 3.2.8.

First we note that, for fixed $\boldsymbol{x} \in X$, $t \mapsto \widehat{\boldsymbol{F}}(t, \boldsymbol{x})$ is locally integrable by Lemma 1 from the proof of Theorem 3.2.8. Clearly $\boldsymbol{x} \mapsto \widehat{\boldsymbol{F}}(t, \boldsymbol{x})$ is locally Lipschitz for $t \in \mathbb{T}$.

Let $(t, \boldsymbol{x}) \in \mathbb{T} \times X$. Let $\alpha \in \mathbb{R}_{>0}$. Then, since $\boldsymbol{\mu}$ is locally essentially bounded, there exists a compact $K_\mu \subseteq U$ such that $\boldsymbol{\mu}(s) \in K_\mu$ for almost every $s \in [t - \alpha, t + \alpha] \cap \mathbb{T}$. Let $\boldsymbol{u} \in K_\mu$ and, by the assumptions on $\boldsymbol{f}$, let $r_{1,u}, r_{2,u}, \alpha_u \in \mathbb{R}_{>0}$ and let

$$g_u \in L^1([t - \alpha_u, t + \alpha_u]; \mathbb{R}_{\geq 0})$$

be such that

$$\|\boldsymbol{f}(s,\boldsymbol{x}',\boldsymbol{u}')\| \leq g_{\boldsymbol{u}}(s), \qquad (s,\boldsymbol{x}',\boldsymbol{u}') \in ([t-\alpha,t+\alpha]\cap\mathbb{T})\times \mathsf{B}^n(r_{1,\boldsymbol{u}},\boldsymbol{x})\times(\mathsf{B}^m(r_{2,\boldsymbol{u}},\boldsymbol{u})\cap U).$$

By compactness of $K_{\boldsymbol{\mu}}$, let $\boldsymbol{u}_1,\dots,\boldsymbol{u}_p \in K_{\boldsymbol{\mu}}$ be such that the balls $\mathsf{B}^m(r_{2,\boldsymbol{u}_j},\boldsymbol{u}_j)$ cover $K_{\boldsymbol{\mu}}$. Define

$$r = \min\{r_{1,\boldsymbol{u}_1},\dots,r_{p,\boldsymbol{u}_p}\}, \quad \alpha' = \min\{\alpha,\alpha_{\boldsymbol{u}_1},\dots,\alpha_{\boldsymbol{u}_p}\},$$

and

$$g(s) = \max\{g_{\boldsymbol{u}_1}(s),\dots,g_{\boldsymbol{u}_p}(s)\}, \qquad s\in[t-\alpha',t+\alpha'].$$

Then

$$\|\widehat{\boldsymbol{F}}(s,\boldsymbol{x}')\| = \|\boldsymbol{f}(s,\boldsymbol{x}',\boldsymbol{\mu}(s))\| \leq g(s), \qquad (s,\boldsymbol{x}')\in([t-\alpha',t+\alpha']\cap\mathbb{T})\times\mathsf{B}^n(r,\boldsymbol{x}).$$

This gives the bound (3.7) for $\widehat{\boldsymbol{F}}$.

Again let $(t,\boldsymbol{x})\in\mathbb{T}\times X$ and let $\alpha\in\mathbb{R}_{>0}$. Let $K_{\boldsymbol{\mu}}\subseteq U$ be compact such that $\boldsymbol{\mu}(s)\in K_{\boldsymbol{\mu}}$ for almost every $s\in[t-\alpha,t+\alpha]\cap\mathbb{T}$. Let $\boldsymbol{u}\in K_{\boldsymbol{\mu}}$ and, by the assumptions on $\boldsymbol{f}$, let $t_{1,\boldsymbol{u}},r_{2,\boldsymbol{u}},\alpha_{\boldsymbol{u}}\in\mathbb{R}_{>0}$ and let

$$L_{\boldsymbol{u}} \in \mathsf{L}^1([t-\alpha_{\boldsymbol{u}},t+\alpha_{\boldsymbol{u}}];\mathbb{R}_{\geq 0})$$

be such that

$$\|\boldsymbol{f}(s,\boldsymbol{x}_1',\boldsymbol{u}')-\boldsymbol{f}(s,\boldsymbol{x}_2',\boldsymbol{u}')\| \leq L_{\boldsymbol{u}}(s)\|\boldsymbol{x}_1'-\boldsymbol{x}_2'\|,$$
$$s\in[t-\alpha_{\boldsymbol{u}},t+\alpha_{\boldsymbol{u}}]\cap\mathbb{T},\ \boldsymbol{x}_1',\boldsymbol{x}_2'\in\mathsf{B}^n(r,\boldsymbol{x}),\ \boldsymbol{u}'\in(\mathsf{B}^m(r_{2,\boldsymbol{u}},\boldsymbol{u})\cap U).$$

By compactness of $K_{\boldsymbol{\mu}}$, let $\boldsymbol{u}_1,\dots,\boldsymbol{u}_p\in K_{\boldsymbol{\mu}}$ be such that the balls $\mathsf{B}^m(r_{2,\boldsymbol{u}_j},\boldsymbol{u}_j)$ cover $K_{\boldsymbol{\mu}}$. Define

$$r = \min\{r_{1,\boldsymbol{u}_1},\dots,r_{p,\boldsymbol{u}_p}\}, \quad \alpha' = \min\{\alpha,\alpha_{\boldsymbol{u}_1},\dots,\alpha_{\boldsymbol{u}_p}\},$$

and

$$L(s) = \max\{L_{\boldsymbol{u}_1}(s),\dots,L_{\boldsymbol{u}_p}(s)\}, \qquad s\in[t-\alpha',t+\alpha'].$$

Then

$$\|\widehat{\boldsymbol{F}}(s,\boldsymbol{x}_1')-\widehat{\boldsymbol{F}}(s,\boldsymbol{x}_2')\| = \|\boldsymbol{f}(s,\boldsymbol{x}_1',\boldsymbol{\mu}(s))-\boldsymbol{f}(s,\boldsymbol{x}_2',\boldsymbol{\mu}(s))\| \leq L(s)\|\boldsymbol{x}_1'-\boldsymbol{x}_2'\|,$$
$$s\in[t-\alpha',t+\alpha']\cap\mathbb{T},\ \boldsymbol{x}_1',\boldsymbol{x}_2'\in\mathsf{B}^n(r,\boldsymbol{x}).$$

This gives the bound (3.8) for $\widehat{\boldsymbol{F}}$, and this shows that $\boldsymbol{F}$ satisfies the conditions of Theorem 3.2.8.

The theorem, then, follows from Theorem 3.2.8. ∎

The theorem now permits an adaptation of the notion of the flow of a differential equation in Section 3.2.1.3 to continuous-time state space systems. Let us undertake this notation here.

6.1.11 Definition (Interval of existence, domain of solutions) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a continuous-time state space system that satisfies the conditions of Theorem 6.1.10 for existence and uniqueness of controlled trajectories.

(i) For $(t_0, x_0, \mu) \in \mathbb{T} \times X \times \mathscr{U}$, denote

$$J_\Sigma(t_0, x_0, \mu) = \cup\{J \subseteq \operatorname{dom}(\mu) \mid \ J \text{ is an interval and there exists } \xi\colon J \to X \text{ such that } (\xi, \mu|J) \in \operatorname{Ctraj}(\Sigma),\ \xi(t_0) = x_0\}.$$

The interval $J_\Sigma(t_0, x_0, \mu)$ is the ***interval of existence*** for the initial value problem

$$\dot{\xi}(t) = f(t, \xi(t), \mu(t)), \quad \xi(t_0) = x_0.$$

(ii) For $\mu \in \mathscr{U}$, the ***domain of solutions*** for Σ for the control μ is

$$D_\Sigma(\mu) = \{(t, t_0, x_0) \in \mathbb{T} \times \mathbb{T} \times X \mid \ t \in J_\Sigma(t_0, x_0, \mu)\}.$$

(iii) The ***domain of solutions*** for Σ is

$$D_\Sigma = \{(t, t_0, x_0, \mu) \in \mathbb{T} \times \mathbb{T} \times X \times \mathscr{U} \mid \ (t, t_0, x_0) \in D_\Sigma(\mu)\}. \qquad \bullet$$

As with ordinary differential equations, we can now introduce the notion of the flow for a continuous-time state space system.

6.1.12 Definition (Flow of a continuous-time state space system) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a continuous-time state space system that satisfies the conditions of Theorem 6.1.10 for existence and uniqueness of controlled trajectories. The ***flow*** of Σ is the map $\Phi^\Sigma\colon D_\Sigma \to X$ defined by asking that $\Phi^\Sigma(t, t_0, x_0, \mu)$ is the solution, evaluated at t, of the initial value problem

$$\dot{\xi}(\tau) = f(\tau, \xi(\tau), \mu(\tau)), \quad \xi(t_0) = x_0. \qquad \bullet$$

The definition, phrased differently, says that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi^\Sigma(t, t_0, x_0, \mu) = f(t, \Phi^\Sigma(t, t_0, x_0, \mu), \mu(t)), \quad \Phi^\Sigma(t_0, t_0, x_0, \mu) = x_0.$$

For $t, t_0 \in \mathbb{T}$ and $\mu \in \mathscr{U}$, it is sometimes convenient to denote

$$D_\Sigma(t, t_0, \mu) = \{x \in X \mid \ (t, t_0, x) \in D_\Sigma(\mu)\},$$

and then

$$\begin{aligned} \Phi^{\Sigma,\mu}_{t,t_0}\colon D_\Sigma(t, t_0, \mu) &\to X \\ x &\mapsto \Phi^\Sigma(t, t_0, x, \mu). \end{aligned}$$

Along similar lines, for $t_0 \in \mathbb{T}$, we denote

$$D_\Sigma(t_0) = \{(t, x, \mu) \in \mathbb{T} \times X \times \mathscr{U} \mid \ (t, t_0, x, \mu) \in D_\Sigma\},$$

and then

$$\begin{aligned}\Phi^{\Sigma}(t_0)\colon D_{\Sigma}(t_0) &\to X\\ (t, x, \mu) &\mapsto \Phi^{\Sigma}(t, t_0, x, \mu).\end{aligned}$$

Finally, for $t, t_0 \in \mathbb{T}$, we denote

$$D_{\Sigma}(t, t_0) = \{(x, \mu) \in X \times \mathscr{U} \mid (t, t_0, x, \mu) \in D_{\Sigma}\},$$

and then

$$\begin{aligned}\Phi^{\Sigma}(t, t_0)\colon D_{\Sigma}(t, t_0) &\to X\\ (x, \mu) &\mapsto \Phi^{\Sigma}(t, t_0, x, \mu).\end{aligned}$$

Let us enumerate some of the more elementary properties of the flow for a continuous-time state space system, just as for an ordinary differential equation.

6.1.13 Proposition (Elementary properties of flows of continuous-time state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a continuous-time state space system that satisfies the conditions of Theorem 6.1.10 for existence and uniqueness of controlled trajectories. Then the following statements hold:*

(i) for each $(\mathrm{t}_0, \mathbf{x}_0, \mu) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$, $(\mathrm{t}_0, \mathrm{t}_0, \mathbf{x}_0, \mu) \in \mathrm{D}_{\Sigma}$ and $\Phi^{\Sigma}(\mathrm{t}_0, \mathrm{t}_0, \mathbf{x}_0, \mu) = \mathbf{x}_0$;

(ii) if $(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \mu) \in \mathrm{D}_{\Sigma}$, then $(\mathrm{t}_3, \mathrm{t}_2, \Phi^{\Sigma}(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}), \mu) \in \mathrm{D}_{\Sigma}$ if and only if $(\mathrm{t}_3, \mathrm{t}_1, \mathbf{x}, \mu) \in \mathrm{D}_{\Sigma}$ and, if this holds, then

$$\Phi^{\Sigma}(\mathrm{t}_3, \mathrm{t}_1, \mathbf{x}, \mu) = \Phi^{\Sigma}(\mathrm{t}_3, \mathrm{t}_2, \Phi^{\Sigma}(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \mu), \mu).$$

(iii) if $(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \mu) \in \mathrm{D}_{\Sigma}$, then $(\mathrm{t}_1, \mathrm{t}_2, \Phi^{\Sigma}(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \mu), \mu) \in \mathrm{D}_{\Sigma}$ and

$$\Phi^{\Sigma}(\mathrm{t}_1, \mathrm{t}_2, \Phi^{\Sigma}(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \mu), \mu) = \mathbf{x}.$$

Proof This follows immediately from Proposition 3.2.12. ■

Useful mnemonics associated with parts (i)–(iii) are:

$$\Phi^{\Sigma,\mu}_{t_0,t_0} = \mathrm{id}_X, \quad (\Phi^{\Sigma,\mu}_{t_2,t_1})^{-1} = \Phi^{\Sigma,\mu}_{t_1,t_2}, \quad \Phi^{\Sigma,\mu}_{t_3,t_2} \circ \Phi^{\Sigma,\mu}_{t_2,t_1} = \Phi^{\Sigma,\mu}_{t_3,t_1}.$$

However, these really are just mnemonics, since they do not account carefully for the domains of the mappings being used.

The following result encodes some less elementary properties of the flow of an ordinary differential equation, including the regularity of the dependence on time, state, and control.

6.1.14 Theorem (Properties of flows of continuous-time state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a continuous-time state space system that satisfies the conditions of Theorem 6.1.10 for existence and uniqueness of controlled trajectories. Then the following statements hold:*

(i) for $(\mathrm{t}_0, \mathbf{x}_0, \mu) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$, $\mathrm{J}_{\Sigma}(\mathrm{t}_0, \mathbf{x}_0, \mu)$ is an interval that is a relatively open subset of $\mathbb{T}$;

(ii) *for* $(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$, *the curve*

$$\begin{aligned} \gamma_{(\mathrm{t}_0,\mathbf{x}_0,\boldsymbol{\mu})} \colon \mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) &\to \mathrm{X} \\ \mathrm{t} &\mapsto \Phi^\Sigma(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) \end{aligned}$$

is well-defined and absolutely continuous;

(iii) *for* $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}$ *and* $\boldsymbol{\mu} \in \mathscr{U}$, $\mathrm{D}_\Sigma(\mathrm{t}, \mathrm{t}_0, \boldsymbol{\mu})$ *is open in* X;

(iv) *for* $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}$ *and* $\boldsymbol{\mu} \in \mathscr{U}$ *for which* $\mathrm{D}_\Sigma(\mathrm{t}, \mathrm{t}_0, \boldsymbol{\mu}) \neq \varnothing$, $\Phi^{\Sigma,\boldsymbol{\mu}}_{\mathrm{t},\mathrm{t}_0}$ *is a locally bi-Lipschitz homeomorphism onto its image;*

(v) *for* $\mathrm{t}_0 \in \mathbb{T}$, $\mathrm{D}_\Sigma(\mathrm{t}_0)$ *is relatively open in* $\mathbb{T} \times \mathrm{X} \times \mathscr{U}$;

(vi) *for* $\mathrm{t}_0 \in \mathbb{T}$, *the map*

$$\begin{aligned} \Phi^\Sigma(\mathrm{t}_0) \colon \mathrm{D}_\Sigma(\mathrm{t}_0) &\to \mathrm{X} \\ (\mathrm{t}, \mathbf{x}, \boldsymbol{\mu}) &\mapsto \Phi^\Sigma(\mathrm{t}, \mathrm{t}_0, \mathbf{x}, \boldsymbol{\mu}) \end{aligned}$$

is well-defined and continuous;

(vii) D_Σ *is relatively open in* $\mathbb{T} \times \mathbb{T} \times \mathrm{X} \times \mathscr{U}$;

(viii) *the map*

$$\Phi^\Sigma \colon \mathrm{D}_\Sigma \to \mathrm{X}$$

is continuous;

(ix) *for* $(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$ *and for* $\epsilon \in \mathbb{R}_{>0}$, *there exists* $\mathrm{r}, \rho, \alpha \in \mathbb{R}_{>0}$ *such that*

$$\sup \mathrm{J}_\Sigma(\mathrm{t}, \mathbf{x}, \boldsymbol{\mu}) > \sup \mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) - \epsilon, \qquad \inf \mathrm{J}_\Sigma(\mathrm{t}, \mathbf{x}, \boldsymbol{\mu}) < \inf \mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) + \epsilon$$

for all $(\mathrm{t}, \mathbf{x}, \boldsymbol{\mu}) \in (\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha) \cap \mathbb{T} \times \mathsf{B}^{\mathrm{n}}(\mathrm{r}, \mathbf{x}_0) \times \mathsf{B}(\rho, \boldsymbol{\mu}_0)$.

Proof Parts (i)–(iv) follow immediately from Theorem 3.2.13. The remaining parts (v)–(ix) are analogous to the corresponding parts of Theorem 3.2.13, but include the rôle of control in the flow. We shall prove the local version of the theorem—the analogue of Lemma 2 from the proof of Theorem 3.2.13—since the global version follows by an argument very much like the lengthy argument from Theorem 3.2.13.

Let $(t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0) \in \mathbb{T} \times X \times \mathscr{U}$. We shall suppose first that $t_0 \neq \sup \mathbb{T}$.

Let $r' \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^n(r', \boldsymbol{x}_0) \subseteq X$. Let $r = \frac{r'}{2}$ and $\lambda \in (0, 1)$. There exists $\alpha \in \mathbb{R}_{>0}$ such that

$$\int_{t_0}^{t} \|\boldsymbol{f}(s, \boldsymbol{x}, \boldsymbol{\mu}_0(s))\| \, \mathrm{d}s < \frac{r}{2}$$

for $t \in [t_0, t_0 + \alpha]$ and $\boldsymbol{x} \in \mathsf{B}^n(r', \boldsymbol{x}_0)$, and

$$\int_{t_0}^{t} \|\boldsymbol{f}(s, \boldsymbol{x}_1, \boldsymbol{\mu}_0(s)) - \boldsymbol{f}(s, \boldsymbol{x}_2, \boldsymbol{\mu}_0(s))\| \, \mathrm{d}s \leq \lambda \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|$$

for $t \in [t_0, t_0 + \alpha]$ and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathsf{B}^n(r', \boldsymbol{x}_0)$. Now let $\rho \in \mathbb{R}_{>0}$ be such that, if $\boldsymbol{\mu} \in \mathscr{U}$ satisfies

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|_{[t_0, t_0+\alpha], \infty} = \sup\{\|\boldsymbol{\mu}(s) - \boldsymbol{\mu}_0(s)\| \mid s \in [t_0, t_0 + \alpha]\} < \rho,$$

then

$$\int_{t_0}^{t} \|f(s, x, \mu(s)) - f(s, x, \mu_0(s))\| \, \mathrm{d}s < \frac{r}{2}$$

for $t \in [t_0, t_0 + \alpha]$ and $x \in \mathsf{B}^n(r', x_0)$, and

$$\int_{t_0}^{t} \left| \|f(s, x_1, \mu(s)) - f(s, x_2, \mu(s))\| - \|f(s, x_1, \mu_0(s)) - f(s, x_2, \mu_0(s))\| \right| \, \mathrm{d}s < \frac{\lambda}{2} \|x_2 - x_1\|$$

for $t \in [t_0, t_0 + \alpha]$ and $x_1, x_2 \in \mathsf{B}^n(r', x_0)$. This is possible by the assumptions (6.2) and (6.3) on f. Let us denote

$$\mathsf{B}_{[t_0, t_0+\alpha]}(\rho, \mu_0) = \{\mu \in \mathscr{U} \mid \|\mu - \mu_0\|_{[t_0, t_0+\alpha], \infty} < \rho\}.$$

An application of the triangle inequality gives

$$\int_{t_0}^{t} \|f(s, x, \mu(s))\| \, \mathrm{d}s < r$$

for $t \in [t_0, t_0 + \alpha]$, $x \in \mathsf{B}^n(r', x_0)$, and $\mu \in \mathsf{B}_{[t_0, t_0+\alpha]}(\rho, \mu_0)$, and

$$\int_{t_0}^{t} \|f(s, x_2, \mu(s)) - f(s, x_1, \mu(s))\| \, \mathrm{d}s < \lambda \|x_2 - x_1\|$$

for $t \in [t_0, t_0 + \alpha]$, $x_1, x_2 \in \mathsf{B}^n(r', x_0)$, and $\mu \in \mathsf{B}_{[t_0, t_0+\alpha]}(\rho, \mu_0)$.

With these constructions, one can duplicate the Contraction Mapping Theorem arguments from the proof of Theorem 3.2.13 to show that Φ^Σ is a continuous mapping

$$[t_0, t_0 + \alpha] \times \{t_0\} \times \mathsf{B}^n(r, x_0) \times \mathsf{B}_{[t_0, t_0+\alpha]}(\rho, \mu_0) \to X.$$

One can similarly give the same conclusion for the time interval $[t_0 - \alpha, t_0]$, and thus give the local version of the continuity properties of the flow for Σ. As indicated at the beginning of the proof, one can then duplicate the lengthy argument from the proof of Theorem 3.2.13 to give the last five parts of the theorem. ■

6.1.3 Control-affine continuous-time state space systems

We next consider a special class of continuous-time state space systems. The class is worthy of consideration for a few reasons: (1) one can consider for these systems a somewhat broader class of controls, namely those that are locally integrable; (2) this class of systems is a midpoint between general continuous-time state space systems and the linear systems we shall consider in Section 6.6; (3) systems that arise in practice are often of this form.

Here is the definition.

6.1.15 Definition (Control-affine continuous-time state space system) A continuous-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ with $U \subseteq \mathbb{R}^m$ is ***control-affine*** if

(i) there exists $f_0, f_1, \dots, f_m \colon \mathbb{T} \times X \to \mathbb{R}^n$ such that

$$f(t, x, u) = f_0(t, x) + \sum_{a=1}^{m} u_a f_a(t, x),$$

and

(ii) there exists $h_0, h_1, \dots, h_m \colon \mathbb{T} \times X \to \mathbb{R}^k$ such that

$$h(t, x, u) = h_0(t, x) + \sum_{a=1}^{m} u_a h_a(t, x).$$

We call f_0 (resp. h_0) the ***drift dynamics*** (resp. ***drift/output map***) and $f_1, \dots, f_m$ (resp. $h_1, \dots, h_m$) the ***control dynamics*** (resp. ***control/output maps***). •

For a control-affine continuous-time state space system, we shall frequently denote $\mathscr{F} = (f_0, f_1, \dots, f_m)$ and $\mathscr{H} = (h_0, h_1, \dots, h_m)$ and then prescribe such a system by the data $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H})$. Of course, all the notions attached to continuous-time state space systems—e.g., controlled trajectories, controlled outputs, autonomous, proper—can also be attached to those that are control-affine.

Let us state the conditions for existence and uniqueness of controlled trajectories for control-affine systems. Here we see that there is a distinction between the autonomous and nonautonomous cases.

6.1.16 Theorem (Existence and uniqueness of controlled trajectories for control-affine continuous-time state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H})$ be a control-affine continuous-time state space system and assume the following:*

(i) the maps $\mathrm{t} \mapsto \mathbf{f}_\mathrm{a}(\mathrm{t}, \mathbf{x})$, $\mathrm{a} \in \{0, 1, \dots, \mathrm{m}\}$, are measurable for each $\mathbf{x} \in \mathrm{X}$;

(ii) the maps $\mathbf{x} \mapsto \mathbf{f}_\mathrm{a}(\mathrm{t}, \mathbf{x})$, $\mathrm{a} \in \{0, 1, \dots, \mathrm{m}\}$, are locally Lipschitz for each $\mathrm{t} \in \mathbb{T}$;

(iii) for each $(\mathrm{t}, \mathbf{x}) \in \mathbb{T} \times \mathrm{X}$, there exist $\mathrm{r}, \alpha \in \mathbb{R}_{>0}$ and

$$\mathrm{g}, \mathrm{L} \in \mathsf{L}^1([\mathrm{t} - \alpha, \mathrm{t} + \alpha]; \mathbb{R}_{\geq 0})$$

such that

$$\|\mathbf{f}_\mathrm{a}(\mathrm{s}, \mathbf{x}')\| \leq \mathrm{g}(\mathrm{s}), \qquad \mathrm{a} \in \{0, 1, \dots, \mathrm{m}\},\ (\mathrm{s}, \mathbf{x}') \in ([\mathrm{t} - \alpha, \mathrm{t} + \alpha] \cap \mathbb{T} \times \mathsf{B}^\mathrm{n}(\mathrm{r}, \mathbf{x}), \quad (6.4)$$

and

$$\|\mathbf{f}_\mathrm{a}(\mathrm{s}, \mathbf{x}'_1) - \mathbf{f}_\mathrm{a}(\mathrm{s}, \mathbf{x}'_2)\| \leq \mathrm{L}(\mathrm{s})\|\mathbf{x}'_1 - \mathbf{x}'_2\|, \quad \mathrm{a} \in \{0, 1, \dots, \mathrm{m}\},\ \mathrm{s} \in [\mathrm{t} - \alpha, \mathrm{t} + \alpha] \cap \mathbb{T},\ \mathbf{x}'_1, \mathbf{x}'_2 \in \mathsf{B}^\mathrm{n}(\mathrm{r}, \mathbf{x}). \quad (6.5)$$

(iv) $\mathscr{U} \subseteq \mathsf{L}^\infty_{\mathrm{loc}}((\mathbb{T}); \mathrm{U})$.

Then, for $\boldsymbol{\mu} \in \mathscr{U}$ and $(t_0, \mathbf{x}_0) \in \mathbb{T} \times X$, there exists a subinterval $\mathbb{T}' \subseteq \mathbb{T}$, relatively open in $\mathbb{T}$ and with $t_0 \in \operatorname{int}_{\mathbb{T}}(\mathbb{T}')$, and a locally absolutely continuous curve $\boldsymbol{\xi} \colon \mathbb{T}' \to X$ such that $\boldsymbol{\xi}(t_0) = \mathbf{x}_0$ and such that $(\boldsymbol{\xi}, \boldsymbol{\mu}|\mathbb{T}') \in \operatorname{Ctraj}(\Sigma)$. Moreover, if $\mathbb{T}''$ is another such interval and $\boldsymbol{\eta} \colon \mathbb{T}'' \to X$ is another such curve, then $\boldsymbol{\eta}(t) = \boldsymbol{\xi}(t)$ for all $t \in \mathbb{T}'' \cap \mathbb{T}'$.

Proof We can prove the theorem by showing that, for a given $\boldsymbol{\mu} \in \mathscr{U}$, the hypotheses imply those of Theorem 6.1.10. This verification, however, is elementary, and we leave the working out of this to the reader as Exercise 6.1.5. ■

One of the useful features of control-affine continuous-time state space systems is that one can use locally integrable controls for autonomous systems, i.e., those where $f_0, f_1, \dots, f_m$ are independent of time. This is a case that often arises in applications.

6.1.17 Theorem (Existence and uniqueness of controlled trajectories for autonomous control-affine continuous-time state space systems) *Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H})$ be an autonomous control-affine continuous-time state space system and assume that the maps $\mathbf{x} \mapsto \mathbf{f}_a(\mathbf{x})$, $a \in \{0, 1, \dots, m\}$, are locally Lipschitz. Suppose that $\mathscr{U} \subseteq L^1_{\mathrm{loc}}((\mathbb{T}); U)$. Then, for $\boldsymbol{\mu} \in \mathscr{U}$ and $(t_0, \mathbf{x}_0) \in \mathbb{T} \times X$, there exists a subinterval $\mathbb{T}' \subseteq \mathbb{T}$, relatively open in $\mathbb{T}$ and with $t_0 \in \operatorname{int}_{\mathbb{T}}(\mathbb{T}')$, and a locally absolutely continuous curve $\boldsymbol{\xi} \colon \mathbb{T}' \to X$ such that $\boldsymbol{\xi}(t_0) = \mathbf{x}_0$ and such that $(\boldsymbol{\xi}, \boldsymbol{\mu}|\mathbb{T}') \in \operatorname{Ctraj}(\Sigma)$. Moreover, if $\mathbb{T}''$ is another such interval and $\boldsymbol{\eta} \colon \mathbb{T}'' \to X$ is another such curve, then $\boldsymbol{\eta}(t) = \boldsymbol{\xi}(t)$ for all $t \in \mathbb{T}'' \cap \mathbb{T}'$.*

Proof As with the proof of Theorem 6.1.10, we can prove the theorem by showing that, for a given $\boldsymbol{\mu} \in \mathscr{U}$, the hypotheses imply those of Theorem 3.2.8 for the ordinary differential equation $\boldsymbol{F}$ with right-hand side

$$\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) = f_0(\boldsymbol{x}) + \sum_{a=1}^{m} \mu_a(t) f_a(\boldsymbol{x}).$$

This verification, however, is elementary, and we leave the working out of this to the reader as Exercise 6.1.6. ■

Given the preceding theorems, the flow-related concepts and terminology of Definitions 6.1.11 and 6.1.12 apply equally well to control-affine systems, provided that one keeps in mind the classes of controls one can use for autonomous and nonautonomous systems. The elementary properties of flows from Proposition 6.1.13 also immediately apply to all control-affine systems. Additionally, the regularity properties of the flow from Theorem 6.1.14 carries over verbatim for control-affine systems, this by virtue of the fact that a control-affine continuous-time state space system satisfying the hypotheses of Theorem 6.1.16 also satisfies the hypotheses of Theorem 6.1.10. For autonomous control-affine systems, this verbatim transcription does not apply, since the class of controls is different. Nevertheless, the corresponding result does hold, as we now state.

6.1.18 Theorem (Properties of flows of autonomous control-affine continuous-time state space systems) *Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H})$ be an autonomous control-affine continuous-time state space system that satisfies the conditions of Theorem 6.1.17 for existence and uniqueness of controlled trajectories. Then the following statements hold:*

(i) *for $(t_0, x_0, \mu) \in \mathbb{T} \times X \times \mathscr{U}$, $J_\Sigma(t_0, x_0, \mu)$ is an interval that is a relatively open subset of $\mathbb{T}$;*

(ii) *for $(t_0, x_0, \mu) \in \mathbb{T} \times X \times \mathscr{U}$, the curve*

$$\begin{aligned} \gamma_{(t_0,x_0,\mu)} \colon J_\Sigma(t_0, x_0, \mu) &\to X \\ t &\mapsto \Phi^\Sigma(t, t_0, x_0, \mu) \end{aligned}$$

is well-defined and absolutely continuous;

(iii) *for $t, t_0 \in \mathbb{T}$ and $\mu \in \mathscr{U}$, $D_\Sigma(t, t_0, \mu)$ is open in X;*

(iv) *for $t, t_0 \in \mathbb{T}$ and $\mu \in \mathscr{U}$ for which $D_\Sigma(t, t_0, \mu) \neq \emptyset$, $\Phi^{\Sigma,\mu}_{t,t_0}$ is a locally bi-Lipschitz homeomorphism onto its image;*

(v) *for $t_0 \in \mathbb{T}$, $D_\Sigma(t_0)$ is relatively open in $\mathbb{T} \times X \times \mathscr{U}$;*

(vi) *for $t_0 \in \mathbb{T}$, the map*

$$\begin{aligned} \Phi^\Sigma(t_0) \colon D_\Sigma(t_0) &\to X \\ (t, x, \mu) &\mapsto \Phi^\Sigma(t, t_0, x, \mu) \end{aligned}$$

is well-defined and continuous;

(vii) *D_Σ is relatively open in $\mathbb{T} \times \mathbb{T} \times X \times \mathscr{U}$;*

(viii) *the map*

$$\Phi^\Sigma \colon D_\Sigma \to X$$

is continuous;

(ix) *for $(t_0, x_0, \mu_0) \in \mathbb{T} \times X \times \mathscr{U}$ and for $\epsilon \in \mathbb{R}_{>0}$, there exists $r, \rho, \alpha \in \mathbb{R}_{>0}$ such that*

$$\sup J_\Sigma(t, x, \mu) > \sup J_\Sigma(t_0, x_0, \mu_0) - \epsilon, \qquad \inf J_\Sigma(t, x, \mu) < \inf J_\Sigma(t_0, x_0, \mu_0) + \epsilon$$

for all $(t, x, \mu) \in (t_0 - \alpha, t_0 + \alpha) \cap \mathbb{T} \times \mathsf{B}^n(r, x_0) \times \mathsf{B}(\rho, \mu_0)$.

Proof The constructions from the proof of Theorem 6.1.14 can be equally well performed under the hypotheses of the current theorem, and we leave the details of this to the reader in Exercise 6.1.7. ■

Exercises

6.1.1 Show that a continuous-time state space system is not memoryless. (See Example 2.2.31–2 for the definition of a memoryless system.)

6.1.2 For the continuous-time state space systems $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ given and for $t_0 \in \mathbb{T}$, indicate whether they are causal from t_0, strongly causal from t_0, finitely observable from any $\tau \in \mathbb{T}_{>t_0}$, stationary from t_0, strongly stationary from t_0, and/or memoryless.

(a) Take
(i) $X = \mathbb{R}$, (iv) $\mathscr{U} = \mathsf{L}^1_{\text{loc}}((\mathbb{R}); \mathbb{R})$,
(ii) $U = \mathbb{R}$, (v) $f(t, x, u) = tx + u$,
(iii) $\mathbb{T} = \mathbb{R}$, (vi) $h(t, x, u) = \sin(x)$.

(b) Take
(i) $X = \mathbb{R}$, (iv) $\mathscr{U} = \mathsf{L}^1_{\text{loc}}((\mathbb{R}); \mathbb{R})$,
(ii) $U = \mathbb{R}$, (v) $f(t, x, u) = tx + u$,
(iii) $\mathbb{T} = \mathbb{R}$, (vi) $h(t, x, u) = \sin(x)u$.

(c) Take
(i) $X = \mathbb{R}^2$, (iv) $\mathscr{U} = \mathsf{L}^1_{\text{loc}}((\mathbb{R}); \mathbb{R})$,
(ii) $U = \mathbb{R}$, (v) $f(t, (x_1, x_2), u) = (x_1 - x_2, 2x_1 + 3x_2) + (0, u^2)$,
(iii) $\mathbb{T} = \mathbb{R}$, (vi) $h(t, (x_1, x_2), u) = (x_1, x_2)$.

(d) Take
(i) $X = \mathbb{R}^3$, (iv) $\mathscr{U} = \mathsf{L}^1_{\text{loc}}((\mathbb{R}); \mathbb{R})$,
(ii) $U = \mathbb{R}^2$, (v) $f(t, (x_1, x_2, x_3), (u_1, u_2)) = (\cos(t)x_1 - \sin(t)x_2, \sin(t)x_1 + \cos(t)x_2, x_3) + (u_1, 0, u_2)$,
(iii) $\mathbb{T} = \mathbb{R}$,
(vi) $h(t, (x_1, x_2, x_3), (u_1, u_2)) = (x_1 + x_2, x_2 + u)$.

6.1.3 For the continuous-time state space systems $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ given, indicate whether they satisfy the hypotheses of Theorem 6.1.10.

(a) Take
(i) $X = \mathbb{R}$, (iv) $\mathscr{U} = \mathsf{L}^\infty_{\text{loc}}((\mathbb{R}); \mathbb{R})$,
(ii) $U = \mathbb{R}$, (v) $f(t, x, u) = |txu|$,
(iii) $\mathbb{T} = \mathbb{R}$, (vi) $h(t, x, u) = |txu|$.

(b) Take
(i) $X = \mathbb{R}$, (iv) $\mathscr{U} = \mathsf{L}^\infty_{\text{loc}}((\mathbb{R}); \mathbb{R})$,
(ii) $U = \mathbb{R}$, (v) $f(t, x, u) = x + \sqrt{|u|}$,
(iii) $\mathbb{T} = \mathbb{R}$, (vi) $h(t, x, u) = x$.

(c) Take
(i) $X = \mathbb{R}$, (iv) $\mathscr{U} = \mathsf{L}^\infty_{\text{loc}}((\mathbb{R}); \mathbb{R})$,
(ii) $U = \mathbb{R}$,
(iii) $\mathbb{T} = \mathbb{R}$, (v) $f(t, x, u) = \begin{cases} \frac{xu^2}{x^2+u^4}, & (x, u) \neq (0, 0), \\ 0, & (x, u) = (0, 0), \end{cases}$
(vi) $h(t, x, u) = x$.

(d) Take
(i) $X = \mathbb{R}^3$, (iv) $\mathscr{U} = \mathsf{L}^2_{\text{loc}}((\mathbb{R}); \mathbb{R})$,
(ii) $U = \mathbb{R}$, (v) $f(t, x, u) = (x_2, x_3, 0) + (0, 0, u)$,
(iii) $\mathbb{T} = \mathbb{R}$, (vi) $h(t, x, u) = x_1 + x_2 + x_3$.

6.1.4 For each of the following continuous-time state space systems $\Sigma = (X, U, \mathbb{T}; \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ with $\mathscr{U}$ left undetermined, determine the natural choice for $\mathscr{U}$.

(a) Take
  (i) $X = \mathbb{R}$,
  (ii) $U = \mathbb{R}$,
  (iii) $\mathbb{T} = \mathbb{R}$,
  (iv) $f(t, x, u) = u$,
  (v) $h(t, x, u) = x$.

(b) Take
  (i) $X = \mathbb{R}$,
  (ii) $U = \mathbb{R}$,
  (iii) $\mathbb{T} = \mathbb{R}$,
  (iv) $f(t, x, u) = \sin(u)$,
  (v) $h(t, x, u) = x$.

(c) Take
  (i) $X = \mathbb{R}$,
  (ii) $U = \mathbb{R}$,
  (iii) $\mathbb{T} = \mathbb{R}$,
  (iv) $f(t, x, u) = u^2$,
  (v) $h(t, x, u) = x$.

6.1.5 Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H})$ be a control-affine continuous-time state space system satisfying the hypotheses of Theorem 6.1.16. Show that the associated continuous-time state space system satisfies the hypotheses of Theorem 6.1.10.

6.1.6 Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H})$ be a control-affine continuous-time state space system satisfying the hypotheses of Theorem 6.1.17. Show that, for $\boldsymbol{\mu} \in \mathscr{U}$, the ordinary differential equation $\boldsymbol{F}$ with right-hand side

$$\widehat{\boldsymbol{F}}(t, \boldsymbol{x}) = \boldsymbol{f}_0(\boldsymbol{x}) + \sum_{a=1}^{m} \mu_a(t) \boldsymbol{f}_a(\boldsymbol{x}).$$

satisfies the hypotheses of Theorem 3.2.8.

6.1.7 Show that the arguments from the proof of Theorem 6.1.14 can be used to prove Theorem 6.1.18.

6.1.8 Consider the circuit in Figure 6.1 with an ideal diode (the triangle thingy) through which the current is determined by the formula

$$I = I_s \left(\mathrm{e}^{V/\nu V_T} - 1 \right),$$

where V is the voltage drop across the diode and where the physical constants are

I_S	saturation current
V_T	thermal voltage
ν	emission coefficient

Answer the following questions.

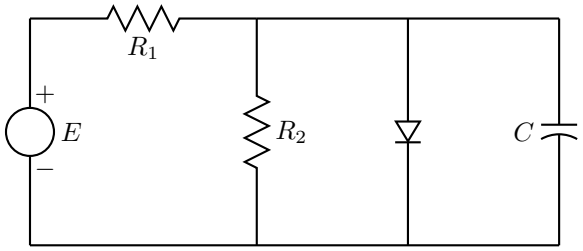


Figure 6.1 Circuit with diode

(a) What is the state space X for the system?
(b) What is the control set U for the system?
(c) What is the time-domain $\mathbb{T}$ for the system?
(d) What is a good choice for the space $\mathscr{U}$ of inputs?
(e) What are the dynamics $\boldsymbol{f}$?
(f) What is the output map $\boldsymbol{h}$?

6.1.9 In Figure 6.2 is depicted a pendulum swinging in the plane and with a motor

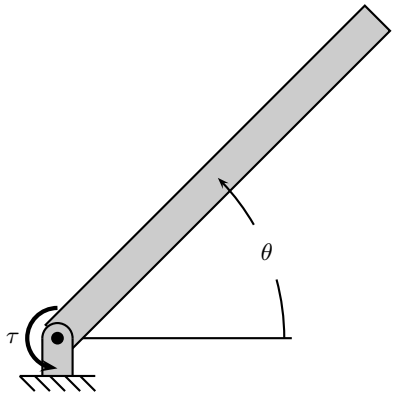


Figure 6.2 Forced pendulum

at the base supplying a torque τ. Suppose that a sensor measures the angle of the pendulum. Let the length of the pendulum be ℓ, let its mass be m, and suppose that it is an homogeneous rod. Answer the following questions.

(a) What is the state space X for the system?
(b) What is the control set U for the system?
(c) What is the time-domain $\mathbb{T}$ for the system?
(d) What is a good choice for the space $\mathscr{U}$ of inputs?
(e) What are the dynamics $\boldsymbol{f}$?
(f) What is the output map $\boldsymbol{h}$?

6.1.10 *(Mini-project)* Consider the simplified bicycle model shown in Figure 6.3. There are two wheels which roll without sliding on the plane. The direction

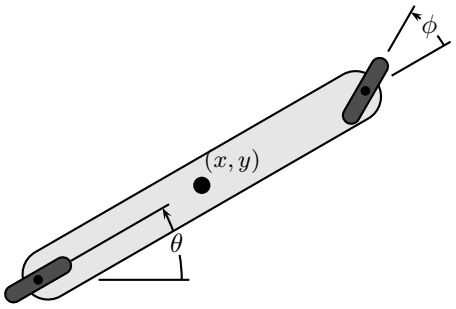


Figure 6.3 A simplified bicycle model

of the "back" wheel is fixed while the direction of the "front" wheel can be controlled via the angle ϕ. The position of the geometric centre (between the point of contact of the two wheels) is denoted by (x, y). The orientation of the bicycle in the plane is determined by the angle θ. The "back" wheel is used to drive the bicycle, and so the forward velocity of the point of contact of the back wheel is something that can be controlled. Also, the rotational velocity of the "front" steering wheel can be controlled. The output is the position of the geometric centre.

We wish to assemble all of this into a continuous-time state space system.

(a) What is the state space X for the system?
(b) What is the control set U for the system?
(c) What is the time-domain $\mathbb{T}$ for the system?
(d) What is a good choice for the space $\mathscr{U}$ of inputs?
(e) What are the dynamics $\boldsymbol{f}$?
(f) What is the output map $\boldsymbol{h}$?

Any physical parameters you require, you should introduce yourself. Answer the following questions about the model.

(g) Is the system model causal?
(h) Is the system model stationary?
(i) Is the system model memoryless?
(j) Is the system model control-affine?

Finally, do some system theoretic explorations as follows.

(k) Do some research and describe three system theoretic problems that arise in a natural way for the problem.
(l) Using a computer package for simulating ordinary differential equations, setup the system for simulation, and see if you can steer the

output from an initial to a desired final value. You should represent the output as a parametric curve in the output space $\mathbb{R}^2$.

6.1.11 *(Mini-project)* Consider the rolling ball on a beam shown in Figure 6.4 and

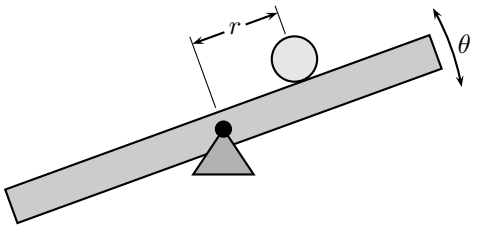


Figure 6.4 A ball rolling on a beam

discussed in [Keshmiri, Jahromi, Mohebbi, Amoozgar, and Xie 2012]. A motor supplies a torque to rotate the beam. A sensor measures the displacement of the ball along the beam.

We wish to assemble all of this into a continuous-time state space system.

(a) What is the state space X for the system?
(b) What is the control set U for the system?
(c) What is the time-domain $\mathbb{T}$ for the system?
(d) What is a good choice for the space $\mathscr{U}$ of inputs?
(e) What are the dynamics $\boldsymbol{f}$?
(f) What is the output map $\boldsymbol{h}$?

Any physical parameters you require, you should introduce yourself. Answer the following questions about the model.

(g) Is the system model causal?
(h) Is the system model stationary?
(i) Is the system model memoryless?
(j) Is the system model control-affine?

Finally, do some system theoretic explorations as follows.

(k) Do some research and describe three system theoretic problems that arise in a natural way for the problem.
(l) Using a computer package for simulating ordinary differential equations, setup the system for simulation, and see if you can devise a way to steer the ball to a displacement of $r = 0$ from an initial displacement of $r = r_0$.

6.1.12 *(Mini-project)* Consider the continuous fermenter depicted in Figure 6.5 and discussed in [Henson and Seborg 1992]. A substrate with concentration S_f is input to the constant volume fermenter and an effluent with cell-mass concentration X, a substrate with concentration S, and the product with

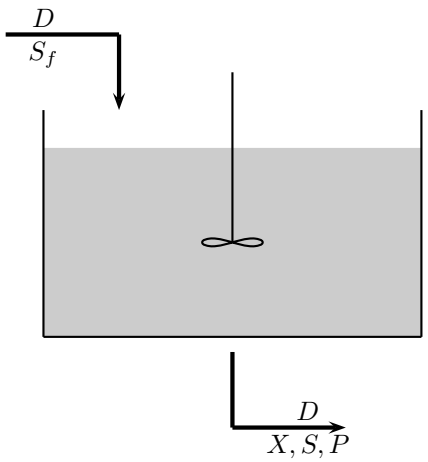


Figure 6.5 Continuous fermenter

concentration P are produced. The dilution rate D is regarded as an input. One may model the process with the ordinary differential equation model

$$\begin{aligned}
\dot{X}(t) &= -D(t)X(t) + \mu(S(t), P(t))X(t), \\
\dot{S}(t) &= D(t)(S_f(t) - S(t)) - Y_{X/S}\mu(S(t), P(t))X(t), \\
\dot{P}(t) &= -D(t)P(t) + (\alpha\mu(S(t), P(t)) + \beta)X(t),
\end{aligned}$$

where the specific growth rate μ is given by

$$\mu(S, P) = \frac{\mu_m\left(1 - \frac{P}{P_m}\right)S}{K_m + S + \frac{S^2}{K_i}},$$

and where the following are constant parameters determined by experiment:

$Y_{X/S}$	cell-mass yield,
α, β	yield parameters,
μ_m	maximum specific growth rate,
P_m	product saturation constant,
K_m	substrate saturation constant,
K_i	substrate inhibition constant.

We wish to assemble all of this into a continuous-time state space system.

(a) What is the state space X for the system?
(b) What is the control set U for the system?
(c) What is the time-domain $\mathbb{T}$ for the system?
(d) What is a good choice for the space $\mathscr{U}$ of inputs?
(e) What are the dynamics $\boldsymbol{f}$?
(f) What is the output map $\boldsymbol{h}$?

Any physical parameters you require, you should introduce yourself. Answer the following questions about the model.

(g) Is the system model causal?
(h) Is the system model stationary?
(i) Is the system model memoryless?
(j) Is the system model control-affine?

Finally, do some system theoretic explorations as follows.

(k) Do some research and describe three system theoretic problems that arise in a natural way for the problem.
(l) Using a computer package for simulating ordinary differential equations, setup the system for simulation, and try out some inputs, while interpreting the outputs.

Section 6.2

Continuous-time input/output systems

The next class of systems we consider are input/output systems, still in the setting of continuous-time systems. We shall see in this section the emergence of a theme in our treatment of system theory, namely that of an input/output system as a continuous mapping between a space of input signals to a space of output signals. We shall also see another these, namely that state space systems can be regarded as input/output systems. Note that this is connected with constructions in general system theory as exemplified by Propositions 2.1.7 and 2.1.13 (for general systems), and Theorem 2.2.20 and Proposition 2.2.49 (for general time systems).

Do I need to read this section? The ideas about input/output systems, and about the connection of such systems to state space systems, that are provided here are a theme in much of our presentation. This theme is enunciated in a somewhat general form for continuous-time systems in this section, and so this section is an important one for what follows. •

6.2.1 Topological constructions for spaces of continuous-time partially defined signals

As we briefly suggested above, input/output systems are maps between spaces of input and output signals. Because of the necessity of allowing signals defined on varying time-domains, cf. Example 2.2.21, this complicates things. Therefore, let us develop some methodology for dealing with this complication. First let us recall from Notation 2.2.7 some notation for distinguished sets of partially defined signals.

6.2.1 Definition (Spaces of partially defined signals with topological structure) Let $\mathbb{T} \subseteq \mathbb{R}$ be a continuous time-domain.

(i) Consider the space

$$\mathsf{C}^0((\mathbb{T}); \mathbb{R}^n) = \{f \in (\mathbb{R}^n)^{(\mathbb{T})} \mid f \in \mathsf{C}^0(\mathrm{dom}(f); \mathbb{R}^n)\},$$

where we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the seminorms

$$\|f\|_{\mathbb{K},\infty} = \sup\{|f_a(t)| \mid t \in \mathbb{K},\ a \in \{1, \dots, n\}\}, \quad \mathbb{K} \subseteq \mathbb{S} \text{ a compact interval.}$$

(ii) Consider the space

$$\mathsf{C}^0_{\mathrm{bdd}}((\mathbb{T}); \mathbb{R}^n) = \{f \in (\mathbb{R}^n)^{(\mathbb{T})} \mid f \in \mathsf{C}^0_{\mathrm{bdd}}(\mathrm{dom}(f); \mathbb{R}^n)\},$$

where we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the norm

$$\|f\|_\infty = \sup\{|f_a(t)| \mid t \in \mathbb{S},\ a \in \{1, \dots, n\}\}.$$

(iii) Consider the space

$$\mathsf{L}^\infty_{\mathrm{loc}}((\mathbb{T});\mathbb{R}^n) = \{\boldsymbol{f} \in (\mathbb{R}^n)^{(\mathbb{T})} \mid \boldsymbol{f} \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathrm{dom}(\boldsymbol{f});\mathbb{R}^n)\},$$

where we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the seminorms

$$\|\boldsymbol{f}\|_{\mathbb{K},\infty} = \max\{\operatorname{ess\,sup}\{|f_a(t)| \mid t \in \mathbb{K}\} \mid a \in \{1,\dots,n\}\}, \quad \mathbb{K} \subseteq \mathbb{S} \text{ a compact interval.}$$

(iv) Consider the space

$$\mathsf{L}^\infty((\mathbb{T});\mathbb{R}^n) = \{\boldsymbol{f} \in (\mathbb{R}^n)^{(\mathbb{T})} \mid \boldsymbol{f} \in \mathsf{L}^\infty(\mathrm{dom}(\boldsymbol{f});\mathbb{R}^n)\},$$

where we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the norm

$$\|\boldsymbol{f}\|_\infty = \max\{\operatorname{ess\,sup}\{|f_a(t)| \mid t \in \mathbb{S}\} \mid a \in \{1,\dots,n\}\}.$$

(v) For $p \in [1,\infty)$, consider the space

$$\mathsf{L}^p_{\mathrm{loc}}((\mathbb{T});\mathbb{R}^n) = \{\boldsymbol{f} \in (\mathbb{R}^n)^{(\mathbb{T})} \mid \boldsymbol{f} \in \mathsf{L}^p_{\mathrm{loc}}(\mathrm{dom}(\boldsymbol{f});\mathbb{R}^n)\},$$

where we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the seminorms

$$\|\boldsymbol{f}\|_{\mathbb{K},p} = \max\left\{\left(\int_{\mathbb{K})} |f_a(t)|^p \,\mathrm{d}t\right)^{1/p} \;\middle|\; a \in \{1,\dots,n\}\right\}, \quad \mathbb{K} \subseteq \mathbb{S} \text{ a compact interval.}$$

(vi) For $p \in [1,\infty)$, consider the space

$$\mathsf{L}^p((\mathbb{T});\mathbb{R}^n) = \{\boldsymbol{f} \in (\mathbb{R}^n)^{(\mathbb{T})} \mid \boldsymbol{f} \in \mathsf{L}^p(\mathrm{dom}(\boldsymbol{f});\mathbb{R}^n)\},$$

where we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the norm

$$\|\boldsymbol{f}\|_p = \max\left\{\left(\int_{\mathbb{S}} |f_a(t)|^p \,\mathrm{d}t\right)^{1/p} \;\middle|\; a \in \{1,\dots,n\}\right\}.$$ •

Note that the preceding sets of partially defined signals are not, themselves, topological spaces. They are merely collections of subsets of signals, each having topologies.

The spaces we shall use are then the following subsets of the preceding spaces.

6.2.2 Definition (Space of partially defined continuous-time signals with topology) Let $\mathbb{T} \subseteq \mathbb{R}$ be a continuous time-domain and let $S \subseteq \mathbb{R}^n$. A ***space of partially defined signals with topology*** is a subset $\mathscr{S}$ of one of the following spaces of partially defined signals:

(i) the space

$$\mathsf{C}^0((\mathbb{T});S) = \{\boldsymbol{f} \in \mathsf{C}^0((\mathbb{T});\mathbb{R}^n) \mid \boldsymbol{f}(t) \in S,\ t \in \mathrm{dom}(\boldsymbol{f})\}$$

equipped with the subspace topology;

(ii) the space

$$\mathsf{C}^0_{\text{bdd}}((\mathbb{T}); S) = \{f \in \mathsf{C}^0_{\text{bdd}}((\mathbb{T}); \mathbb{R}^n) \mid \ f(t) \in S, \ t \in \operatorname{dom}(f)\}$$

equipped with the subspace topology;

(iii) the space

$$\mathsf{L}^\infty_{\text{loc}}((\mathbb{T}); S) = \{f \in \mathsf{L}^\infty_{\text{loc}}((\mathbb{T}); \mathbb{R}^n) \mid \ \text{for each compact sub-time-domain } \mathbb{K} \subseteq \operatorname{dom}(f), \text{ there exists a relatively compact } K \subseteq S \text{ such that } f(t) \in K, \text{ a.e. } t \in \mathbb{K}\}$$

equipped with the subspace topology;[2]

(iv) the space

$$\mathsf{L}^\infty((\mathbb{T}); S) = \{f \in \mathsf{L}^\infty((\mathbb{T}); \mathbb{R}^n) \mid \ \text{there exists a relatively compact } K \subseteq S \text{ such that } f(t) \in K, \text{ a.e. } t \in \operatorname{dom}(f)\}$$

equipped with the subspace topology;[3]

(v) for $p \in [1, \infty)$, the space

$$\mathsf{L}^p_{\text{loc}}((\mathbb{T}); S) = \{f \in \mathsf{L}^p_{\text{loc}}((\mathbb{T}); \mathbb{R}^n) \mid \ f(t) \in S, \ t \in \operatorname{dom}(f)\}$$

equipped with the subspace topology;

(vi) for $p \in [1, \infty)$, the space

$$\mathsf{L}^p((\mathbb{T}); S) = \{f \in \mathsf{L}^p((\mathbb{T}); \mathbb{R}^n) \mid \ f(t) \in S, \ t \in \operatorname{dom}(f)\}$$

equipped with the subspace topology.

If $\operatorname{dom}(f) = \mathbb{T}$ for every $f \in \mathscr{S}$, then $\mathscr{S}$ is a ***space of continuous-time signals with topology***. •

Given a space $\mathscr{S}$ of partially defined signals with topology with time-domain $\mathbb{T}$ and given a sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, we shall use the notation

$$\mathscr{S}(\mathbb{S}) = \{f \in \mathscr{S} \mid \ \operatorname{dom}(f) = \mathbb{S}\}.$$

6.2.2 Definitions and system theoretic properties

With suitable notions of spaces of partially defined signals at hand, we can give a suitable definition of an input/output system.

[2]There is a seeming lack of symmetry in this definition, as is does not altogether match our other definitions. However, it *does* match if one keeps in mind that $f \in \mathsf{L}^\infty(\mathbb{T}; \mathbb{R}^n)$ if and only if there exists a compact $K \subseteq \mathbb{R}^n$ for which $f(t) \in K$ for almost every $t \in \mathbb{T}$. It is the compactness of the set in which the control takes its values that is crucial, not just its boundedness.

[3]Ibid.

6.2.3 Definition (Continuous-time input/output system) A ***continuous-time input/output system*** is a quintuple $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$, where

(i) $U \subseteq \mathbb{R}^m$ (the ***input set***),

(ii) $\mathbb{T} \subseteq \mathbb{R}$ is an interval (the ***time-domain***),

(iii) $\mathscr{U} \subseteq U^{(\mathbb{T})}$ is a space of partially defined signals with topology (the ***input signals***),

(iv) $\mathscr{Y} \subseteq (\mathbb{R}^k)^{(\mathbb{T})}$ is a space of partially defined signals with topology (the ***output signals***), and

(v) $g\colon \mathscr{U} \to \mathscr{Y}$ has the following properties:

  (a) for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, the restriction of g to $\mathscr{U}(\mathbb{S})$, denoted by $g_{\mathbb{S}}$, takes values in $\mathscr{Y}(\mathbb{S})$;

  (b) if $\mathbb{S}, \mathbb{S}' \subseteq \mathbb{T}$ are sub-time-domains with $\mathbb{S}' \subseteq \mathbb{S}$, then $g_{\mathbb{S}}|\mathscr{U}(\mathbb{S}') = g_{\mathbb{S}'}$;

  (c) $g_{\mathbb{S}}$ is continuous for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$.

Moreover,

(iv) a pair (μ, η) with $\mu \in \mathscr{U}(\mathbb{S})$ and $\eta = g_{\mathbb{S}}(\mu)$ is a ***behaviour*** for Σ, and we denote by $\mathscr{B}(\Sigma)$ the set of behaviours. •

6.2.4 Remark (Restriction in continuous-time input/output systems) Note that we *do not* require that, if $\mathbb{S}, \mathbb{S}' \subseteq \mathbb{T}$ are sub-time-domains with $\mathbb{S}' \subseteq \mathbb{S}$ and if $\mu \in \mathscr{U}(\mathbb{S})$, then $\mu|\mathbb{S}' \in \mathscr{U}(\mathbb{S}')$. What we *do* require is that, if $\mu|\mathbb{S}' \in \mathscr{U}(\mathbb{S}')$, then

$$g_{\mathbb{S}'}(\mu|\mathbb{S}') = g_{\mathbb{S}}(\mu)|\mathbb{S}'.$$

If a continuous-time input/output system does have then property that $\mu|\mathbb{S}' \in \mathscr{U}(\mathbb{S}')$ for every pair of sub-time-domains satisfying $\mathbb{S}' \subseteq \mathbb{S}$ and for every $\mu \in \mathscr{U}(\mathbb{S})$, we shall say that the system is ***closed under restriction***.

Note that, by not requiring that continuous-time input/output systems be closed under restriction, we allow the common situation where all inputs and outputs are considered only as signals defined on the entire time-domain. That is to say, for a system like that, we have $\mathscr{U}(\mathbb{S}) = \varnothing$ and $\mathscr{Y}(\mathbb{S}) = \varnothing$ for every strict sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$. •

Let us connect some of the general systems ideas from Chapter 2 to our concept of a continuous-time inout/output system. Along the way, we shall give a few elementary examples of such systems. In Section 6.2.3 we shall see that all continuous-time state space systems are also continuous-time input.output systems.

We begin by making the connection to the basic types of general systems.

6.2.5 Remarks (Continuous-time input/output systems as general systems) We let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a continuous-time input/output system.

1. A continuous-time input/output system is a general input/output system as per Definition 2.1.3. To see this, take
   (a) "$\mathscr{U} = \mathscr{U}$," i.e., the inputs for the general input/output system are the same as the inputs for the continuous-time state space system,
   (b) "$\mathscr{Y} = \mathscr{Y}$," i.e., the outputs for the general input/output system are the same as the outputs for the continuous-time state space system, and
   (c) $\mathscr{B} = \{(\mu, g(\mu)) \mid \mu \in \mathscr{U}\}$, i.e., a continuous-time input/output system is a functional input/output system, as per Definition 2.1.4.
2. A continuous-time input/output system is, more specifically, a general time system as per Definition 2.2.9. To see this, take
   (a) "$U = U$," i.e., the input set for the general time system is the same as the input set for the continuous-time input/output system,
   (b) $Y = \mathbb{R}^k$, i.e., the output set for the general time system is $\mathbb{R}^k$,
   (c) "$\mathscr{U} = \mathscr{U}$," i.e., the admissible input signals for the general input/output system are the same as the input signals for the continuous-time input/output system,
   (d) $\mathscr{Y} = (\mathbb{R}^k)^{(\mathbb{T})}$, i.e., the admissible output signals for the general input/output system are the partial $\mathbb{R}^k$-valued functions on $\mathbb{T}$, and
   (e) $\mathscr{B} = \{(\mu, g(\mu)) \mid \mu \in \mathscr{U}\}$ i.e., the behaviours for the general time system input/output pairs for the continuous-time input/output system. •

Let us now consider the matter of output completeness and completeness for continuous-time input/output systems.

6.2.6 Remarks (Completeness for continuous-time input/output systems) We let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a continuous-time input/output system.

1. *Continuous-time input/output systems are output complete:* Let $\mu \in \mathscr{U}$ and let $(I, \preceq)$ be a totally ordered set, and let $(\eta_i)_{i \in I}$ be a family of outputs satisfying conditions (a)–(f) of Definition 2.2.12. Note that
$$\eta_i(t) = g(\mu)(t), \qquad t \in \operatorname{dom}(\eta_i).$$
Now let $\mathbb{S} = \cup_{i \in I} \operatorname{dom}(\eta_i)$ and let $\eta \colon \mathbb{S} \to \mathbb{R}^k$ be such that $\eta_{\operatorname{dom}(\eta_i)} = \eta_i$, $i \in I$. Then, if $t \in \mathbb{S}$, we must have $t \in \operatorname{dom}(\eta_i)$ for some $i \in I$. Therefore,
$$\eta(t) = \eta_i(t) = g(\mu)(t).$$
As this holds for every $t \in \operatorname{dom}(\eta)$, we conclude output completeness.
2. *Generally, a continuous-time input/output system is not complete:* In Theorem 6.2.10 we shall see that continuous-time state space systems are continuous-time input/output systems. Thus Example 2.2.21 gives an example of a continuous-time input/output system that is not complete. •

We know from general results, i.e., Theorem 2.2.20, a complete continuous-time input/output system has a dynamical systems representation specified by some response family and some family of state transition maps. Moreover, the proof of Theorem 2.2.20 gives an explicit construction of such a dynamical systems representation. The difficulty is that, in any given example, the resulting dynamical systems representation will not be meaningful (whatever might be the meaning of "meaningful"). Indeed, the matter of constructing a meaningful dynamical systems representation is something that, typically, one should think carefully about.

Now let us consider the various attributes for general time systems from Section 2.2, as they pertain to continuous-time input/output systems. We shall see that these notions do not hold, generally, and so are assumptions that must be made if one needs them. In order to connect the general time system discussion of Section 2.2 to the systems we consider here, let us make suitable definitions for the appropriate notions.

First we consider causality, where the definition captures the idea that the output at time t depends only on the input prior to time t.

6.2.7 Definition (Causality for continuous-time input/output systems) Let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a continuous-time input/output system.

(i) The system Σ is ***causal*** if, for every $\mu_1, \mu_2 \in \mathscr{U}$ with $\operatorname{dom}(\mu_1) = \operatorname{dom}(\mu_2)$ and for every $t \in \operatorname{dom}(\mu_1) = \operatorname{dom}(\mu_2)$,

$$\mu_1|(\mathbb{T}_{\leq t} \cap \operatorname{dom}(\mu_1)) = \mu_2|(\mathbb{T}_{\leq t} \cap \operatorname{dom}(\mu_2)) \implies g(\mu_1)(t) = g(\mu_2)(t).$$

(ii) The system Σ is ***strongly causal*** if, for every $\mu_1, \mu_2 \in \mathscr{U}$ with $\operatorname{dom}(\mu_1) = \operatorname{dom}(\mu_2)$ and for every $t \in \operatorname{dom}(\mu_1) = \operatorname{dom}(\mu_2)$,

$$\mu_1|(\mathbb{T}_{< t} \cap \operatorname{dom}(\mu_1)) = \mu_2|(\mathbb{T}_{< t} \cap \operatorname{dom}(\mu_2)) \implies g(\mu_1)(t) = g(\mu_2)(t). \quad \bullet$$

Next we consider stationarity. We let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a continuous-time input/output system. As we saw in Section 2.2.8, stationarity has to do, roughly, with shift-invariance. To make this clear, let us first carefully think about what we mean by shifting. Let X be a set and let $\mathscr{X} \subseteq X^{(\mathbb{T})}$ be a collection of partially defined signals. Let $a \in \mathbb{R}$. If $\xi \in \mathscr{X}$, denote by $\tau_a^* \xi$ the signal with domain

$$\operatorname{dom}(\tau_a^* \xi) = \{t \in \mathbb{T} \mid t - a \in \operatorname{dom}(\xi)\}$$

and given by $\tau_a^* \xi(t) = \xi(t - a)$. Note that we may well have $\operatorname{dom}(\tau_a^* \xi) = \emptyset$, in which case $\tau_a^* \xi$ is not defined, by convention.

With this notation, we have the following definitions regarding stationarity.

6.2.8 Definition (Stationarity for continuous-time input/output systems) Let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a continuous-time input/output system with $\sup \mathbb{T} = \infty$.

(i) The system Σ is ***stationary*** if $\tau_a^*(\mathscr{U}) \subseteq \mathscr{U}$ for every $a \in \mathbb{R}_{>0}$ and if, for every $\mu \in \mathscr{U}$,

$$g(\tau_a^* \mu) = \tau_a^* g(\mu).$$

(ii) The system Σ is ***strongly stationary*** if it is stationary and if, for every $a \in \mathbb{R}_{>0}$ and every $\mu \in \mathscr{U}$, there exists $\mu' \in \mathscr{U}$ such that

$$g(\mu) = g(\tau_a^* \mu').$$ •

With these definitions, we can make the following remarks.

6.2.9 Remarks (System theoretic attributes of continuous-time input/output systems) We let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a continuous-time input/output system.

1. Σ *is generally not causal:* To see this, we give a simple counterexample. We take $U = \mathbb{R}$, $\mathbb{T} = \mathbb{R}$, and let $\mathscr{U} = \mathsf{C}^0(\mathbb{R}; \mathbb{R})$, i.e., inputs are all continuous $\mathbb{R}$-valued functions on $\mathbb{R}$. We also take $\mathscr{Y} = \mathsf{C}^0(\mathbb{R}; \mathbb{R})$. The topologies for $\mathscr{U}$ and $\mathscr{Y}$ are as defined in Definition 6.2.2(i). Now define $g \colon \mathscr{U} \to \mathscr{Y}$ by $g(\mu)(t) = \mu(-t)$. Because we are only considering signals defined on all of $\mathbb{R}$, conditions (v)(a) and (v)(b) of a continuous-time input/output system are immediately satisfied. We claim that condition (v)(c) is also satisfied. Indeed, let $(\mu_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $\mathsf{C}^0(\mathbb{R}; \mathbb{R})$ converging to $\mu \in \mathsf{C}^0(\mathbb{R}; \mathbb{R})$. Let $\mathbb{K} \subseteq \mathbb{R}$ be a compact interval and let $\epsilon \in \mathbb{R}_{>0}$. Let
$$-\mathbb{K} = \{-t \mid t \in \mathbb{K}\}.$$
Then there exists $N \in \mathbb{Z}_{>0}$ such that
$$|\mu(t) - \mu_j(t)| < \epsilon, \qquad t \in -\mathbb{K},\ j \geq N.$$
Then we immediately have
$$|g(\mu)(t) - g(\mu_j)(t)| < \epsilon, \qquad t \in \mathbb{K},\ j \geq N,$$
giving convergence of $(g(\mu_j))_{j \in \mathbb{Z}_{>0}}$ to $g(\mu)$, and so giving continuity of g.
Now we show that the system is not causal. Let $\mu_1, \mu_2 \in \mathscr{U}$ be defined by
$$\mu_1(t) = \begin{cases} 1, & t \in \mathbb{R}_{<0}, \\ 0, & t \in \mathbb{R}_{\geq 0}, \end{cases} \qquad \mu_2(t) = 1, \qquad t \in \mathbb{R}.$$
Let $t \in \mathbb{R}_{<0}$ and note that
$$\mu_1 | \mathbb{R}_{\leq t} = \mu_2 | \mathbb{R}_{\leq t}.$$
However,
$$g(\mu_1)(t) = \mu_1(-t) = 0, \qquad g(\mu_2)(t) = \mu_2(-t) = 1,$$
and this demonstrates the lack of causality.
One can show that, if Σ is closed under restriction as defined in Remark 6.2.4, then this implies that the system is causal. We invite the reader to prove this assertion as Exercise 6.2.1.
2. Σ *is generally not past determined:* This follows since, as proved in Proposition 2.2.35, past determined systems are causal.

3. Σ *is finitely observable:* This is a consequence of the fact that Σ, as a general input/output system, is functional.
4. Σ *is not generally stationary:* To see this, we note that continuous-time state space systems are continuous-time input-output systems by Theorem 6.2.10. Therefore, since continuous-time state space systems are not generally stationary (as we pointed out in Remark 6.1.8–4).
5. Σ *is not generally linear:* Presumably, since in Section 6.7 we shall specifically consider linear continuous-time input/output systems, it is not the case that all continuous-time input/output systems are linear. To see this, one need only produce a counterexample, and such an example can be seen in Example 2.2.21, and the lack of linearity here is a reflection of the fact that, in the formula (2.4) for controlled trajectories, the input μ does not appear linearly. •

6.2.3 Continuous-time state space systems as continuous-time input/output systems

As we saw in our discussion above of the system theoretic attributes for continuous-time input/output systems, these systems were capable of exhibiting characteristics that are not possible for continuous-time state space systems. In this section we show how the various classes of continuous-time state space systems are also continuous-time input/output systems.

First let us informally associate to a continuous-time state space system its candidate input/output system. Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a continuous-time state space system. If one thinks about the controlled outputs for Σ, one sees that these behaviours do not form the basis for a continuous-time input/output system since there are multiple outputs for a single input. To rectify this, one should choose an initial condition. Thus let $(t_0, x_0) \in \mathbb{T} \times X$. Then we can try to associate a continuous-time input/output system Σ for this initial condition data by the quintuple $\Sigma_{\mathrm{i/o}}(t_0, x_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$, where

1. "$U = U$,"
2. "$\mathbb{T} = \mathbb{T}$,"
3. "$\mathscr{U} = \mathscr{U}$,"
4. $\mathscr{Y} \subseteq (\mathbb{R}^k)^{(\mathbb{T})}$, and
5. $g(\mu)(t) = h(t, \Phi^{\Sigma}(t, t_0, x_0, \mu), \mu(t))$ for $t \in J_{\Sigma}(t_0, x_0, \mu)$.

This does not quite yet define a continuous-time input/output system since we must prescribe the structure of a space of partially defined signals with topology to both $\mathscr{U}$ and $\mathscr{Y}$. As we shall see, the appropriate such structure depends on the character of the system.

The following result characterises the various cases in which one can make the preceding association precise. In a few of the cases considered in the result, we assume that $U \subseteq \mathbb{R}^m$ is locally compact. We refer the reader to Definition II-1.2.65 for the definition, and we recall from Example II-1.2.66–1 and II-2 that open and

closed sets are locally compact. As we shall see in Lemma 2 in the proof, local compactness allows us to prove that controls eventually take values in a relatively compact subset of U.

6.2.10 Theorem (Continuous-time input/output systems from continuous-time state space systems) *Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a continuous-time state space system, and consider the following cases.*

The most general case: *Assume that Σ satisfies the hypotheses of Theorem 6.1.10 and that*

(i) $U \subseteq \mathbb{R}^m$ *is locally compact,*

(ii) *the map* $t \mapsto \mathbf{h}(t, \mathbf{x}, \mathbf{u})$ *is measurable for each* $(\mathbf{x}, \mathbf{u}) \in X \times U$,

(iii) *the map* $(\mathbf{x}, \mathbf{u}) \mapsto \mathbf{h}(t, \mathbf{x}, \mathbf{u})$ *is continuous for each* $t \in \mathbb{T}$, *and*

(iv) *for each* $(t, \mathbf{x}, \mathbf{u}) \in \mathbb{T} \times X \times U$, *there exist* $r_1, r_2, \alpha \in \mathbb{R}_{>0}$ *and*

$$g \in L^1([t - \alpha, t + \alpha]; \mathbb{R}_{\geq 0})$$

such that

$$\|\mathbf{h}(s, \mathbf{x}', \mathbf{u}')\| \leq g(s), \qquad (s, \mathbf{x}', \mathbf{u}') \in ([t - \alpha, t + \alpha] \cap \mathbb{T}) \times B^n(r_1, \mathbf{x}) \times B^n(r_2, \mathbf{u}).$$

Then $\Sigma_{i/o}(t_0, \mathbf{x}_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$, with $\mathbf{g}$ as defined above, defines a continuous-time input/output system for $\mathscr{U} \subseteq L^\infty_{loc}((\mathbb{T}); U)$ the space of partially defined signals with topology as in Definition 6.2.2(iii) and for $\mathscr{Y} = L^1_{loc}((\mathbb{T}); \mathbb{R}^k)$ the space of partially defined signals with topology as in Definition 6.2.2(v).

The output autonomous case: *Assume that Σ satisfies the hypotheses of Theorem 6.1.10 and that*

(i) $U \subseteq \mathbb{R}^m$ *is locally compact,*

(ii) Σ *is output autonomous, and*

(iii) *the map* $(\mathbf{x}, \mathbf{u}) \mapsto \mathbf{h}(\mathbf{x}, \mathbf{u})$ *is continuous.*

Then $\Sigma_{i/o}(t_0, \mathbf{x}_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$, with $\mathbf{g}$ as defined above, defines a continuous-time input/output system for $\mathscr{U} \subseteq L^\infty_{loc}((\mathbb{T}); U)$ the space of partially defined signals with topology as in Definition 6.2.2(iii) and for $\mathscr{Y} = L^\infty_{loc}((\mathbb{T}); \mathbb{R}^k)$ the space of partially defined signals with topology as in Definition 6.2.2(iii).

The output autonomous, proper case: *Assume that Σ satisfies the hypotheses of Theorem 6.1.10 and that*

(i) Σ *is output autonomous and proper, and*

(ii) *the map* $\mathbf{x} \mapsto \mathbf{h}(\mathbf{x})$ *is continuous.*

Then $\Sigma_{i/o}(t_0, \mathbf{x}_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$, with $\mathbf{g}$ as defined above, defines a continuous-time input/output system for $\mathscr{U} \subseteq L^\infty_{loc}((\mathbb{T}); U)$ the space of partially defined signals with topology as in Definition 6.2.2(iii) and $\mathscr{Y} = C^0((\mathbb{T}); \mathbb{R}^k)$ the space of partially defined signals with topology as in Definition 6.2.2(i).

The general control-affine case: *Assume that Σ satisfies the hypotheses of Theorem 6.1.16 and that*

(i) *the maps* $t \mapsto \mathbf{h}_a(t, \mathbf{x})$, $a \in \{0, 1, \dots, m\}$, *are measurable for each* $\mathbf{x} \in X$,

(ii) *the maps* $\mathbf{x} \mapsto \mathbf{h}_a(t, \mathbf{x})$, $a \in \{0, 1, \dots, m\}$, *are continuous for each* $t \in \mathbb{T}$, *and*

(iii) *for each* $(t, \mathbf{x}) \in \mathbb{T} \times X$, *there exist* $r, \alpha \in \mathbb{R}_{>0}$ *and*

$$g \in L^1_{\mathrm{loc}}([t - \alpha, t + \alpha]; \mathbb{R}_{\geq 0})$$

such that

$$\|\mathbf{h}_a(s, \mathbf{x}')\| \leq g(s), \qquad a \in \{0, 1, \dots, m\},\ (s, \mathbf{x}') \in ([t - \alpha, t + \alpha] \cap \mathbb{T} \times B^n(r, \mathbf{x}).$$

Then $\Sigma_{i/o}(t_0, \mathbf{x}_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$, *with* $\mathbf{g}$ *as defined above, defines a continuous-time input/output system for* $\mathscr{U} \subseteq L^\infty_{\mathrm{loc}}((\mathbb{T}); U)$ *the space of partially defined signals with topology as in Definition 6.2.2(iii) and for* $\mathscr{Y} = L^1_{\mathrm{loc}}((\mathbb{T}); \mathbb{R}^k)$ *the space of partially defined signals with topology as in Definition 6.2.2(v).*

The control-affine output autonomous case: *Assume that* Σ *satisfies the hypotheses of Theorem 6.1.16 and that*

(i) Σ *is output autonomous, and*

(ii) *the maps* $\mathbf{x} \mapsto \mathbf{h}_a(\mathbf{x})$, $a \in \{0, 1, \dots, m\}$, *are continuous.*

Then $\Sigma_{i/o}(t_0, \mathbf{x}_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$, *with* $\mathbf{g}$ *as defined above, defines a continuous-time input/output system for* $\mathscr{U} \subseteq L^\infty_{\mathrm{loc}}((\mathbb{T}); U)$ *the space of partially defined signals with topology as in Definition 6.2.2(iii) and for* $\mathscr{Y} = L^\infty_{\mathrm{loc}}((\mathbb{T}); \mathbb{R}^k)$ *the space of partially defined signals with topology as in Definition 6.2.2(iii).*

The control-affine output autonomous, proper case: *Assume that* Σ *satisfies the hypotheses of Theorem 6.1.16 and that*

(i) Σ *is output autonomous and proper, and*

(ii) *the map* $\mathbf{x} \mapsto \mathbf{h}_0(\mathbf{x})$ *is continuous.*

Then $\Sigma_{i/o}(t_0, \mathbf{x}_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$, *with* $\mathbf{g}$ *as defined above, defines a continuous-time input/output system for* $\mathscr{U} \subseteq L^\infty_{\mathrm{loc}}((\mathbb{T}); U)$ *the space of partially defined signals with topology as in Definition 6.2.2(iii) and* $\mathscr{Y} = C^0((\mathbb{T}); \mathbb{R}^k)$ *the space of partially defined signals with topology as in Definition 6.2.2(i).*

The control-affine autonomous case: *Assume that* Σ *satisfies the hypotheses of Theorem 6.1.17 and that,*

(i) Σ *is autonomous and*

(ii) *the maps* $\mathbf{x} \mapsto \mathbf{h}_a(\mathbf{x})$, $a \in \{0, 1, \dots, m\}$, *are continuous.*

Then $\Sigma_{i/o}(t_0, \mathbf{x}_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$, *with* $\mathbf{g}$ *as defined above, defines a continuous-time input/output system for* $\mathscr{U} \subseteq L^1_{\mathrm{loc}}((\mathbb{T}); U)$ *the space of partially defined signals with topology as in Definition 6.2.2(v) and for* $\mathscr{Y} = L^1_{\mathrm{loc}}((\mathbb{T}); \mathbb{R}^k)$ *the space of partially defined signals with topology as in Definition 6.2.2(v).*

The control-affine, autonomous, proper case: *Assume that* Σ *satisfies the hypotheses of Theorem 6.1.17 and that*

(i) Σ *is autonomous and proper, and*

(ii) the map $\mathbf{x} \mapsto \mathbf{h}_0(\mathbf{x})$ *is continuous.*

Then $\Sigma_{\mathrm{i/o}}(\mathrm{t}_0, \mathbf{x}_0) = (\mathrm{U}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$*, with* $\mathbf{g}$ *as defined above, defines a continuous-time input/output system for* $\mathscr{U} \subseteq \mathsf{L}^1_{\mathrm{loc}}((\mathbb{T}); \mathrm{U})$ *the space of partially defined signals with topology as in Definition 6.2.2(v) and for* $\mathscr{Y} = \mathsf{C}^0((\mathbb{T}); \mathbb{R}^k)$ *the space of partially defined signals with topology as in Definition 6.2.2(i).*

Proof The following lemma records an essential part of the proof.

1 Lemma *Let* $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ *be a continuous-time state space system satisfying the hypotheses of either of Theorems 6.1.10, 6.1.16, or 6.1.17 for existence and uniqueness of controlled trajectories. Let* $(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$*,* $\mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0}$ *satisfy* $(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) \in \mathrm{D}_\Sigma$*, let* $(\boldsymbol{\mu}_\mathrm{j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *be a sequence in* $\mathscr{U}$ *converging to* $\boldsymbol{\mu}_0$ *with respect to the seminorm* $\|\cdot\|_{[\mathrm{t}_0,\mathrm{t}],\infty}$ *(in the case of Theorem 6.1.10 or 6.1.16) or with respect to the seminorm* $\|\cdot\|_{[\mathrm{t}_0,\mathrm{t}],1}$ *(in the case of Theorem 6.1.17). Then the following statements hold:*

(i) there exists $\mathrm{N} \in \mathbb{Z}_{>0}$ *such that* $(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_\mathrm{j}) \in \mathrm{D}_\Sigma$ *for* $\mathrm{j} \geq \mathrm{N}$*;*

(ii) the sequence

$$\mathrm{s} \mapsto \Phi^\Sigma(\mathrm{s}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_\mathrm{j}), \qquad \mathrm{j} \in \mathbb{Z}_{>0},$$

of mappings in $\mathsf{C}^0([\mathrm{t}_0, \mathrm{t}]; \mathrm{X})$ *converges uniformly to*

$$\mathrm{s} \mapsto \Phi^\Sigma(\mathrm{s}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0).$$

Proof We will prove the result assuming that Σ satisfies the hypotheses of Theorem 6.1.10. The proof carries over directly to the other two cases, given that (1) Theorem 6.1.14 applies also to control-affine systems satisfying the hypotheses of Theorem 6.1.16 and (2) the conclusion (ix) of Theorem 6.1.18 holds for systems satisfying the hypotheses of Theorem 6.1.17.

The first assertion if a direct consequence of part (ix) of Theorem 6.1.14. We must, therefore, prove the uniform convergence conclusion of the second assertion.

Let $\epsilon \in \mathbb{R}_{>0}$. For $s \in [t_0, t]$, let us denote $\boldsymbol{x}_s = \Phi^\Sigma(s, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0)$. As we argued in the proof of Theorem 6.1.14, there exists $\alpha_s, r_s, \rho_s \in \mathbb{R}_{>0}$ such that

$$\|\Phi^\Sigma(\tau, s, \boldsymbol{x}, \boldsymbol{\mu}) - \boldsymbol{x}_s\| < \frac{\epsilon}{2}, \qquad \tau \in [s - \alpha_s, s + \alpha_s] \cap [t_0, t],\ \boldsymbol{x} \in \mathsf{B}^n(r_s, \boldsymbol{x}_s),\ \boldsymbol{\mu} \in \mathsf{B}_{[t_0,t]}(\rho_s, \boldsymbol{\mu}_0).$$

Let $N_s \in \mathbb{Z}_{>0}$ be sufficiently large that

$$\boldsymbol{\mu}_j \in \mathsf{B}_{[t_0,t]}(\rho_s, \boldsymbol{\mu}_0), \quad \Phi^\Sigma(s, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) \in \mathsf{B}^n(r_s, \boldsymbol{x}_s),$$

for $j \geq N_s$. Now, by compactness of $[t_0, t]$, let $s_1, \ldots, s_k \in [t_0, t]$ be such that

$$[t_0, t] \subseteq \cup_{l=1}^k [s_l - \alpha_{s_l}, s_l + \alpha_{s_l}].$$

Let $N = \max\{N_{s_1}, \ldots, N_{s_k}\}$. Let $s \in [t_0, t]$ and let $l \in \{1, \ldots, k\}$ be such that $s \in [s_l - \alpha_{s_l}, s_l + \alpha_{s_l}]$. Then, for $j \geq N$,

$$\begin{aligned}
&\|\Phi^\Sigma(s, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) - \Phi^\Sigma(s, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0)\| \\
&\qquad = \|\Phi^\Sigma(s, s_l, \Phi^\Sigma(s_l, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j), \boldsymbol{\mu}_j) - \Phi^\Sigma(s, s_l, \Phi^\Sigma(s_l, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \boldsymbol{\mu}_0)\| \\
&\qquad \leq \|\Phi^\Sigma(s, s_l, \Phi^\Sigma(s_l, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j), \boldsymbol{\mu}_j) - \boldsymbol{x}_{s_l}\| + \|\Phi^\Sigma(s, s_l, \Phi^\Sigma(s_l, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \boldsymbol{\mu}_0) - \boldsymbol{x}_{s_l}\| \\
&\qquad \leq \frac{\epsilon}{2} + \frac{\epsilon}{2},
\end{aligned}$$

giving the desired uniform convergence. ▼

In all cases in the theorem, we do the following. Let $\boldsymbol{\mu}_0 \in \mathscr{U}$ and let $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$ in the appropriate topology as in Definition 6.2.2. This means that we can assume that $\boldsymbol{\mu}_j$, $j \in \mathbb{Z}_{\geq 0}$, have a common domain, say $\mathbb{S}$, and that, for every compact $\mathbb{K} \subseteq \mathbb{S}$, we have

$$\lim_{j\to\infty} \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_0\|_{\mathbb{K}} = 0,$$

where $\|\cdot\|_{\mathbb{K}}$ means the appropriate seminorm from Definition 6.2.2 for the input signals. Given this, we shall show that, for every compact interval $\mathbb{L} \subseteq \mathbb{S}$, we have

$$\lim_{j\to\infty} \|\boldsymbol{\eta}_j - \boldsymbol{\eta}_0\|_{\mathbb{L}} = 0,$$

where

$$\boldsymbol{\eta}_j(t) = \boldsymbol{h}(t, \Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j), \boldsymbol{\mu}_j(t)), \qquad j \in \mathbb{Z}_{\geq 0},\ t \in \mathbb{S},$$

and where $\|\cdot\|_{\mathbb{L}}$ means the appropriate seminorm from Definition 6.2.2 for the output signals.

The following lemma will aid us in this proof strategy.

2 Lemma *Let $\mathbb{K} \subseteq \mathbb{R}$ be a compact time-domain, let $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{m}}$ be locally compact, and let $(\boldsymbol{\mu}_{\mathrm{j}})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ be a sequence in $\mathsf{L}^{\infty}(\mathbb{K};\mathrm{U})$ converging to $\boldsymbol{\mu}_0 \in \mathsf{L}^{\infty}(\mathbb{K};\mathrm{U})$. Then there exists $\mathrm{N} \in \mathbb{Z}_{>0}$, a subset $\mathrm{Z} \subseteq \mathbb{K}$ of measure zero, and a relatively compact subset $\mathrm{L} \subseteq \mathrm{U}$ such that $\boldsymbol{\mu}_{\mathrm{j}}(\mathrm{t}) \in \mathrm{L}$ for $\mathrm{j} \geq \mathrm{N}$ and $\mathrm{t} \in \mathbb{K} \setminus \mathrm{Z}$.*

Proof By definition of $\mathsf{L}^{\infty}(\mathbb{K}; U)$, there exists a relatively compact $L' \subseteq U$ and $Z_0 \subseteq \mathbb{K}$ such that $\boldsymbol{\mu}_0(t) \in L'$ for $t \in \mathbb{K} \setminus Z_0$. For $\boldsymbol{u} \in L'$, there is a relative neighbourhood $N_{\boldsymbol{u}} \subseteq U$ of $\boldsymbol{u}$. Thus there exists $\epsilon_{\boldsymbol{u}} \in \mathbb{R}_{>0}$ such that $\mathsf{B}^m(3\epsilon_{\boldsymbol{u}}, \boldsymbol{u}) \subseteq N_{\boldsymbol{u}}$ and so $\overline{\mathsf{B}}^m(2\epsilon_{\boldsymbol{u}}, \boldsymbol{u}) \subseteq N_{\boldsymbol{u}}$. By compactness of L', let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in L'$ be such that $L' = \cup_{j=1}^k \mathsf{B}^m(\epsilon_{\boldsymbol{u}_j}, \boldsymbol{u}_j)$. By the Lebesgue Number Lemma, let $r \in \mathbb{R}_{>0}$ be such that, for each $\boldsymbol{u} \in L'$, there exists $j \in \{1, \ldots, k\}$ for which $\mathsf{B}^m(r, \boldsymbol{u}) \subseteq \mathsf{B}^m(\epsilon_{\boldsymbol{u}_j}, \boldsymbol{u}_j)$. ref

Define $L = \cup_{j=1}^k \overline{\mathsf{B}}^m(\epsilon_{\boldsymbol{u}_j}, \boldsymbol{u}_j)$. Note that L is a compact subset of $\mathbb{R}^m$ and so is a relatively compact subset of U by Proposition II-1.2.59. By Proposition III-3.8.50, let $Z_1 \subseteq \mathbb{K}$ be such that $(\boldsymbol{\mu}_j|(\mathbb{K} \setminus Z_1))_{j\in\mathbb{Z}_{>0}}$ converges uniformly to $\boldsymbol{\mu}_0|(\mathbb{K} \setminus Z_1)$. Let $N \in \mathbb{Z}_{>0}$ be sufficiently large that

$$\|\boldsymbol{\mu}_0(t) - \boldsymbol{\mu}_j(t)\| < r, \qquad j \geq N,\ t \in (\mathbb{K} \setminus Z_1).$$

The lemma follows with N and L as defined, and with $Z = Z_0 \cup Z_1$. ▼

The most general case: Let $\boldsymbol{\mu}_0 \in \mathscr{U}$ and let $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$ in the topology as in Definition 6.2.2(iii). Let $\mathbb{L} \subseteq \mathbb{S}$ be compact. Let $t_1 \in \mathbb{S}$ be such that $\mathbb{L} \subseteq [t_0, t_1]$. By Lemma 2, suppose that there exists $L \subseteq U$ relatively compact and $Z \subseteq [t_0, t_1]$ of measure zero such that $\boldsymbol{\mu}_j(t) \in L$ for every $j \in \mathbb{Z}_{\geq 0}$ and $t \in [t_0, t_1] \setminus Z$. By the uniform convergence of Lemma 1, we can also assume that there is a compact $K \subseteq X$ such that

$$\Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) \in K, \qquad j \in \mathbb{Z}_{\geq 0},\ t \in [t_0, t_1].$$

Let $\boldsymbol{p} = (s, \boldsymbol{x}, \boldsymbol{u}) \in [t_0, t_1] \times K \times L$ and let $\alpha_{\boldsymbol{p}}, r_{1,\boldsymbol{p}}, r_{2,\boldsymbol{p}} \in \mathbb{R}_{>0}$ and

$$g_{\boldsymbol{p}} \in \mathsf{L}^1([s - \alpha_{\boldsymbol{p}}, s + \alpha_{\boldsymbol{p}}]; \mathbb{R}_{\geq 0})$$

be such that

$$\|\boldsymbol{h}(s', \boldsymbol{x}', \boldsymbol{u}')\| \leq g_{\boldsymbol{p}}(s'), \qquad (s', \boldsymbol{x}', \boldsymbol{u}') \in [s - \alpha_{\boldsymbol{p}}, s + \alpha_{\boldsymbol{p}}] \times \mathsf{B}^n(r_{1,\boldsymbol{p}}, \boldsymbol{x}) \times \mathsf{B}^m(r_{2,\boldsymbol{p}}, \boldsymbol{u}).$$

By compactness of $[t_0, t_1] \times K \times L$, let $\boldsymbol{p}_j = (s_j, \boldsymbol{x}_j, \boldsymbol{u}_j) \in [t_0, t_1] \times K \times L$ such that

$$[t_0, t_1] \times K \times L \subseteq \cup_{j=1}^k ([s_j - \alpha_{\boldsymbol{p}_j}, s_j + \alpha_{\boldsymbol{p}_j}] \times \mathsf{B}^n(r_{1,\boldsymbol{p}_j}, \boldsymbol{p}_j) \times \mathsf{B}^m(r_{2,\boldsymbol{p}_j}, \boldsymbol{u}_j))$$

Let $g \in \mathsf{L}^1([t_0, t_1]; \mathbb{R}_{\geq 0})$ be such that

$$g(s) = \max\{g_{\boldsymbol{p}_j}(s) \mid \ j \in \{1, \dots, k\} \text{ are such that } s \in [s_j - \alpha_{\boldsymbol{p}_j}, s_j + \alpha_{\boldsymbol{p}_j}]\}.$$

Then, for $(t, \boldsymbol{x}, \boldsymbol{u}) \in [t_0, t_1] \times K \times L$, we have

$$\|\boldsymbol{h}(t, \boldsymbol{x}, \boldsymbol{u})\|_{\mathbb{R}^n} \leq g(t).$$

Therefore, we can use the Dominated Convergence Theorem, continuity of $\boldsymbol{h}$ with respect to $\boldsymbol{x}$ and $\boldsymbol{u}$, and continuity of the flow with respect to control (Theorem 6.1.14(viii)) to conclude that

$$\lim_{j\to\infty} \|\boldsymbol{\eta}_j(t) - \boldsymbol{\eta}_0(t)\| = \lim_{j\to\infty} \int_{t_0}^{t_1} \|\boldsymbol{h}(t, \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j), \boldsymbol{\mu}_j(t)) - \boldsymbol{h}(t, \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \boldsymbol{\mu}_0(t))\| \, \mathrm{d}t = 0,$$

as desired.

The output autonomous, proper case: Let $\boldsymbol{\mu}_0 \in \mathscr{U}$ and let $(\boldsymbol{\mu}_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$ in the topology as in Definition 6.2.2(iii). Let $\mathbb{L} \subseteq \mathbb{S}$ be compact. Let $t_1 \in \mathbb{S}$ be such that $\mathbb{L} \subseteq [t_0, t_1]$. By the uniform convergence of Lemma 1, there is a compact set $K \subseteq X$ such that

$$\{\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) \mid \ t \in [t_0, t_1]\} \subseteq \operatorname{int}(K)$$

for sufficiently large $j \in \mathbb{Z}_{>0}$. We can assume, therefore, that this inclusion holds for some compact K and for every $j \in \mathbb{Z}_{>0}$, without loss of generality. Since $\boldsymbol{h} \colon X \to \mathbb{R}^k$ is continuous, it is uniformly continuous when restricted to K, by the Heine–Cantor Theorem (Theorem II-1.3.33). Therefore, for $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{x}_1, \boldsymbol{x}_2 \in K$ satisfy $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| < \delta$, then $\|\boldsymbol{h}(\boldsymbol{x}_1) - \boldsymbol{h}(\boldsymbol{x}_2)\| < \epsilon$. By Lemma 1, choose $N \in \mathbb{Z}_{>0}$ such that

$$\|\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) - \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0)\| < \delta, \qquad t \in [t_0, t_1], \ j \geq N.$$

Then, for $j \geq N$ and $t \in [t_0, t_1]$, we have

$$\|\boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j)) - \boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0))\| < \epsilon.$$

This gives the desired result that

$$\lim_{j\to\infty} \|\boldsymbol{\eta}_j - \boldsymbol{\eta}_0\|_{\mathbb{L},\infty} = 0.$$

The output autonomous case: Again, let $\boldsymbol{\mu}_0 \in \mathscr{U}$ and let $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$ in the topology as in Definition 6.2.2(iii). Let $\mathbb{L} \subseteq \mathbb{S}$ be compact. Let $t_1 \in \mathbb{S}$ be such that $\mathbb{L} \subseteq [t_0, t_1]$. By Lemma 2, let $L \subseteq U$ be relatively compact and let $Z \subseteq [t_0, t_1]$ have measure zero such that $\boldsymbol{\mu}_j(t) \in L$ for every $t \in [t_0, t_1] \setminus Z$. As in the preceding two parts of the proof, we can suppose that there is a compact set $K \subseteq X$ such that

$$\{\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) \mid t \in [t_0, t_1]\} \subseteq \operatorname{int}(K), \qquad j \in \mathbb{Z}_{\geq 0}.$$

Since $\boldsymbol{h}$ is continuous, $\boldsymbol{h}|K \times L$ is uniformly continuous as in the preceding part of the proof. Therefore, for $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \|\boldsymbol{u}_1 - \boldsymbol{u}_2\| < \delta \quad \implies \quad \|\boldsymbol{h}(\boldsymbol{x}_1, \boldsymbol{u}_1) - \boldsymbol{h}(\boldsymbol{x}_2, \boldsymbol{u}_2)\| < \frac{\epsilon}{2}.$$

Choose $N \in \mathbb{Z}_{>0}$ such that

$$\|\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) - \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0)\|, \|\boldsymbol{\mu}_j(t) - \boldsymbol{\mu}_0(t)\| < \delta, \qquad j \geq N,\ t \in [t_0, t_1] \setminus Z.$$

Then we have, for $j \geq N$ and $t \in [t_0, t_1] \setminus Z$,

$$\begin{aligned}
&\|\boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j), \boldsymbol{\mu}_j) - \boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \boldsymbol{\mu}_0)\| \\
&\qquad \leq \|\boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j), \boldsymbol{\mu}_j) - \boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \boldsymbol{\mu}_j)\| \\
&\qquad\quad + \|\boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \boldsymbol{\mu}_j) - \boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \boldsymbol{\mu}_0)\| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

This gives the desired conclusion that

$$\lim_{j\to\infty} \|\boldsymbol{\eta}_j - \boldsymbol{\eta}_0\|_{\mathbb{L},\infty} = 0,$$

where

$$\boldsymbol{\eta}_j(t) = \boldsymbol{h}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j), \boldsymbol{\mu}_j), \qquad j \in \mathbb{Z}_{\geq 0},\ t \in \mathbb{S}.$$

This is the desired conclusion.

The general control-affine case: Let $\boldsymbol{\mu}_0 \in \mathscr{U}$ and let $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$ in the topology as in Definition 6.2.2(iii). Let $\mathbb{L} \subseteq \mathbb{S}$ be compact. Let $t_1 \in \mathbb{S}$ be such that $\mathbb{L} \subseteq [t_0, t_1]$. By Proposition III-3.8.50, let $Z \subseteq [t_0, t_1]$ have measure zero such that $(\boldsymbol{\mu}_j|([t_0, t_1] \setminus Z))_{j\in\mathbb{Z}_{>0}}$ converges uniformly to $\boldsymbol{\mu}_0|([t_0, t_1] \setminus Z)$. Let $K \subseteq X$ be compact such that

$$\{\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) \mid t \in [t_0, t_1]\} \subseteq \operatorname{int}(K), \qquad j \in \mathbb{Z}_{\geq 0}.$$

We can argue as in the proof of the most general case above (but the argument is simpler since we do not have dependence on control) that there exists $g \in \mathsf{L}^1([t_0, t_1]; \mathbb{R}_{\geq 0})$ such that

$$\|\boldsymbol{h}_a(t, \boldsymbol{x})\| \leq g(t), \qquad (t, \boldsymbol{x}) \in [t_0, t_1] \times K.$$

Let $A \in \mathbb{R}_{>0}$ be such that

$$|\mu_{j,a}(t)| \leq A, \qquad t \in [t_0, t_1] \setminus Z,\ j \in \mathbb{Z}_{>0},\ a \in \{0, 1, \ldots, m\},$$

(this is possible since the sequence $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ converges uniformly on $[t_0, t_1] \setminus Z$). Then

$$\|\boldsymbol{h}(t, \boldsymbol{x}, \boldsymbol{u})\| \leq A(1 + mg(t))$$

if $t \in [t_0, t_1] \setminus Z$, $\boldsymbol{x} \in K$, and if $\boldsymbol{u}$ satisfies $|u_a| < A$, $a \in \{1, \dots, m\}$. One can now complete the proof just as in the most general case.

The control-affine output autonomous case: Let $\boldsymbol{\mu}_0 \in \mathscr{U}$ and let $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$ in the topology as in Definition 6.2.2(iii). Let $\mathbb{L} \subseteq \mathbb{S}$ be compact. Let $t_1 \in \mathbb{S}$ be such that $\mathbb{L} \subseteq [t_0, t_1]$. By Proposition III-3.8.50, let $Z \subseteq [t_0, t_1]$ have measure zero such that $(\boldsymbol{\mu}_j|([t_0, t_1] \setminus Z))_{j\in\mathbb{Z}_{>0}}$ converges uniformly to $\boldsymbol{\mu}_0|([t_0, t_1] \setminus Z)$. Let $K \subseteq X$ be compact such that

$$\{\Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) \mid t \in [t_0, t_1]\} \subseteq \operatorname{int}(K), \qquad j \in \mathbb{Z}_{\geq 0}.$$

Since $\boldsymbol{h}_a$, $a \in \{0, 1, \dots, m\}$, are continuous, they are uniformly continuous when restricted to K. Let $A, B \in \mathbb{R}_{>0}$ be such that

$$|\mu_{j,a}(t)| \leq A, \qquad t \in [t_0, t_1] \setminus Z,\ j \in \mathbb{Z}_{>0},\ a \in \{0, 1, \dots, m\},$$

(this is possible since the sequence $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ converges uniformly on $[t_0, t_1] \setminus Z$) and

$$\|\boldsymbol{h}_a(\boldsymbol{x})\| \leq B, \qquad a \in \{0, 1, \dots, m,\ \boldsymbol{x} \in K.$$

(by compactness of K and continuity of $\boldsymbol{h}_a$, $a \in \{0, 1, \dots, m\}$). By uniform continuity, let $\delta \in \mathbb{R}_{>0}$ be such that, if $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| < \delta$, then

$$\|\boldsymbol{h}_0(\boldsymbol{x}_1) - \boldsymbol{h}_0(\boldsymbol{x}_2)\| < \frac{\epsilon}{4}.$$

Also suppose that δ has the property that, if $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| < \delta$ then

$$\|\boldsymbol{h}_a(\boldsymbol{x}_1) - \boldsymbol{h}_a(\boldsymbol{x}_2)\| \leq \frac{mA\epsilon}{8}, \qquad a \in \{1, \dots, m\}.$$

and, if $\|\boldsymbol{u}_1 - \boldsymbol{u}_2\| < \delta$, then

$$|u_{1,a} - u_{2,a}| < \frac{mB\epsilon}{8}, \qquad a \in \{1, \dots, m\}.$$

Then, if $\boldsymbol{x}_1, \boldsymbol{x}_2 \in K$ and $\boldsymbol{u}_1, \boldsymbol{u}_2$ satisfy $|u_{1,a}|, |u_{2,a}| \leq A$ for $a \in \{1, \dots, m\}$, and $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \|\boldsymbol{u}_1 - \boldsymbol{u}_2\| < \delta$, we have

$$\begin{aligned}
\|\boldsymbol{h}(\boldsymbol{x}_1, \boldsymbol{u}_1) - \boldsymbol{h}(\boldsymbol{x}_2, \boldsymbol{u}_2)\| &= \left\| \boldsymbol{h}_0(\boldsymbol{x}_1) - \boldsymbol{h}_0(\boldsymbol{x}_2) + \sum_{a=1}^{m} (u_{1,a}\boldsymbol{h}_a(\boldsymbol{x}_1) - u_{2,a}\boldsymbol{h}_a(\boldsymbol{x}_2)) \right\| \\
&\leq \|\boldsymbol{h}_0(\boldsymbol{x}_1) - \boldsymbol{h}_0(\boldsymbol{x}_2)\| + \sum_{a=1}^{m} |u_{1,a}| \|\boldsymbol{h}_a(\boldsymbol{x}_1) - \boldsymbol{h}_a(\boldsymbol{x}_2)\| \\
&\quad + \sum_{a=1}^{m} |u_{1,a} - u_{2,a}| \|\boldsymbol{h}_a(\boldsymbol{x}_2)\| \\
&\leq \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}.
\end{aligned}$$

The remainder of the proof then goes exactly as in the general output autonomous case.

The control-affine output autonomous, proper case: This follows from the general output autonomous, proper case, since the hypotheses in the current case directly imply those of the general output autonomous, proper case when specialised to control-affine systems.

The control-affine autonomous, proper case: This follows from Lemma 1 in the same manner as does the general output autonomous, proper case.

The control-affine autonomous case: let $\boldsymbol{\mu}_0 \in \mathscr{U}$ and let $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$ in the topology as in Definition 6.2.2(v) with $p = 1$. Let $\mathbb{L} \subseteq \mathbb{S}$ be compact. Let $t_1 \in \mathbb{S}$ be such that $\mathbb{L} \subseteq [t_0, t_1]$. By Lemma 1, we can take N sufficiently large that

$$|h_{0,l}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j)) - h_{0,l}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j))| < \frac{\epsilon}{3(t_1 - t_0)}, \qquad t \in [t_0, t_1],\ l \in \{1, \dots, k\}.$$

Since the sequence $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ converges to $\boldsymbol{\mu}_0$, the sequence is bounded by Proposition III-3.2.4. Thus there exists $M_1 \in \mathbb{R}_{>0}$ such that

$$\int_{t_0}^{t_1} |\mu_{j,a}(t)| \, \mathrm{d}t \le M_1, \qquad j \in \mathbb{Z}_{\ge 0},\ a \in \{1, \dots, m\}.$$

Similarly, by the uniform convergence of Lemma 1, there exists $M_2 \in \mathbb{R}_{>0}$ such that

$$\|\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j)\| \le M_2, \qquad j \in \mathbb{Z}_{\ge 0},\ t \in [t_0, t_1].$$

We can now choose N sufficiently large such that

$$|h_{a,l}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j)) - h_{a,l}(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0))| < \frac{\epsilon}{3mM_1(t_1 - t_0)},$$
$$j \ge N,\ a \in \{1, \dots, m\},\ l \in \{1, \dots, k\}, t \in [t_0, t_1],$$

and that

$$\int_{t_0}^{t_1} |\mu_{j,a}(t) - \mu_{0,a}(t)| \, \mathrm{d}t < \frac{\epsilon}{3mM_2}, \qquad j \ge N,\ a \in \{1, \dots, m\}.$$

Then we have, for $j \geq N$ and $l \in \{1, \dots, k\}$,

$$\begin{aligned}
&\int_{t_0}^{t_1} |h_l(\Phi^\Sigma(t, t_0, x_0, \mu_j), \mu_j) - h_l(\Phi^\Sigma(t, t_0, x_0, \mu_0), \mu_0)| \, \mathrm{d}t \\
&\qquad \leq \int_{t_0}^{t_1} |h_{0,l}(\Phi^\Sigma(t, t_0, x_0, \mu_j)) - h_{0,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0))| \, \mathrm{d}t \\
&\qquad\quad + \sum_{a=1}^{m} \int_{t_0}^{t_1} |\mu_{j,a}(t) h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_j)) - \mu_{0,l}(t) h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0))| \, \mathrm{d}t \\
&\qquad \leq \int_{t_0}^{t_1} |h_{0,l}(\Phi^\Sigma(t, t_0, x_0, \mu_j)) - h_{0,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0))| \, \mathrm{d}t \\
&\qquad\quad + \sum_{a=1}^{m} \int_{t_0}^{t_1} |\mu_{j,a}(t) h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_j)) - \mu_{j,a}(t) h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0))| \, \mathrm{d}t \\
&\qquad\quad + \sum_{a=1}^{m} \int_{t_0}^{t_1} |\mu_{j,a}(t) h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0)) - \mu_{0,a}(t) h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0))| \, \mathrm{d}t \\
&\qquad \leq \int_{t_0}^{t_1} |h_{0,l}(\Phi^\Sigma(t, t_0, x_0, \mu_j)) - h_{0,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0))| \, \mathrm{d}t \\
&\qquad\quad + \sum_{a=1}^{m} \int_{t_0}^{t_1} |\mu_{j,a}(t)| |h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_j)) - h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0))| \, \mathrm{d}t \\
&\qquad\quad + \sum_{a=1}^{m} \int_{t_0}^{t_1} |\mu_{j,a}(t) - \mu_{0,a}(t)| |h_{a,l}(\Phi^\Sigma(t, t_0, x_0, \mu_0))| \, \mathrm{d}t \\
&\qquad \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}.
\end{aligned}$$

This gives the desired conclusion that

$$\lim_{j \to \infty} \|\eta_j - \eta_0\|_{\mathbb{L},1} = 0,$$

where

$$\eta_j(t) = h(\Phi^\Sigma(t, t_0, x_0, \mu_j), \mu_j), \qquad j \in \mathbb{Z}_{\geq 0}, \ t \in \mathbb{S}.$$

This is the desired conclusion. ∎

6.2.4 Continuous-time differential input/output systems

Exercises

6.2.1 Let $\Sigma = (U, \mathbb{T}, \mathcal{U}, \mathcal{Y}, g)$ be a continuous-time input/output system. Show that, if Σ is closed under restriction as defined in Remark 6.2.4, then Σ is causal.

6.2.2 For the continuous-time input/output systems $\Sigma = (U, \mathbb{T}, \mathcal{U}, \mathcal{Y}, g)$ given, indicate whether they are causal, strongly causal, finitely observable from any $\tau \in \mathbb{T}_{>t_0}$, stationary, strongly stationary, and/or memoryless.

(a) Take
(i) $U = \mathbb{R}$,
(ii) $\mathbb{T} = \mathbb{R}$,
(iii) $\mathscr{U} = \mathsf{L}^1(\mathbb{R};\mathbb{R})$,
(iv) $\mathscr{Y} = \{0,1\}^{\mathbb{R}} \cap \mathsf{L}^\infty(\mathbb{R};\mathbb{R})$,
(v) $g(\mu)(t) = \begin{cases} 1, & \int_{-\infty}^t \mu(\tau)\,\mathrm{d}\tau \geq 1, \\ 0, & \text{otherwise.} \end{cases}$

(b) Take
(i) $U = \mathbb{R}$,
(ii) $\mathbb{T} = \mathbb{R}$,
(iii) $\mathscr{U} = \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$,
(iv) $\mathscr{Y} = \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$,
(v) $g(\mu)(t) = \mu(t^2)$.

(c) Take
(i) $U = \mathbb{R}$,
(ii) $\mathbb{T} = \mathbb{R}$,
(iii) $\mathscr{U} = \mathsf{C}^0(\mathbb{R};\mathbb{R})$,
(iv) $\mathscr{Y} = \mathsf{C}^0(\mathbb{R};\mathbb{R})$,
(v) $g(\mu)(t) = \mu(0)$.

(d) Take
(i) $U = \mathbb{R}$,
(ii) $\mathbb{T} = \mathbb{R}$,
(iii) $\mathscr{U} = \mathsf{L}^1(\mathbb{R};\mathbb{R})$,
(iv) $\mathscr{Y} = \mathsf{C}^0(\mathbb{R};\mathbb{R})$,
(v) $g(\mu)(t) = \int_{-\infty}^t \sin(\tau)\mu(\tau)\,\mathrm{d}\tau$.

6.2.3 Let $U, Y \subseteq \mathbb{R}$ be open sets, let $\mathbb{T} \subseteq \mathbb{R}$ be a continuous time-domain, let $n \in \mathbb{Z}_{>0}$, and let

$$F\colon \mathbb{T} \times Y \times \mathsf{L}^{\leq n}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) \times U \times \mathsf{L}^{\leq n-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) \to \mathbb{R}.$$

(See the beginning of Section 3.1.3 for notation.) Suppose that, for each

$$(t, y, y^{(1)}, \dots, y^{(n-1)}, u, u^{(1)}, \dots, u^{(n-1)}) \in \mathbb{T} \times Y \times \mathsf{L}^{\leq n-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}) \times U \times \mathsf{L}^{\leq n-1}_{\mathrm{sym}}(\mathbb{R};\mathbb{R}),$$

we can solve the equation

$$F(t, y, y^{(1)}, \dots, y^{(n-1)}, y^{(n)}, u, u^{(1)}, \dots, u^{(n-1)}) = 0$$

uniquely for $y^{(n)}$ and denote the unique solution by

$$y^{(n)} = \widehat{F}(t, y, y^{(1)}, \dots, y^{(n-1)}, u, u^{(1)}, \dots, u^{(n-1)}).$$

Suppose that $\widehat{F}$ is continuously differentiable[4] and consider the differential equation

$$\frac{\mathrm{d}^n\eta}{\mathrm{d}t^n}(t) = \widehat{F}\left(t, \eta(t), \frac{\mathrm{d}\eta}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^{n-1}\eta}{\mathrm{d}t^{n-1}}(t), \mu(t), \frac{\mathrm{d}\mu}{\mathrm{d}t}(t), \dots, \frac{\mathrm{d}^{n-1}\mu}{\mathrm{d}t^{n-1}}(t)\right).$$

Answer the following questions.

[4] We assume continuous differentiability for simplicity; less stringent hypotheses are possible, since we just need hypotheses that ensure the existence and uniqueness conditions of Theorem 3.2.8 are satisfied.

(a) Show that this determines a general time system as per Definition 2.2.9. Clearly identify the spaces of input and output signals.
(b) Argue that a natural choice of states for this system is

$$\xi_j(t) = \frac{\mathrm{d}^j \eta}{\mathrm{d}t^j}(t), \qquad j \in \{0, 1, \dots, n-1\}.$$

(c) Derive a continuous-time state space system for which the input/output relation is the same as the general time system from part (a) and for which the states are as in part (b).

6.2.4 For the given continuous-time state space systems

$$\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h}),$$

with $\mathscr{U}$ left unprescribed, and for $(t_0, \boldsymbol{x}_0) \in \mathbb{T} \times X$, indicate the appropriate input and output spaces $\mathscr{U}$ and $\mathscr{Y}$ that ensure that $\Sigma_{\mathrm{i/o}}(t_0, \boldsymbol{x}_0)$ is a continuous-time input/output system, i.e., for which the input/output map $\boldsymbol{g}$ has the continuity property of Definition 6.2.3(c).

Here are the systems with $\mathscr{U}$ left unspecified.

(a) Take
  (i) $X = \mathbb{R}$,
  (ii) $U = \mathbb{R}$,
  (iii) $\mathbb{T} = \mathbb{R}$,
  (iv) $f(t, x, u) = \tanh(t)x^2 + \sin(x)u$,
  (v) $h(t, x, u) = \mathsf{1}_{\geq 0}(t)\cos(x) + u$.
(b) Take
  (i) $X = \mathbb{R}$,
  (ii) $U = \mathbb{R}$,
  (iii) $\mathbb{T} = \mathbb{R}$,
  (iv) $f(t, x, u) = \tan^{-1}(t)x^2 + x\sin(x)u$,
  (v) $h(t, x, u) = x$.
(c) Take
  (i) $X = \mathbb{R}$,
  (ii) $U = \mathbb{R}$,
  (iii) $\mathbb{T} = \mathbb{R}$,
  (iv) $f(t, x, u) = \sin(t)x + u^2$,
  (v) $h(t, x, u) = x$.
(d) Take
  (i) $X = \mathbb{R}$,
  (ii) $U = \mathbb{R}$,
  (iii) $\mathbb{T} = \mathbb{R}$,
  (iv) $f(t, x, u) = \sin(t)xu$,
  (v) $h(t, x, u) = tx + u$.
(e) Take
  (i) $X = \mathbb{R}$,
  (ii) $U = \mathbb{R}$,
  (iii) $\mathbb{T} = \mathbb{R}$,
  (iv) $f(t, x, u) = x^2 + u$,
  (v) $h(t, x, u) = x + \sin(x)u$.
(f) Take

(i) $X = \mathbb{R}$, (iv) $f(t, x, u) = \tanh(t)x^2 + \sin(x)u$,
(ii) $U = \mathbb{R}$, (v) $h(t, x, u) = x + u$.
(iii) $\mathbb{T} = \mathbb{R}$,

(g) Take
(i) $X = \mathbb{R}$, (iv) $f(t, x, u) = x^2 + \cos(x)u$,
(ii) $U = \mathbb{R}$, (v) $h(t, x, u) = x + 3$.
(iii) $\mathbb{T} = \mathbb{R}$,

(h) Take
(i) $X = \mathbb{R}$, (iv) $f(t, x, u) = 2t + x^2u$,
(ii) $U = \mathbb{R}$, (v) $h(t, x, u) = x \sin(u)$.
(iii) $\mathbb{T} = \mathbb{R}$,

6.2.5 For an integrable function $f\colon [0, T] \to \mathbb{R}$ defined on the compact interval $[0, T]$, the ***mean*** is

$$\operatorname{mean}(f) = \frac{1}{T} \int_0^T f(t) \, \mathrm{d}t$$

and the ***standard deviation*** is

$$\operatorname{stddev}(f) = \frac{1}{T} \int_0^T (f(t) - \operatorname{mean}(f))^2 \, \mathrm{d}t.$$

Let us adopt the convention that $\operatorname{mean}(\mu) = \mu(0)$ and $\operatorname{stddev}(\mu)(0) = 0$ when $T = 0$.

Suppose that, given $\mu \in \mathsf{C}^0(\mathbb{R}_{\geq 0}; \mathbb{R})$, we define functions

$$\operatorname{mean}(\mu), \operatorname{stddev}(\mu) \in \mathsf{C}^0(\mathbb{R}_{\geq 0}; \mathbb{R})$$

by

$$\operatorname{mean}(\mu)(t) = \operatorname{mean}(\mu|[0, t]), \quad \operatorname{stddev}(\mu)(t) = \operatorname{stddev}(\mu|[0, t]).$$

Answer the following questions.

(a) Make this record of "running mean" and "running standard deviation" into a continuous-time input/output system $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \boldsymbol{g})$.
(b) Is the system causal? strongly causal? stationary? strongly stationary? memoryless?

6.2.6 We consider the input/output system that models the input of "accelerator pedal" and the output of "car velocity." We assume that the thrust applied to the car is proportional to the accelerator pedal input, so that the thrust force is αu is u represents the throttle angle. We suppose that there is an aerodynamic drag force proportional to the square of velocity, $-\beta v^2$, if v is the car velocity.

Answer the following questions.

(a) Provide a differential equation that models the velocity $t \mapsto v(t)$ given the throttle angle $t \mapsto \mu(t)$.
(b) Show that the mapping $v \mapsto v$ defines a continuous-time input/output system.
(c) Determine its system theoretic properties, i.e., is it causal? strongly causal? finitely observable? stationary? strongly stationary? memoryless?
(d) What is the state space for the system?
(e) What is the control set for the system?
(f) What is the time-domain for the system?
(g) What is a reasonable choice for the space $\mathscr{U}$ of input signals?
(h) If the maximum throttle angle is $u_{\max}$, what is the maximum speed $v_{\max}$ attainable by the car?
(i) Using the technique of separable ordinary differential equations from Section 4.1.1, obtain an explicit formula for $v(t)$, assuming that at time 0 the car has velocity v_0 and that the throttle angle is constant u_0 throughout.

Section 6.3

Discrete-time state space systems

In this and the next section, we turn our attention from continuous-time systems to discrete-time systems. We shall conduct a program for discrete-time systems that mirrors that in Sections 6.1 and 6.2 for continuous-time systems. We shall transition to notation differing from that for difference equations in Section 3.3, and Chapters 4 and 5. In order to merge our discussion with the presentation of signal theory in Chapters IV-1 and IV-7, and use "Δ" for the sampling interval, rather than "h" for the discretisation. For $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$, we shall denote

$$\mathbb{T}_{\text{free}} = \{t \in \mathbb{T} \mid t + \Delta \in \mathbb{T}\}$$

which will serve the rôle of the free domain $\mathbb{T}_F$ for difference equations in Section 3.3.

Do I need to read this section? As with the continuous-time state space systems presented in Section 6.1, the material in this section provides important context for linear systems, which will be our main focus in Chapters 7 and 8. •

6.3.1 Definitions and system theoretic properties

Let us introduce the basic object of study, recalling from Section 2.2.2 the notation concerning partially defined functions on time-domains.

6.3.1 Definition (Discrete-time state space system) A ***discrete-time state space system*** is a sextuple $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$, where

(i) $X \subseteq \mathbb{R}^n$ is an open set (the ***state space***),
(ii) $U \subseteq \mathbb{R}^m$ (the ***control set***),
(iii) $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ is a sub-time-domain (the ***time-domain***),
(iv) $\mathscr{U} \subseteq U^{(\mathbb{T})}$ (the ***control functions*** or ***controls***),
(v) $f\colon \mathbb{T} \times X \times U \to \mathbb{R}^n$ (the ***dynamics***), and
(vi) $h\colon \mathbb{T} \times X \times U \to \mathbb{R}^k$ (the ***output map***).

Associated with a discrete-time state space system Σ we have the following notions:

(vii) a ***controlled trajectory*** for Σ is a pair (ξ, μ), where $\mu \in \mathscr{U}$ and where $\xi \in X^{\text{dom}(\mu)}$ are such that

$$\xi(t + \Delta) = f(t, \xi(t), \mu(t)), \qquad t \in \text{dom}(\mu)_{\text{free}}; \tag{6.6}$$

(viii) a ***controlled output*** for Σ is a pair (η, μ), where $\mu \in \mathscr{U}$ and where $\eta \in (\mathbb{R}^k)^{\text{dom}(\mu)}$ satisfies

$$\eta(t) = h(t, \xi(t), \mu(t)), \qquad t \in \text{dom}(\mu),$$

for some controlled trajectory (ξ, μ).

We denote by $\mathrm{Ctraj}(\Sigma)$ the set of controlled trajectories and by $\mathrm{Cout}(\Sigma)$ the set of controlled outputs. •

Of course, since controlled trajectories are defined by solutions to a difference equation, one must make considerations for discrete-time state space systems that account the matter of existence and uniqueness of trajectories. As with the differences between ordinary differential and ordinary difference equations in this regard, there will be differences here between continuous- and discrete-time systems. Such considerations will be developed in the next section. Here we shall consider the system theoretic attributes of discrete-time state space systems; that is, we make reference to the general system theory of Chapter 2, and see which attributes apply to the systems of this section. In doing this, we will not consider the logical interrelations between the various notions, since part of the point of the discussion here is to see how one applies the definitions of Chapter 2.

Let us consider a few attributes of discrete-time state space systems that often arise in practice.

6.3.2 Definition (Autonomous, proper, invertible discrete-time state space systems) A discrete-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ is

(i) ***autonomous*** if there exists

$$\boldsymbol{f}_0 \colon X \times U \to \mathbb{R}^n, \quad \boldsymbol{h}_0 \colon X \times U \to \mathbb{R}^k$$

such that

$$\boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{u}), \quad \boldsymbol{h}(t, \boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{h}_0(\boldsymbol{x}, \boldsymbol{u})$$

for every $(t, \boldsymbol{x}, \boldsymbol{u}) \in \mathbb{T} \times X \times U$, is

(ii) ***proper*** if there exists $\boldsymbol{h}_0 \colon X \times U \to \mathbb{R}^k$ such that

$$\boldsymbol{h}(t, \boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{h}_0(t, \boldsymbol{x})$$

for every $(t, \boldsymbol{x}, \boldsymbol{u}) \in \mathbb{T} \times X \times U$, and is

(iii) ***invertible*** if the ordinary difference equation $\boldsymbol{F}_\mu$ with right-hand side

$$\begin{aligned} \widehat{\boldsymbol{F}}_\mu \colon \mathrm{dom}(\mu) \times X &\to \mathbb{R}^n \\ (t, \boldsymbol{x}) &\mapsto \boldsymbol{f}(t, \boldsymbol{x}, \mu(t)) \end{aligned}$$

is invertible for every $\mu \in \mathscr{U}$. •

If only $\boldsymbol{f}$ (resp. $\boldsymbol{h}$) satisfies the conditions for the system to be autonomous, we shall say that the system is ***dynamically autonomous*** (resp. ***output autonomous***).

We shall see the system theoretic significance of these notions shortly.

Indeed, we next indicate whether/how a discrete-time state space system is a system of the various types introduced in Chapter 2.

6.3.3 Remarks (Discrete-time state space systems as general systems) We let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a discrete-time state space system.

1. A discrete-time state space system is a general input/output system as per Definition 2.1.3. To see this, take
   (a) "$\mathscr{U} = \mathscr{U}$," i.e., the inputs for the general input/output system are the same as the controls for the discrete-time state space system,
   (b) $\mathscr{Y} = (\mathbb{R}^k)^{(\mathbb{T})}$, i.e., the outputs for the general input/output system are the partial $\mathbb{R}^k$-valued functions on $\mathbb{T}$, and
   (c) $\mathscr{B} = \mathrm{Cout}(\Sigma)$, i.e., the behaviours for the general input/output system are exactly the controlled outputs for the discrete-time state space system.
2. A discrete-time state space system is, more specifically, a general time system as per Definition 2.2.9. To see this, take
   (a) "$U = U$," i.e., the input set for the general time system is the same as the control set for the discrete-time state space system,
   (b) $Y = \mathbb{R}^k$, i.e., the output set for the general time system is $\mathbb{R}^k$,
   (c) "$\mathscr{U} = \mathscr{U}$," i.e., the admissible input signals for the general input/output system are the same as the controls for the discrete-time state space system,
   (d) $\mathscr{Y} = (\mathbb{R}^k)^{(\mathbb{T})}$, i.e., the admissible output signals for the general input/output system are the partial $\mathbb{R}^k$-valued functions on $\mathbb{T}$, and
   (e) $\mathscr{B} = \mathrm{Cout}(\Sigma)$, i.e., the behaviours for the general time system are exactly the controlled outputs for the discrete-time state space system. •

Next we consider the issue of various forms of completeness for discrete-time systems, as introduced in a general setting in Section 2.2.4.

6.3.4 Remarks (Completeness for discrete-time state space systems) We let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a discrete-time state space system. In our constructions here, we make use of the notation for flows of discrete-time state space systems introduced in Definition 6.3.11.

1. *Discrete-time state space systems are output complete:* Let $\boldsymbol{\mu} \in \mathscr{U}$ and let $(I, \preceq)$ be a totally ordered set, and let $(\boldsymbol{\eta}_i)_{i \in I}$ be a family of outputs satisfying conditions (a)–(f) of Definition 2.2.12. Note that

$$\boldsymbol{\eta}_i(t) = \boldsymbol{h}(t, \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(t)), \qquad t \in \mathrm{dom}(\boldsymbol{\eta}_i).$$

   Now let $\mathbb{S} = \cup_{i \in I} \mathrm{dom}(\boldsymbol{\eta}_i)$ and let $\boldsymbol{\eta} \colon \mathbb{S} \to \mathbb{R}^k$ be such that $\boldsymbol{\eta}_{\mathrm{dom}(\boldsymbol{\eta}_i)} = \boldsymbol{\eta}_i$, $i \in I$. Then, if $t \in \mathbb{S}$, we must have $t \in \mathrm{dom}(\boldsymbol{\eta}_i)$ for some $i \in I$. Therefore,

$$\boldsymbol{\eta}(t) = \boldsymbol{\eta}_i(t) = \boldsymbol{h}(t, \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(t)).$$

   As this holds for every $t \in \mathrm{dom}(\boldsymbol{\eta})$, we conclude output completeness.

2. *Generally, a discrete-time state space system is not complete:* As with difference equations, the reasons for lack of completeness are not as exotic for discrete-time state space systems as for continuous-time state space systems. Indeed, completeness has to do merely with whether f is X-valued. •

Next let us give the form for the general time system representations of Section 2.2 for discrete-time state space systems.

6.3.5 Remarks (General time system representations for discrete-time state space systems) We let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a discrete-time state space system. In our constructions here, we make use of the notation for flows of discrete-time state space systems introduced in Definition 6.3.11. In the following, we suppose that Σ is complete.

1. Σ *has an initial response function:* Let $t_0 \in \mathbb{T}$ and let $(\eta, \mu) \in \mathrm{Cout}(\Sigma)$ with

$$\eta(t) = h(t, \xi(t), \mu(t)), \qquad t \in \mathrm{dom}(\mu),$$

for $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma)$. Suppose that $t_0 \in \mathrm{dom}(\mu)$. Then

$$\xi(t) = \Phi^{\Sigma}(t, t_0, x_0, \mu)$$

for some $x_0 \in \mathrm{dom}(\mu) \times X$. We can then denote

$$\rho^{\Sigma}_{t_0}(x_0, \mu)(t) = h(t, \Phi^{\Sigma}(t, t_0, x, \mu), \mu(t)),$$

and this defines the initial response function $\rho^{\Sigma}_{t_0}$ from t_0 with initial state object X. One has to verify the conditions of Definition 2.2.14, and this is straightforward. Note that we require completeness in order to ensure the existence of Φ^{Σ} for all arguments.

2. Σ *has a family of state transition maps:* Let $t_0 \in \mathbb{T}$ and, given $t_1, t_2 \in \mathbb{T}_{\geq t_0}$, we take $X_{t_1} = X_{t_2} = X$ and define

$$\Phi_{t_2,t_1}(\mu, x_1) = \Phi^{\Sigma}(t_2, t_1, x_1, \mu),$$

defining the family of state transition maps. The properties of flows enunciated in Proposition 3.2.12 ensure that the conditions of Definition 2.2.15 are satisfied, and we leave the elementary verification of this to the reader. (Indeed, it is the conclusions of Proposition 3.2.12 that explain the conditions of Definition 2.2.15.)

Again, we see that completeness is required.

3. Σ *has a dynamical system representation:* The response function and the family of state transition maps above combine to give a dynamical systems representation at $t_0 \in \mathbb{T}$, as per Definition 2.2.19. One can readily verify the conditions of Definition 2.2.19.

One can show that this dynamical system representation is full if and only if Σ is invertible. This is a consequence of definition of invertibility in Definition 3.4.5.

As in the preceding two items, completeness is obviously required.

4. Σ *has a state space representation:* As output function at $t_0 \in \mathbb{T}$, as per Definition 2.2.24, is simply given by

$$\gamma^{\Sigma}_{t,t_0}(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{h}(t, \boldsymbol{x}, \boldsymbol{u}).$$

One readily verifies that the conditions of Definition 2.2.24 are satisfied. •

Now let us see which of the general time system theoretic attributes of Section 2.2 are held by a discrete-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$. In order to make the connections to the general time system notions of Section 2.2 to the specific case here, we translate these notions into language applicable to the class of system we consider here. In our definitions, we make use of the notation for flows of discrete-time state space systems introduced in Definition 6.3.11.

We begin with causality.

6.3.6 Proposition (Causality for discrete-time state space systems) *Let* $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ *be a discrete-time state space system and let* $\mathrm{t}_0 \in \mathbb{T}$.

(i) The system Σ *is* ***causal*** *from* t_0 *if, for every* $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathscr{U}$ *and every* $\mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0} \cap \operatorname{dom}(\boldsymbol{\mu}_1) \cap \operatorname{dom}(\boldsymbol{\mu}_2)$,

$$\mu_1|[\mathrm{t}_0, \mathrm{t}] = \mu_2|[\mathrm{t}_0, \mathrm{t}] \implies \mathbf{h}(\mathrm{t}, \Phi^{\Sigma}(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_1)) = \mathbf{h}(\mathrm{t}, \Phi^{\Sigma}(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_2))$$

for every $\mathbf{x}_0 \in \mathrm{X}$.

(ii) The system Σ *is* ***strongly causal*** *from* t_0 *if, for every* $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathscr{U}$ *and every* $\mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0} \cap \operatorname{dom}(\boldsymbol{\mu}_1) \cap \operatorname{dom}(\boldsymbol{\mu}_2)$,

$$\mu_1|[\mathrm{t}_0, \mathrm{t}) = \mu_2|[\mathrm{t}_0, \mathrm{t}) \implies \mathbf{h}(\mathrm{t}, \Phi^{\Sigma}(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_1)) = \mathbf{h}(\mathrm{t}, \Phi^{\Sigma}(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_2))$$

for every $\mathbf{x}_0 \in \mathrm{X}$. •

Now we consider stationarity.

6.3.7 Proposition (Stationarity for discrete-time state space systems) *Let* $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ *be a discrete-time state space system and let* $\mathrm{t}_0 \in \mathbb{T}$.

(i) The system Σ *is* ***stationary*** *from* t_0 *if* $\tau^{*}_{\mathrm{t}_0, \mathrm{t}_0+\mathrm{a}}(\mathscr{U}) \subseteq \mathscr{U}$ *for every* $\mathrm{a} \in \mathbb{Z}_{>0}(\Delta)$ *and if, for every* $\boldsymbol{\mu} \in \mathscr{U}$ *and every* $\mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0} \cap \operatorname{dom}(\boldsymbol{\mu})$,

$$\mathbf{h}(\mathrm{t}+\mathrm{a}, \Phi^{\Sigma}(\mathrm{t}+\mathrm{a}, \mathrm{t}_0+\mathrm{a}, \mathbf{x}_0, \tau^{*}_{\mathrm{t}_0, \mathrm{t}_0+\mathrm{a}}\boldsymbol{\mu}), \tau^{*}_{\mathrm{t}_0, \mathrm{t}_0+\mathrm{a}}\boldsymbol{\mu}(\mathrm{t})) = \mathbf{h}(\mathrm{t}, \Phi^{\Sigma}(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(\mathrm{t}))$$

for every $\mathrm{a} \in \mathbb{Z}_{>0}(\Delta)$ *and every* $\mathbf{x}_0 \in \mathrm{X}$.

(ii) The system Σ *is* ***strongly stationary*** *from* t_0 *if it is stationary from* t_0 *and if, for every* $\mathrm{a} \in \mathbb{Z}_{>0}(\Delta)$, *every* $\mathbf{x}_0 \in \mathrm{X}$, *and every* $\boldsymbol{\mu} \in \mathscr{U}$, *there exists* $\mathbf{x}'_0 \in \mathrm{X}$ *such that*

$$\mathbf{h}(\mathrm{t}, \Phi^{\Sigma}(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(\mathrm{t}))\mathbf{h}(\mathrm{t}+\mathrm{a}, \Phi^{\Sigma}(\mathrm{t}+\mathrm{a}, \mathrm{t}_0+\mathrm{a}, \mathbf{x}'_0, \tau^{*}_{\mathrm{t}_0, \mathrm{t}_0+\mathrm{a}}\boldsymbol{\mu}(\mathrm{t})), \tau^{*}_{\mathrm{t}_0, \mathrm{t}_0+\mathrm{a}}\boldsymbol{\mu}(\mathrm{t})).$$ •

Note that a consequence of this definition of stationarity is that $\sup \mathbb{T} = \infty$.

With these definitions, we have the following statements.

6.3.8 Remarks (System theoretic attributes of discrete-time state space systems) We let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a discrete-time state space system. In our constructions here, we make use of the notation for flows of discrete-time state space systems introduced in Definition 6.3.11.

1. Σ *is causal and sometimes strongly causal:* Let $t_0 \in \mathbb{T}$. It follows from the formula (6.6) for controlled trajectories that, if controls $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ agree on $[t_0, t]$, then the controlled trajectories for Σ on $[t_0, t]$ agree. Thus Σ is causal from t_0. If, additionally, $\boldsymbol{h}$ is independent of U, i.e., Σ is proper, then we claim that Σ is also strongly causal from t_0. Indeed, if $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ agree on $[t_0, t)$, then the controlled trajectories with the same initial condition agree on $[t_0, t]$ as a consequence of (6.6). Note, however, that if $\boldsymbol{h}$ *does* depend on control, then we generally have causality, but not strong causality.
2. Σ *is sometimes past determined:* First of all, the definition of being past-determined requires completeness, so one needs to assume completeness to make any statements about past-determinacy. Thus we do this. This ensures that the the first part of the definitions of past-determined and strong past-determined holds.
   Now let $t_0 \in \mathbb{T}$. For the second parts of these definitions, considerations such as those for causality above allow us to conclude that Σ is past-determined from any $\tau \in \mathbb{T}_{>t_0}$, and is strongly past-determined if $\boldsymbol{h}$ is independent of control.
3. Σ *is finitely observable:* Let $t_0 \in \mathbb{T}$ and let $\tau \in \mathbb{T}_{>t_0}$. Then we see that Σ is finitely observable from τ. Indeed, a controlled trajectory on $[t_0, \tau)$, for a fixed control $\boldsymbol{\mu} \in \mathscr{U}$, is uniquely determined by the initial state. This uniqueness then applies to uniqueness for times greater than τ.
4. *Conditions for* Σ *to be stationary:* Generally, a discrete-time state space system is not stationary. However, it is most common to consider systems that *are* stationary, so we consider such systems here. First of all, the definition of stationarity from $t_0 \in \mathbb{T}$ requires that $\tau^*_{t-t_0}(\mathscr{U}_{\geq t_0}) = \mathscr{U}_{\geq t_0}$, i.e., the set of controls is shift-invariant. Then one sees that Σ is stationary if it is autonomous. The argument for this follows along the lines of that for doing Exercise 3.3.5, and we leave the working out of this to the reader. Note that, unlike their continuous-time counterparts, discrete-time state space systems are not generally strongly stationary. This is a result of their not having the same "reversibility" property that continuous-time state space systems possess, cf. the fact that in Proposition 6.3.12(iii), one requires invertibility of the system. See Exercise 6.3.2 for a specific example of this.
5. Σ *is not generally linear:* Presumably, since in Section 6.8 we shall specifically consider linear discrete-time state space systems, it is not the case that all discrete-time state space systems are linear. To see this, one need only produce a counterexample, and such examples abound; see Exercise 6.3.3. •

6.3.2 Existence and uniqueness of controlled trajectories, and flows for discrete-time state space systems

We now turn our attention to the matter of existence and uniqueness of controlled trajectories for discrete-time state space systems. As with ordinary difference equations when compared to ordinary differential equations, there are far fewer technicalities to concern ourselves with for discrete-time state space systems as compared to the conditions of Theorem 6.1.10. First of all, we do not have to worry ourselves with the precise character of controls, although we *will* concern ourselves with topologies for spaces of controls when we consider discrete-time state space systems as discrete-time input/output systems in Section 6.4.3.

Let us jump right to the statement of the existence and uniqueness results.

6.3.9 Theorem (Existence and uniqueness of controlled trajectories for discrete-time state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a discrete-time state space system. Then, for $\boldsymbol{\mu} \in \mathscr{U}$ and $(\mathrm{t}_0, \mathbf{x}_0) \in \mathbb{T} \times \mathrm{X}$, there exists a sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}$ and $\boldsymbol{\xi} \colon \mathbb{T}' \to \mathrm{X}$ such that $\boldsymbol{\xi}(\mathrm{t}_0) = \mathbf{x}_0$ and such that $(\boldsymbol{\xi}, \boldsymbol{\mu}|\mathbb{T}') \in \mathrm{Ctraj}(\Sigma)$. Moreover, if $\mathbb{T}''$ is another such sub-time-domain and $\boldsymbol{\eta} \colon \mathbb{T}'' \to \mathrm{X}$ is another such mapping, then $\boldsymbol{\eta}(\mathrm{t}) = \boldsymbol{\xi}(\mathrm{t})$ for all $\mathrm{t} \in \mathbb{T}'' \cap \mathbb{T}'$. Finally, if $\mathbf{f}$ takes values in X, then the preceding conclusions hold for any sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}_{\geq \mathrm{t}_0}$.*

Proof The theorem follows in the same manner as does Theorem 3.4.2, since the mapping

$$\mathrm{dom}(\boldsymbol{\mu}) \times X \ni (x, \boldsymbol{x}) \mapsto \boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{\mu}(t)) \in \mathbb{R}^n$$

is an ordinary difference equation. ■

As with ordinary difference equations, a controlled trajectory for a discrete-time state space system will generally be defined only for times in $\mathbb{T}_{\geq t_0}$ if one prescribes an initial condition at t_0. This is a consequence of the fact that a general discrete-time ordinary difference equation is not invertible. If the system *is* invertible and complete, however, then every controlled trajectory associated with a control $\boldsymbol{\mu}$ will exist on the entirety of $\mathrm{dom}(\boldsymbol{\mu})$, cf. Theorem 3.4.6.

The theorem now permits an adaptation of the notion of the flow of a difference equation in Section 3.4.1.2 to discrete-time state space systems. Let us undertake this notation here.

6.3.10 Definition (Interval of existence, domain of solutions) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a discrete-time state space system.

(i) For $(t_0, \boldsymbol{x}_0, \boldsymbol{\mu}) \in \mathbb{T} \times X \times \mathscr{U}$, denote

$$J_\Sigma(t_0, \boldsymbol{x}_0, \boldsymbol{\mu}) = \cup \{J \subseteq \mathrm{dom}(\boldsymbol{\mu}) \mid \begin{array}{l} J \text{ is a sub-time-domain and there exists} \\ \qquad \boldsymbol{\xi} \colon J \to X \text{ such that } (\boldsymbol{\xi}, \boldsymbol{\mu}|J) \in \mathrm{Ctraj}(\Sigma),\ \boldsymbol{\xi}(t_0) = \boldsymbol{x}_0\}. \end{array}$$

The sub-time-domain $J_\Sigma(t_0, \boldsymbol{x}_0, \boldsymbol{\mu})$ is the ***interval of existence*** for the initial value problem

$$\boldsymbol{\xi}(t + \Delta) = \boldsymbol{f}(t, \boldsymbol{\xi}(t), \boldsymbol{\mu}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x}_0.$$

(ii) For $\boldsymbol{\mu} \in \mathscr{U}$, the ***domain of solutions*** for Σ for the control $\boldsymbol{\mu}$ is

$$D_\Sigma(\boldsymbol{\mu}) = \{(t, t_0, \boldsymbol{x}_0) \in \mathbb{T} \times \mathbb{T} \times X \mid t \in J_\Sigma(t_0, \boldsymbol{x}_0, \boldsymbol{\mu})\}.$$

(iii) The ***domain of solutions*** for Σ is

$$D_\Sigma = \{(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}) \in \mathbb{T} \times \mathbb{T} \times X \times \mathscr{U} \mid (t, t_0, \boldsymbol{x}_0) \in D_\Sigma(\boldsymbol{\mu})\}. \qquad \bullet$$

We can now introduce the notion of a flow for a discrete-time state space system.

6.3.11 Definition (Flow of a discrete-time state space system) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a discrete-time state space system. The ***flow*** of Σ is the map $\Phi^\Sigma \colon D_\Sigma \to X$ defined by asking that $\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu})$ is the solution, evaluated at t, of the initial value problem

$$\boldsymbol{\xi}(\tau + \Delta) = \boldsymbol{f}(\tau, \boldsymbol{\xi}(\tau), \boldsymbol{\mu}(\tau)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x}_0. \qquad \bullet$$

The definition, phrased differently, says that

$$\Phi^\Sigma(t + \Delta, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}) = \boldsymbol{f}(t, \Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}), \boldsymbol{\mu}(t)), \quad \Phi^\Sigma(t_0, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}) = \boldsymbol{x}_0.$$

For $t, t_0 \in \mathbb{T}$ and $\boldsymbol{\mu} \in \mathscr{U}$, it is sometimes convenient to denote

$$D_\Sigma(t, t_0, \boldsymbol{\mu}) = \{\boldsymbol{x} \in X \mid (t, t_0, \boldsymbol{x}) \in D_\Sigma(\boldsymbol{\mu})\},$$

and then

$$\begin{aligned} \Phi^{\Sigma, \boldsymbol{\mu}}_{t, t_0} \colon D_\Sigma(t, t_0, \boldsymbol{\mu}) &\to X \\ \boldsymbol{x} &\mapsto \Phi^\Sigma(t, t_0, \boldsymbol{x}, \boldsymbol{\mu}). \end{aligned}$$

Along similar lines, for $t_0 \in \mathbb{T}$, we denote

$$D_\Sigma(t_0) = \{(t, \boldsymbol{x}, \boldsymbol{\mu}) \in \mathbb{T} \times X \times \mathscr{U} \mid (t, t_0, \boldsymbol{x}, \boldsymbol{\mu}) \in D_\Sigma\},$$

and then

$$\begin{aligned} \Phi^\Sigma(t_0) \colon D_\Sigma(t_0) &\to X \\ (t, \boldsymbol{x}, \boldsymbol{\mu}) &\mapsto \Phi^\Sigma(t, t_0, \boldsymbol{x}, \boldsymbol{\mu}). \end{aligned}$$

Finally, for $t, t_0 \in \mathbb{T}$, we denote

$$D_\Sigma(t, t_0) = \{(\boldsymbol{x}, \boldsymbol{\mu}) \in X \times \mathscr{U} \mid (t, t_0, \boldsymbol{x}, \boldsymbol{\mu}) \in D_\Sigma\},$$

and then

$$\begin{aligned} \Phi^\Sigma(t, t_0) \colon D_\Sigma(t, t_0) &\to X \\ (\boldsymbol{x}, \boldsymbol{\mu}) &\mapsto \Phi^\Sigma(t, t_0, \boldsymbol{x}, \boldsymbol{\mu}). \end{aligned}$$

Let us enumerate some of the more elementary properties of the flow for a discrete-time state space system, just as for an ordinary difference equation.

6.3.12 Proposition (Elementary properties of flows of discrete-time state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a discrete-time state space system. Then the following statements hold:*

(i) for each $(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$, $(\mathrm{t}_0, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) \in \mathrm{D}_\Sigma$ and $\Phi^\Sigma(\mathrm{t}_0, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) = \mathbf{x}_0$;

(ii) if, for $\mathrm{t}_1, \mathrm{t}_2 \in \mathbb{T}$ with $\mathrm{t}_1 \leq \mathrm{t}_2$, $(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \boldsymbol{\mu}) \in \mathrm{D}_\Sigma$, then, for $\mathrm{t}_3 \in \mathbb{T}$ with $\mathrm{t}_2 \leq \mathrm{t}_3$, $(\mathrm{t}_3, \mathrm{t}_2, \Phi^\Sigma(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}), \boldsymbol{\mu}) \in \mathrm{D}_\Sigma$ if and only if $(\mathrm{t}_3, \mathrm{t}_1, \mathbf{x}, \boldsymbol{\mu}) \in \mathrm{D}_\Sigma$ and, if this holds, then

$$\Phi^\Sigma(\mathrm{t}_3, \mathrm{t}_1, \mathbf{x}, \boldsymbol{\mu}) = \Phi^\Sigma(\mathrm{t}_3, \mathrm{t}_2, \Phi^\Sigma(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \boldsymbol{\mu}), \boldsymbol{\mu}).$$

(iii) if Σ is invertible and if $(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \boldsymbol{\mu}) \in \mathrm{D}_\Sigma$, then $(\mathrm{t}_1, \mathrm{t}_2, \Phi^\Sigma(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \boldsymbol{\mu}), \boldsymbol{\mu}) \in \mathrm{D}_\Sigma$ and

$$\Phi^\Sigma(\mathrm{t}_1, \mathrm{t}_2, \Phi^\Sigma(\mathrm{t}_2, \mathrm{t}_1, \mathbf{x}, \boldsymbol{\mu}), \boldsymbol{\mu}) = \mathbf{x}.$$

Proof This follows immediately from Proposition 3.4.7. ■

Useful mnemonics associated with parts (i)–(iii) are:

$$\Phi^{\Sigma,\boldsymbol{\mu}}_{t_0,t_0} = \mathrm{id}_\mathrm{X}, \quad (\Phi^{\Sigma,\boldsymbol{\mu}}_{t_2,t_1})^{-1} = \Phi^{\Sigma,\boldsymbol{\mu}}_{t_1,t_2}, \quad \Phi^{\Sigma,\boldsymbol{\mu}}_{t_3,t_2} \circ \Phi^{\Sigma,\boldsymbol{\mu}}_{t_2,t_1} = \Phi^{\Sigma,\boldsymbol{\mu}}_{t_3,t_1}.$$

However, these really are just mnemonics, since they do not account carefully for the domains of the mappings being used. Moreover, the second requires invertibility of the system, and the third, generally, must respect the order $t_1 \leq t_2 \leq t_3$.

The matter of regularity of flows for discrete-time state space systems is, like the corresponding theory for ordinary difference equations when compared to ordinary differential equations, substantially simpler than that for continuous-time state space systems given in Theorem 6.1.14.

6.3.13 Theorem (Properties of flows of discrete-time state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a discrete-time state space system and assume that $\mathbf{f}$ is continuous. Then the following statements hold:*

(i) for $(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$, $\mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu})$ is a sub-time-domain of $\mathbb{T}$;

(ii) for $(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$, the curve

$$\begin{aligned} \gamma_{(\mathrm{t}_0,\mathbf{x}_0,\boldsymbol{\mu})} \colon \mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) &\to \mathrm{X} \\ \mathrm{t} &\mapsto \Phi^\Sigma(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}) \end{aligned}$$

is well-defined and continuous;

(iii) for $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}$ with $\mathrm{t} \geq \mathrm{t}_0$ and for $\boldsymbol{\mu} \in \mathscr{U}$, $\mathrm{D}_\Sigma(\mathrm{t}, \mathrm{t}_0, \boldsymbol{\mu})$ is open;

(iv) for $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}$ with $\mathrm{t} \geq \mathrm{t}_0$ and for $\boldsymbol{\mu} \in \mathscr{U}$ for which $\mathrm{D}_\Sigma(\mathrm{t}, \mathrm{t}_0, \boldsymbol{\mu}) \neq \emptyset$, $\Phi^\Sigma_{\mathrm{t},\mathrm{t}_0}$ is continuous;

(v) for $\mathrm{t}_0 \in \mathbb{T}$, $\mathrm{D}_\Sigma(\mathrm{t}_0)$ is relatively open in $\mathbb{T} \times \mathrm{X} \times \mathscr{U}$;

(vi) for $\mathrm{t}_0 \in \mathbb{T}$, the map

$$\begin{aligned} \Phi^\Sigma(\mathrm{t}_0) \colon \mathrm{D}_\Sigma(\mathrm{t}_0) &\to \mathrm{X} \\ (\mathrm{t}, \mathbf{x}, \boldsymbol{\mu}) &\mapsto \Phi^\Sigma(\mathrm{t}, \mathrm{t}_0, \mathbf{x}, \boldsymbol{\mu}) \end{aligned}$$

is well-defined and continuous;

(vii) D_Σ *is relatively open in* $\mathbb{T} \times \mathbb{T} \times \mathrm{X} \times \mathscr{U}$;

(viii) *the map*

$$\Phi^\Sigma \colon \mathrm{D}_\Sigma \to \mathrm{X}$$

is continuous;

(ix) *for* $(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$ *and for* $\epsilon \in \mathbb{R}_{>0}$, *there exists* $\mathrm{r}, \rho \in \mathbb{R}_{>0}$ *such that*

$$\sup \mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}, \boldsymbol{\mu}) > \sup \mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) - \epsilon, \quad \inf \mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}, \boldsymbol{\mu}) < \inf \mathrm{J}_\Sigma(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) + \epsilon,$$

for all $(\mathbf{x}, \boldsymbol{\mu}) \in \mathsf{B}^{\mathrm{n}}(\mathrm{r}, \mathbf{x}_0) \times \mathsf{B}(\rho, \boldsymbol{\mu}_0)$.

Proof Parts (i)–(iv) follow from Theorem 3.4.8.

(v) Let $(t, \boldsymbol{x}, \boldsymbol{\mu}) \in D_\Sigma(t_0)$. Thus

$$\Phi^\Sigma(t, t_0, \boldsymbol{x}, \boldsymbol{\mu}) \in X.$$

For $\tau \in \mathbb{T}$, define

$$\begin{aligned} \Phi_\tau \colon X \times U &\to X \\ (\boldsymbol{y}, \boldsymbol{u}) &\mapsto \boldsymbol{f}(\tau, \boldsymbol{x}, \boldsymbol{u}) \end{aligned}$$

so that

$$\Phi^\Sigma(\tau + \Delta, \tau, \boldsymbol{x}, \boldsymbol{\mu}) = \Phi_\tau(\boldsymbol{x}, \boldsymbol{\mu}(\tau)).$$

Thus

$$\begin{aligned} \Phi^\Sigma(t_0 + \Delta, t_0, \boldsymbol{x}, \boldsymbol{\mu}) &= \Phi_{t_0}(\boldsymbol{x}, \boldsymbol{\mu}(t_0)), \\ \Phi^\Sigma(t_0 + 2\Delta, t_0, \boldsymbol{x}, \boldsymbol{\mu}) &= \Phi_{t_0+\Delta}(\Phi_{t_0}(\boldsymbol{x}, \boldsymbol{\mu}(t_0)), \boldsymbol{\mu}(t_0 + \Delta)), \\ &\vdots \\ \Phi^\Sigma(t, t_0, \boldsymbol{x}, \boldsymbol{\mu}) &= \Phi_{t-\Delta}(\Phi_{t-2\Delta}(\cdots(\Phi_{t_0+\Delta}(\Phi_{t_0}(\boldsymbol{x}, \boldsymbol{\mu}(t_0)), \boldsymbol{\mu}(t_0 + \Delta)), \cdots), \\ &\qquad \boldsymbol{\mu}(t - 2\Delta)), \boldsymbol{\mu}(t - \Delta)) \end{aligned}$$

This shows that the domain of

$$(\boldsymbol{x}, \boldsymbol{\mu}) \mapsto \Phi^\Sigma(t, t_0, \boldsymbol{x}, \boldsymbol{\mu})$$

is $X \times U^{(t-t_0)/\Delta}$. Continuity of $\boldsymbol{f}$ and openness of X gives a neighbourhood N of

$$(\boldsymbol{x}, (\boldsymbol{\mu}(t_0), \boldsymbol{\mu}(t_0 + \Delta), \ldots, \boldsymbol{\mu}(t - \Delta)))$$

that maps to X. This gives the neighbourhood $\{t\} \times N$ in $D_\Sigma(t_0)$ that maps to X, keeping in mind that the topology on $\mathbb{T}$ is the discrete topology.

(vi) This was proved in the preceding part of the proof.

(vii) The proof here can be carried out as was the proof of part (v).

(viii) This follows from part (vii) in the same manner as part (vi) follows from part (v).

(ix) In this discrete-time case, the assertion will follow if we can show that, for $t, t_0 \in \mathbb{T}$ with $t \geq t_0$, for $\boldsymbol{\mu} \in \mathscr{U}$, and for $\boldsymbol{x} \in D_\Sigma(t, t_0, \boldsymbol{\mu})$, there is a neighbourhood N of $(\boldsymbol{x}, \boldsymbol{\mu})$ in $X \times \mathscr{U}$ such that, if $(\boldsymbol{x}', \boldsymbol{\mu}') \in N$, then $\boldsymbol{x}' \in D_\Sigma(t, t_0, \boldsymbol{\mu}')$. This, however, follows from part (v). ■

6.3.3 Control-affine discrete-time state space systems

We next consider a special class of discrete-time state space systems. The class is worthy of consideration for a few reasons: (1) one can consider for these systems a somewhat broader class of controls, namely those that are locally integrable; (2) this class of systems is a midpoint between general discrete-time state space systems and the linear systems we shall consider in Section 6.8; (3) systems that arise in practice are often of this form.

Here is the definition.

6.3.14 Definition (Control-affine discrete-time state space system) A discrete-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ with $U \subseteq \mathbb{R}^m$ is ***control-affine*** if

(i) there exists $f_0, f_1, \dots, f_m \colon \mathbb{T} \times X \to \mathbb{R}^n$ such that

$$f(t, x, u) = f_0(t, x) + \sum_{a=1}^{m} u_a f_a(t, x),$$

and

(ii) there exists $h_0, h_1, \dots, h_m \colon \mathbb{T} \times X \to \mathbb{R}^k$ such that

$$h(t, x, u) = h_0(t, x) + \sum_{a=1}^{m} u_a h_a(t, x).$$

We call f_0 (resp. h_0) the ***drift dynamics*** (resp. ***drift/output map***) and $f_1, \dots, f_m$ (resp. $h_1, \dots, h_m$) the ***control dynamics*** (resp. ***control/output maps***). •

For a control-affine discrete-time state space system, we shall frequently denote $\mathscr{F} = (f_0, f_1, \dots, f_m)$ and $\mathscr{H} = (h_0, h_1, \dots, h_m)$ and then prescribe such a system by the data $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H})$. Of course, all the notions attached to discrete-time state space systems—e.g., controlled trajectories, controlled outputs, autonomous, proper—can also be attached to those that are control-affine.

Moreover, because there are no distinctions between locally essentially bounded controls and locally integrable controls such as we have for continuous-time state space systems, there are no special cases one needs to consider for control-affine discrete-time state space systems. Thus the existence and uniqueness result Theorem 6.3.9 applies to control-affine discrete-time state space systems, and cannot be improved upon or generalised in any useful way. Similarly, the definition of flow from Definition 6.3.11, and the properties of this flow enunciated in Proposition 6.3.12 and Theorem 6.3.13 hold for control-affine discrete-time state space systems, and cannot be usefully improved or generalised.

Exercises

6.3.1 Show that a discrete-time state space system is not memoryless. (See Example 2.2.31–2 for the definition of a memoryless system.)

6.3.2 Consider the discrete-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ with

(i) $X = \mathbb{R}^2$,
(ii) $U = \mathbb{R}$,
(iii) $\mathbb{T} = \mathbb{Z}_{\geq 0}$,
(iv) $\mathscr{U} = \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}; \mathbb{R})$,
(v) $\boldsymbol{f}(t, (x_1, x_2), u) = (u, 0)$,
(vi) $\boldsymbol{h}(t, (x_1, x_2), u) = x_2$.

Show that Σ is stationary but not strongly stationary.

6.3.3 For the discrete-time state space systems $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ given and for $t_0 \in \mathbb{T}$, indicate whether they are causal from t_0, strongly causal from t_0, finitely observable from any $\tau \in \mathbb{T}_{>t_0}$, stationary from t_0, strongly stationary from t_0, and/or memoryless.

(a) Take
(i) $X = \mathbb{R}$,
(ii) $U = \mathbb{R}$,
(iii) $\mathbb{T} = \mathbb{Z}$,
(iv) $\mathscr{U} = \ell_{\text{loc}}((\mathbb{Z}); \mathbb{R})$,
(v) $f(t, x, u) = x^2 u$,
(vi) $h(t, x, u) = 1$.

(b) Take
(i) $X = \mathbb{R}$,
(ii) $U = \mathbb{R}$,
(iii) $\mathbb{T} = \mathbb{Z}_{>0}$,
(iv) $\mathscr{U} = \ell_{\text{loc}}((\mathbb{Z}_{>0}); \mathbb{R})$,
(v) $f(t, x, u) = t^{-1}x + u^2$,
(vi) $h(t, x, u) = x + u$.

(c) Take
(i) $X = \mathbb{R}^3$,
(ii) $U = \mathbb{R}^2$,
(iii) $\mathbb{T} = \mathbb{Z}$,
(iv) $\mathscr{U} = \ell_{\text{loc}}((\mathbb{Z}); \mathbb{R})$,
(v) $f(t, (x_1, x_2, x_3), (u_1, u_2)) = (x_2, x_3, 0) + (u_1, 0, u_2)$,
(vi) $h(t, (x_1, x_2, x_3), (u_1, u_2)) = (x_1 + x_2, u)$.

(d) Take
(i) $X = \mathbb{R}^2$,
(ii) $U = \mathbb{R}$,
(iii) $\mathbb{T} = \mathbb{Z}$,
(iv) $\mathscr{U} = \ell_{\text{loc}}((\mathbb{Z}); \mathbb{R})$,
(v) $f(t, (x_1, x_2), u) = (2x_1 - x_2, 4x_1 - 3x_2) + (0, u)$,
(vi) $h(t, (x_1, x_2), u) = (2x_1, \frac{1}{2}x_2)$.

The next exercise concerns the connection between discrete-time state space systems and deterministic finite state automata described in Example 2.2.11–2. In making this connection, we will relax the requirement that the state space for a discrete-time state space system be open; this is not much of an alteration since the requirement that the state space be open is made mainly to be consistent with continuous-time state space systems, where openness of the state space is essential.

6.3.4 Answer the following questions.

(a) Show how, given a deterministic finite state automaton $(Q, Y, \Lambda, \delta, \gamma)$, one can associate a discrete-time state space system.

(b) What properties should a discrete-time state space system

$$\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$$

have in order to be a deterministic finite state automaton?

6.3.5 *(Mini-project)* Consider a ball bouncing on a table that undergoes a prescribed vertical motion, as depicted in Figure 6.6 and discussed in [Holmes 1982].

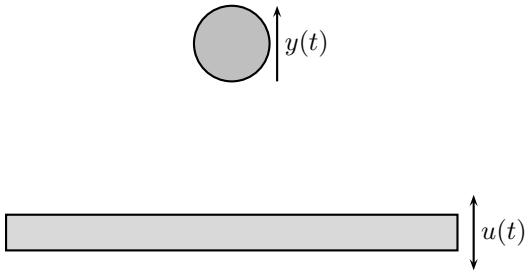


Figure 6.6 Ball bouncing on oscillating table

We let y be the vertical displacement of the ball and let u be the vertical displacement of the table. Make the following assumptions.

1. An impact at time t_0 is modelled by requiring that

$$\dot{y}(t_0+) - u(t_0) = \alpha(u(t_0) - \dot{y}(t_0-)),$$

   where

$$\dot{y}(t_0-) = \lim_{t \uparrow t_0} \dot{y}(t), \quad \dot{y}(t_0+) = \lim_{t \downarrow t_0} \dot{y}(t),$$

   and where $\alpha \in (0, 1]$ is the coefficient of restitution.
2. The mass of the table is so large that the impact of the ball does not affect the motion of the table.
3. The height achieved by the ball after an impact is so large compared to the oscillation of the table that we can assume that, if $\dot{y}(t_j+)$ is the departing velocity after an impact at time t_j, then the arrival velocity for the next impact at time t_{j+1} is $\dot{y}(t_{j+1}-) = -\dot{y}(t_j+)$.

The state are the time τ of impact and the departing velocity v after an impact. The output is the maximum height h attained after an impact.

We wish to assemble all of this into a discrete-time state space system.

(a) What is the state space X for the system?
(b) What is the control set U for the system?
(c) What is the time-domain $\mathbb{T}$ for the system?
(d) What is a good choice for the space $\mathscr{U}$ of inputs?
(e) What are the dynamics $\boldsymbol{f}$?
(f) What is the output map $\boldsymbol{h}$?

Any physical parameters you require, you should introduce yourself. Answer the following questions about the model.

(g) Is the system model causal?
(h) Is the system model stationary?

(i) Is the system model memoryless?
(j) Is the system model control-affine?

Finally, do some system theoretic explorations as follows.

(k) Do some research and describe three system theoretic problems that arise in a natural way for the problem.
(l) Using a computer package for simulating ordinary difference equations, setup the system for simulation, and try some harmonic inputs to see what behaviour you observe.

6.3.6 *(Mini-project)* Consider, as in [Nishimura and Stachurski 2004], a model for a two-sector economy, where one sector produces a good x that is purely consumed and a good y that is purely a capital good. Inputs to the economy are (1) labour ℓ_{con} and ℓ_{cap} to the consumption and capital sectors, respectively and (2) capital c_{con} and c_{cap} to the consumption and capital sectors, respectively. Inputs of capital are made one period prior to production and inputs of labour are made in the same period as production. The measured output is a function u of x that measures the utility of a consumer when she consumes x units of the consumption good. Make the following assumptions.

1. If $C = c_{\text{con}} + c_{\text{cap}}$ is the aggregate capital input, then the capital good y at the end of period k is determined by the following gross accumulation formula:
$$y(k) = C(k) - (1 - \delta)C(k - 1),$$
for a depreciation $\delta \in (0, 1)$.
2. If L is the total labour force, i.e., $L = \ell_{\text{con}} + \ell_{\text{cap}}$, we assume that this is constant.

We wish to assemble all of this into a discrete-time state space system.

(a) What is the state space X for the system?
(b) What is the control set U for the system?
(c) What is the time-domain $\mathbb{T}$ for the system?
(d) What is a good choice for the space $\mathscr{U}$ of inputs?
(e) What are the dynamics $\boldsymbol{f}$?
(f) What is the output map $\boldsymbol{h}$?

Answer the following questions about the model.

(g) Is the system model causal?
(h) Is the system model stationary?
(i) Is the system model memoryless?
(j) Is the system model control-affine?

Finally, do some system theoretic explorations as follows.

(k) Do some research and describe three system theoretic problems that arise in a natural way for the problem.

(l) Using a computer package for simulating ordinary differential equations, setup the system for simulation, and try to maximise the output.

Section 6.4

Discrete-time input/output systems

The next class of systems we consider are input/output systems, now in the setting of discrete-time systems. We shall continue in this section to see a theme in our treatment of system theory, namely that of an input/output system as a continuous mapping between a space of input signals to a space of output signals. We shall also see another these, namely that state space systems can be regarded as input/output systems. Note that this is connected with constructions in general system theory as exemplified by Propositions 2.1.7 and 2.1.13 (for general systems), and Theorem 2.2.20 and Proposition 2.2.49 (for general time systems).

Do I need to read this section? The ideas about input/output systems, and about the connection of such systems to state space systems, that are provided here are a theme in much of our presentation. This theme is enunciated in a somewhat general form for discrete-time systems in this section, and so this section is an important one for what follows. •

6.4.1 Topological constructions for spaces of discrete-time partially defined signals

As we briefly suggested above, input/output systems are maps between spaces of input and output signals. Because of the necessity of allowing signals defined on varying time-domains, cf. Example 2.2.21, this complicates things. Therefore, let us develop some methodology for dealing with this complication.

6.4.1 Definition (Spaces of partially defined signals with topology) Let $\mathbb{T} \subseteq \mathbb{R}$ be a discrete time-domain.

(i) Consider the space

$$\ell^\infty_{\mathrm{loc}}((\mathbb{T}); \mathbb{R}^n) = (\mathbb{R}^n)^{(\mathbb{T})}$$

is a space of partially defined signals with topology when we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the seminorms

$$\|f\|_{\mathbb{K},\infty} = \max\{\sup\{|f_a(t)| \mid t \in \mathbb{K}\} \mid a \in \{1, \dots, n\}\},$$

$\mathbb{K} \subseteq \mathbb{S}$ a bounded sub-time-domain.

(ii) The space

$$\ell^\infty((\mathbb{T}); \mathbb{R}^n) = \{f \in (\mathbb{R}^n)^{(\mathbb{T})} \mid f \in \ell^\infty(\mathrm{dom}(f); \mathbb{R}^n)\}$$

is a space of partially defined signals with topology when we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the norm

$$\|f\|_\infty = \max\{\sup\{|f_a(t)| \mid t \in \mathbb{S}\} \mid a \in \{1, \dots, n\}\}.$$

(iii) For $p \in [1, \infty)$, the space

$$\ell^p_{\mathrm{loc}}((\mathbb{T}); \mathbb{R}^n) = (\mathbb{R}^n)^{(\mathbb{T})}$$

is a space of partially defined signals with topology when we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the seminorms

$$\|\boldsymbol{f}\|_{\mathbb{K},p} = \max\left\{\left(\sum_{t\in\mathbb{K}} |f_a(t)|^p\right)^{1/p} \;\middle|\; a \in \{1, \ldots, n\}\right\},$$

$\mathbb{K} \subseteq \mathbb{S}$ a bounded sub-time-domain.

(iv) For $p \in [1, \infty)$, the space

$$\ell^p((\mathbb{T}); \mathbb{R}^n) = \{\boldsymbol{f} \in (\mathbb{R}^n)^{(\mathbb{T})} \mid \boldsymbol{f} \in \ell^p(\mathrm{dom}(\boldsymbol{f}); \mathbb{R}^n)\}$$

is a space of partially defined signals with topology when we equip $\mathrm{dom}^{-1}(\mathbb{S})$ with the topology defined by the norm

$$\|\boldsymbol{f}\|_p = \max\left\{\left(\sum_{t\in\mathbb{S}} |f_a(t)|^p\right)^{1/p} \;\middle|\; a \in \{1, \ldots, n\}\right\}.$$ •

Note that the preceding sets of partially defined signals are not, themselves, topological spaces. They are merely collections of subsets of signals, each having topologies. There is an important distinction with the discrete-time case, when compared to the continuous-time case. This is that there is no meaningful distinction between the spaces $\ell^p_{\mathrm{loc}}((\mathbb{T}); \mathbb{R}^n)$, $p \in [1, \infty]$. This is because, for these spaces of discrete-time signal spaces, the notions of convergence are identical, and so too, therefore, are the notions of continuity we shall consider for mappings between these spaces. For this reason, when working with spaces of "locally integrable" discrete-time signals, we will always just take $p = \infty$, cf. the discussion in Section IV-1.2.5.

The spaces we shall use are then the following subsets of the preceding spaces.

6.4.2 Definition (Space of partially defined discrete-time signals with topology) We let $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ be a sub-time-domain and let $S \subseteq \mathbb{R}^n$. A ***space of partially defined signals with topology*** is a subset of one of the following spaces of partially defined signals:

(i) the space

$$\ell^\infty_{\mathrm{loc}}((\mathbb{T}); S) = \{\boldsymbol{f} \in \ell^\infty_{\mathrm{loc}}((\mathbb{T}); \mathbb{R}^n) \mid \boldsymbol{f}(t) \in S,\ t \in \mathrm{dom}(\boldsymbol{f})\}$$

with the subspace topology;

(ii) the space

$$\ell^\infty((\mathbb{T}); S) = \{f \in \ell^\infty((\mathbb{T}); \mathbb{R}^n) \mid \ f(t) \in S, \ t \in \operatorname{dom}(f)\}$$

equipped with the subspace topology;

(iii) for $p \in [1, \infty)$, the space

$$\ell^p_{\text{loc}}((\mathbb{T}); S) = \{f \in \ell^p_{\text{loc}}((\mathbb{T}); \mathbb{R}^n) \mid \ f(t) \in S, \ t \in \operatorname{dom}(f)\}$$

equipped with the subspace topology;

(iv) for $p \in [1, \infty)$, the space

$$\ell^p((\mathbb{T}); S) = \{f \in \ell^p((\mathbb{T}); \mathbb{R}^n) \mid \ f(t) \in S, \ t \in \operatorname{dom}(f)\}$$

equipped with the subspace topology.

If $\operatorname{dom}(f) = \mathbb{T}$ for every $f \in \mathscr{S}$, then $\mathscr{S}$ is a ***space of discrete-time signals with topology***. •

Given a space $\mathscr{S}$ of partially defined signals with topology with time-domain $\mathbb{T}$ and given a sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, we shall use the notation

$$\mathscr{S}(\mathbb{S}) = \{f \in \mathscr{S} \mid \ \operatorname{dom}(f) = \mathbb{S}\}.$$

6.4.2 Definitions and system theoretic properties

With suitable notions of spaces of partially defined signals at hand, we can give a suitable definition of an input/output system.

6.4.3 Definition (Discrete-time input/output system) A ***discrete-time input/output system*** is a quintuple $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$, where

(i) $U \subseteq \mathbb{R}^m$ (the ***input set***),

(ii) $\mathbb{T} \subseteq \mathbb{R}$ is an interval (the ***time-domain***),

(iii) $\mathscr{U} \subseteq U^{(\mathbb{T})}$ is a space of partially defined signals with topology (the ***input signals***),

(iv) $\mathscr{Y} \subseteq (\mathbb{R}^k)^{(\mathbb{T})}$ is a space of partially defined signals with topology (the ***output signals***), and

(v) $g \colon \mathscr{U} \to \mathscr{Y}$ has the following properties:

  (a) for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, the restriction of g to $\mathscr{U}(\mathbb{S})$, denoted by $g_{\mathbb{S}}$, takes values in $\mathscr{Y}(\mathbb{S})$;

  (b) if $\mathbb{S}, \mathbb{S}' \subseteq \mathbb{T}$ are sub-time-domains with $\mathbb{S}' \subseteq \mathbb{S}$, then $g_{\mathbb{S}}|\mathscr{U}(\mathbb{S}') = g_{\mathbb{S}'}$;

  (c) $g_{\mathbb{S}}$ is continuous for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$.

Moreover,

(xi) a pair (μ, η) with $\mu \in \mathscr{U}(\mathbb{S})$ and $\eta = g_{\mathbb{S}}(\mu)$ is a ***behaviour*** for Σ, and we denote by $\mathscr{B}(\Sigma)$ the set of behaviours. •

6.4.4 Remark (Restriction in discrete-time input/output systems) Note that we *do not* require that, if $\mathbb{S}, \mathbb{S}' \subseteq \mathbb{T}$ are sub-time-domains with $\mathbb{S}' \subseteq \mathbb{S}$ and if $\boldsymbol{\mu} \in \mathscr{U}(\mathbb{S})$, then $\boldsymbol{\mu}|\mathbb{S}' \in \mathscr{U}(\mathbb{S}')$. What we *do* require is that, if $\boldsymbol{\mu}|\mathbb{S}' \in \mathscr{U}(\mathbb{S}')$, then

$$g_{\mathbb{S}'}(\boldsymbol{\mu}|\mathbb{S}') = g_{\mathbb{S}}(\boldsymbol{\mu})|\mathbb{S}'.$$

If a discrete-time input/output system does have then property that $\boldsymbol{\mu}|\mathbb{S}' \in \mathscr{U}(\mathbb{S}')$ for every pair of sub-time-domains satisfying $\mathbb{S}' \subseteq \mathbb{S}$ and for every $\boldsymbol{\mu} \in \mathscr{U}(\mathbb{S})$, we shall say that the system is ***closed under restriction***.

Note that, by not requiring that discrete-time input/output systems be closed under restriction, we allow the common situation where all inputs and outputs are considered only as signals defined on the entire time-domain. That is to say, for a system like that, we have $\mathscr{U}(\mathbb{S}) = \varnothing$ and $\mathscr{Y}(\mathbb{S}) = \varnothing$ for every strict sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$. •

Let us connect some of the general systems ideas from Chapter 2 to our concept of a discrete-time inout/output system. Along the way, we shall give a few elementary examples of such systems. In Section 6.2.3 we shall see that all discrete-time state space systems are also discrete-time input.output systems.

We begin by making the connection to the basic types of general systems.

6.4.5 Remarks (Discrete-time input/output systems as general systems) We let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a discrete-time input/output system.

1. A discrete-time input/output system is a general input/output system as per Definition 2.1.3. To see this, take
   (a) "$\mathscr{U} = \mathscr{U}$," i.e., the inputs for the general input/output system are the same as the inputs for the discrete-time state space system,
   (b) "$\mathscr{Y} = \mathscr{Y}$," i.e., the outputs for the general input/output system are the same as the outputs for the discrete-time state space system, and
   (c) $\mathscr{B} = \{(\boldsymbol{\mu}, g(\boldsymbol{\mu})) \mid \boldsymbol{\mu} \in \mathscr{U}\}$, i.e., a discrete-time input/output system is a functional input/output system, as per Definition 2.1.4.
2. A discrete-time input/output system is, more specifically, a general time system as per Definition 2.2.9. To see this, take
   (a) "$U = U$," i.e., the input set for the general time system is the same as the input set for the discrete-time input/output system,
   (b) $Y = \mathbb{R}^k$, i.e., the output set for the general time system is $\mathbb{R}^k$,
   (c) "$\mathscr{U} = \mathscr{U}$," i.e., the admissible input signals for the general input/output system are the same as the input signals for the discrete-time input/output system,
   (d) $\mathscr{Y} = (\mathbb{R}^k)^{(\mathbb{T})}$, i.e., the admissible output signals for the general input/output system are the partial $\mathbb{R}^k$-valued functions on $\mathbb{T}$, and
   (e) $\mathscr{B} = \{(\boldsymbol{\mu}, g(\boldsymbol{\mu})) \mid \boldsymbol{\mu} \in \mathscr{U}\}$ i.e., the behaviours for the general time system input/output pairs for the discrete-time input/output system. •

Let us now consider the matter of output completeness and completeness for discrete-time input/output systems.

6.4.6 Remarks (Completeness for discrete-time input/output systems) We let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a discrete-time input/output system.

1. *Discrete-time input/output systems are output complete:* Let $\boldsymbol{\mu} \in \mathscr{U}$ and let $(I, \preceq)$ be a totally ordered set, and let $(\boldsymbol{\eta}_i)_{i \in I}$ be a family of outputs satisfying conditions (a)–(f) of Definition 2.2.12. Note that

$$\boldsymbol{\eta}_i(t) = g(\boldsymbol{\mu})(t), \qquad t \in \operatorname{dom}(\boldsymbol{\eta}_i).$$

   Now let $\mathbb{S} = \cup_{i \in I} \operatorname{dom}(\boldsymbol{\eta}_i)$ and let $\boldsymbol{\eta} \colon \mathbb{S} \to \mathbb{R}^k$ be such that $\boldsymbol{\eta}_{\operatorname{dom}(\boldsymbol{\eta}_i)} = \boldsymbol{\eta}_i$, $i \in I$. Then, if $t \in \mathbb{S}$, we must have $t \in \operatorname{dom}(\boldsymbol{\eta}_i)$ for some $i \in I$. Therefore,

$$\boldsymbol{\eta}(t) = \boldsymbol{\eta}_i(t) = g(\boldsymbol{\mu})(t).$$

   As this holds for every $t \in \operatorname{dom}(\boldsymbol{\eta})$, we conclude output completeness.

2. *Generally, a discrete-time input/output system is not complete:* In Theorem 6.4.10 we shall see that discrete-time state space systems are discrete-time input/output systems. Thus any discrete-time state space system that is not complete will furnish us with a discrete-time input/output system that is not complete. •

We know from general results, i.e., Theorem 2.2.20, a complete discrete-time input/output system has a dynamical systems representation specified by some response family and some family of state transition maps. Moreover, the proof of Theorem 2.2.20 gives an explicit construction of such a dynamical systems representation. The difficulty is that, in any given example, the resulting dynamical systems representation will not be meaningful (whatever might be the meaning of "meaningful"). Indeed, the matter of constructing a meaningful dynamical systems representation is something that, typically, one should think carefully about.

Now let us consider the various attributes for general time systems from Section 2.2, as they pertain to discrete-time input/output systems. We shall see that these notions do not hold, generally, and so are assumptions that must be made if one needs them. In order to connect the general time system discussion of Section 2.2 to the systems we consider here, let us make suitable definitions for the appropriate notions.

First we consider causality, where the definition captures the idea that the output at time t depends only on the input prior to time t.

6.4.7 Definition (Causality for discrete-time input/output systems) Let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a discrete-time input/output system.

(i) The system Σ is ***causal*** if, for every $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathscr{U}$ with $\operatorname{dom}(\boldsymbol{\mu}_1) = \operatorname{dom}(\boldsymbol{\mu}_2)$ and for every $t \in \operatorname{dom}(\boldsymbol{\mu}_1) = \operatorname{dom}(\boldsymbol{\mu}_2)$,

$$\boldsymbol{\mu}_1|(\mathbb{T}_{\preceq t} \cap \operatorname{dom}(\boldsymbol{\mu}_1)) = \boldsymbol{\mu}_1|(\mathbb{T}_{\preceq t} \cap \operatorname{dom}(\boldsymbol{\mu}_2)) \implies g(\boldsymbol{\mu}_1)(t) = g(\boldsymbol{\mu}_2)(t).$$

(ii) The system Σ is ***strongly causal*** if, for every $\mu_1, \mu_2 \in \mathscr{U}$ with $\operatorname{dom}(\mu_1) = \operatorname{dom}(\mu_2)$ and for every $t \in \operatorname{dom}(\mu_1) = \operatorname{dom}(\mu_2)$,

$$\mu_1|(\mathbb{T}_{<t} \cap \operatorname{dom}(\mu_1)) = \mu_1|(\mathbb{T}_{<t} \cap \operatorname{dom}(\mu_2)) \implies g(\mu_1)(t) = g(\mu_2)(t). \qquad \bullet$$

Next we consider stationarity. We let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a discrete-time input/output system. As we saw in Section 2.2.8, stationarity has to do, roughly, with shift-invariance. To make this clear, let us first carefully think about what we mean by shifting. Let X be a set and let $\mathscr{X} \subseteq X^{(\mathbb{T})}$ be a collection of partially defined signals. Let $a \in \mathbb{Z}(\Delta)$. If $\xi \in \mathscr{X}$, denote by $\tau_a^* \xi$ the signal with domain

$$\operatorname{dom}(\tau_a^* \xi) = \{t \in \mathbb{T} \mid t - a \in \operatorname{dom}(\xi)\}$$

and given by $\tau_a^* \xi(t) = \xi(t - a)$. Note that we may well have $\operatorname{dom}(\tau_a^* \xi) = \varnothing$, in which case $\tau_a^* \xi$ is not defined, by convention.

With this notation, we have the following definitions regarding stationarity.

6.4.8 Definition (Stationarity for discrete-time input/output systems) Let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a discrete-time input/output system with $\sup \mathbb{T} = \infty$.

(i) The system Σ is ***stationary*** if $\tau_a^*(\mathscr{U}) \subseteq \mathscr{U}$ for every $a \in \mathbb{Z}_{>0}(\Delta)$ and if, for every $\mu \in \mathscr{U}$,

$$g(\tau_a^* \mu) = \tau_a^* g(\mu).$$

(ii) The system Σ is ***strongly stationary*** if it is stationary and if, for every $a \in \mathbb{Z}_{>0}(\Delta)$ and every $\mu \in \mathscr{U}$, there exists $\mu' \in \mathscr{U}$ such that

$$g(\mu) = g(\tau_a^* \mu'). \qquad \bullet$$

With these definitions, we can make the following remarks.

6.4.9 Remarks (System theoretic attributes of discrete-time input/output systems) We let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a discrete-time input/output system.

1. Σ *is generally not causal:* To see this, we give a simple counterexample.
   We take $U = \mathbb{R}$, $\mathbb{T} = \mathbb{Z}$, and let $\mathscr{U} = \ell^\infty_{\mathrm{loc}}(\mathbb{Z}; \mathbb{R})$, i.e., inputs are all $\mathbb{R}$-valued functions on $\mathbb{Z}$. We also take $\mathscr{Y} = \ell^\infty_{\mathrm{loc}}(\mathbb{Z}; \mathbb{R})$. The topologies for $\mathscr{U}$ and $\mathscr{Y}$ are as defined in Definition 6.4.2–(i). Now define $g \colon \mathscr{U} \to \mathscr{Y}$ by $g(\mu)(t) = \mu(-t)$. Because we are only considering signals defined on all of $\mathbb{Z}$, conditions (v)(a) and (v)(b) of a discrete-time input/output system are immediately satisfied. We claim that condition (v)(c) is also satisfied. Indeed, let $(\mu_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $\ell^\infty_{\mathrm{loc}}(\mathbb{Z}; \mathbb{R})$ converging to $\mu \in \ell^\infty_{\mathrm{loc}}(\mathbb{R}; \mathbb{R})$. Let $\mathbb{K} \subseteq \mathbb{R}$ be a bounded sub-time-domain and let $\epsilon \in \mathbb{R}_{>0}$. Let

   $$-\mathbb{K} = \{-t \mid t \in \mathbb{K}\}.$$

   Then there exists $N \in \mathbb{Z}_{>0}$ such that

   $$|\mu(t) - \mu_j(t)| < \epsilon, \qquad t \in -\mathbb{K},\ j \geq N.$$

Then we immediately have

$$|g(\mu)(t) - g(\mu_j)(t)| < \epsilon, \qquad t \in \mathbb{K},\ j \geq N,$$

giving convergence of $(g(\mu_j))_{j\in\mathbb{Z}_{>0}}$ to $g(\mu)$, and so giving continuity of g.
Now we show that the system is not causal. Let $\mu_1, \mu_2 \in \mathscr{U}$ be defined by

$$\mu_1(t) = \begin{cases} 1, & t \in \mathbb{Z}_{<0}, \\ 0, & t \in \mathbb{Z}_{\geq 0}, \end{cases} \qquad \mu_2(t) = 1, \qquad t \in \mathbb{Z}.$$

Let $t \in \mathbb{Z}_{<0}$ and note that

$$\mu_1|\mathbb{R}_{\leq t} = \mu_2|\mathbb{R}_{\leq t}.$$

However,

$$g(\mu_1)(t) = \mu_1(-t) = 0, \qquad g(\mu_2)(t) = \mu_2(-t) = 1,$$

and this demonstrates the lack of causality.

2. Σ *is generally not past determined:* This follows since, as proved in Proposition 2.2.35, past determined systems are causal.
3. Σ *is finitely observable:* This is a consequence of the fact that Σ, as a general input/output system, is functional.
4. Σ *is not generally stationary:* To see this, we note that discrete-time state space systems are discrete-time input-output systems by Theorem 6.4.10. Therefore, since discrete-time state space systems are not generally stationary (as we pointed out in Remark 6.1.8–4).
5. Σ *is not generally linear:* Presumably, since in Section 6.9 we shall specifically consider linear discrete-time input/output systems, it is not the case that all discrete-time input/output systems are linear. To see this, one need only produce a counterexample, and we leave the elementary construction of such an example to the reader, cf. Exercise 6.3.3. •

6.4.3 Discrete-time state space systems as discrete-time input/output systems

As we saw in our discussion above of the system theoretic attributes for discrete-time input/output systems, these systems were capable of exhibiting characteristics that are not possible for discrete-time state space systems. In this section we show how the various classes of discrete-time state space systems are also discrete-time input/output systems.

First let us informally associate to a discrete-time state space system its candidate input/output system. Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a discrete-time state space system. If one thinks about the controlled outputs for Σ, one sees that these behaviours do not form the basis for a discrete-time input/output system since there are multiple outputs for a single input. To rectify this, one should choose an initial condition.

Thus let $(t_0, x_0) \in \mathbb{T} \times X$. Then we can try to associate a discrete-time input/output system Σ for this initial condition data by the quintuple $\Sigma_{\mathrm{i/o}}(t_0, x_0) = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$, where

1. "$U = U$,"
2. "$\mathbb{T} = \mathbb{T}$,"
3. "$\mathscr{U} = \mathscr{U}$,"
4. $\mathscr{Y} \subseteq (\mathbb{R}^k)^{(\mathbb{T})}$, and
5. $g(\mu)(t) = h(t, \Phi^\Sigma(t, t_0, x_0, \mu), \mu(t))$ for $t \in J_\Sigma(t_0, x_0, \mu)$.

This does not quite yet define a discrete-time input/output system since we must prescribe the structure of a space of partially defined signals with topology to both $\mathscr{U}$ and $\mathscr{Y}$. As we shall see, the appropriate such structure depends on the character of the system.

The following result characterises how one can make the preceding association precise. Note that, in contrast with Theorem 6.2.10, we do not need to have many separate cases, since the topologies $\ell^p_{\mathrm{loc}}(\mathbb{T}; \mathsf{V})$ agree for $p \in [1, \infty]$.

6.4.10 Theorem (Discrete-time input/output systems from discrete-time state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a discrete-time state space system. Assume that $\mathbf{f}$ is continuous and that $\mathbf{h}$ is output autonomous and a continuous mapping from $\mathrm{X} \times \mathrm{U}$ to $\mathbb{R}^k$. Let $(\mathrm{t}_0, \mathbf{x}_0) \in \mathbb{T} \times \mathrm{X}$. Then $\Sigma_{\mathrm{i/o}}(\mathrm{t}_0, \mathbf{x}_0) = (\mathrm{U}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathbf{g})$, with $\mathbf{g}$ as defined above, defines a discrete-time input/output system, where*

(i) $\mathscr{U} \subseteq \ell^\infty_{\mathrm{loc}}((\mathbb{T}); \mathrm{U})$ is the space of partially defined signals with topology as in Definition 6.4.2–(i) and

(ii) $\mathscr{Y} = \ell^\infty_{\mathrm{loc}}((\mathbb{T}); \mathbb{R}^k)$ is the space of partially defined signals with topology as in Definition 6.4.2–(i).

Proof The following lemma records an essential part of the proof.

1 Lemma *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a discrete-time state space system for which $\mathbf{f}$ is continuous. Let $(\mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) \in \mathbb{T} \times \mathrm{X} \times \mathscr{U}$, $\mathrm{t} \in \mathbb{T}_{\geq \mathrm{t}_0}$ satisfy $(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0) \in \mathrm{D}_\Sigma$, let $(\boldsymbol{\mu}_\mathrm{j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$ with respect to the seminorm $\|\cdot\|_{[\mathrm{t}_0, \mathrm{t}], \infty}$. Then the following statements hold:*

(i) there exists $\mathrm{N} \in \mathbb{Z}_{>0}$ such that $(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_\mathrm{j}) \in \mathrm{D}_\Sigma$ for $\mathrm{j} \geq \mathrm{N}$;

(ii) the sequence

$$\mathrm{s} \mapsto \Phi^\Sigma(\mathrm{s}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_\mathrm{j}), \qquad \mathrm{j} \in \mathbb{Z}_{>0},$$

of mappings in $\ell^\infty([\mathrm{t}_0, \mathrm{t}]; \mathrm{X})$ converges to

$$\mathrm{s} \mapsto \Phi^\Sigma(\mathrm{s}, \mathrm{t}_0, \mathbf{x}_0, \boldsymbol{\mu}_0).$$

Proof The first assertion is a direct consequence of part (ix) of Theorem 6.3.13. We must, therefore, prove the convergence conclusion of the second assertion.

As we saw in the proof of Theorem 6.3.13(v), $\mathscr{U}_{[t_0,t]} \simeq U^{(t-t_0)/\Delta}$. Moreover, convergence in $\ell^\infty([t_0, t]; U)$ is the same as normal convergence in $U^{(t-t_0)/\Delta}$, thinking of the

latter as a subset of $(\mathbb{R}^m)^{(t-t_0)/\Delta}$. Also as we saw in the proof of Theorem 6.3.13(v), the mapping

$$(\boldsymbol{\mu}(t_0), \boldsymbol{\mu}(t_0 + \Delta), \dots, \boldsymbol{\mu}(t - \Delta)) \mapsto \Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu})$$

is continuous. Therefore, by this continuity,

$$\lim_{j\to\infty} \Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) = \Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0).$$

The lemma follows since convergence in $\ell^\infty_{\mathrm{loc}}([t_0, t]; X)$ is the same as normal convergence in $X^{(t-t_0)/\Delta}$, thinking of this as a subset of $(\mathbb{R}^n)^{(t-t_0)/\Delta}$. ▼

Let $(\boldsymbol{\mu}_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathscr{U}$ converging to $\boldsymbol{\mu}_0$. Thus $\mathrm{dom}(\boldsymbol{\mu}_j) = \mathbb{S}$, $j \in \mathbb{Z}_{\geq 0}$, and that, for every finite sub-time-domain $\mathbb{K} \subseteq \mathbb{S}$, we have

$$\lim_{j\to\infty} \mu_{j,a}(t) = \mu_{0,a}(t), \qquad a \in \{1, \dots, m\},\ t \in \mathbb{K},$$

keeping in mind that convergence in $\ell^\infty(\mathbb{K}; U)$ is normal convergence, as in the proof of the lemma.

Let $\mathbb{L} \subseteq \mathbb{S}$ be bounded and let $t_1 \in \mathbb{S}$ be such that $\mathbb{K} \subseteq [t_0, t_1]$. Then we have, by the lemma,

$$\lim_{j\to\infty} \Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j) = \Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \qquad t \in [t_0, t_1].$$

By continuity of $\boldsymbol{h}$, this gives

$$\lim_{j\to\infty} \boldsymbol{h}(\Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_j), \boldsymbol{\mu}_j) = \boldsymbol{h}(\Phi^{\Sigma}(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0), \boldsymbol{\mu}_0), \qquad t \in [t_0, t_1].$$

Since convergence in $\ell^\infty([t_0, t_1]; \mathbb{R}^k)$ is normal convergence in $(\mathbb{R}^k)^{(t_t-t_0)/\Delta}$, this gives the desired conclusion. ■

We point out that we have focussed in the preceding result on the case when $\boldsymbol{h}$ is independent of time. Just like in the continuous-time case, if $\boldsymbol{h}$ does depend on time, then it is the nature of this time-dependence that will determine the manner of the continuity properties of the input/output map. As such, these cases have to be treated in a more case-by-case manner. The results we give are interesting and useful general results.

6.4.4 Discrete-time difference input/output systems

Exercises

6.4.1 For the discrete-time input/output systems $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \boldsymbol{g})$ given, indicate whether they are causal, strongly causal, finitely observable from any $\tau \in \mathbb{T}_{>t_0}$, stationary, strongly stationary, and/or memoryless.

(a) Take

(i) $U = \mathbb{R}$,

(ii) $\mathbb{T} = \mathbb{Z}$,

(iii) $\mathscr{U} = \ell^1(\mathbb{Z}; \mathbb{R})$,

(iv) $\mathscr{Y} = \{0, 1\}^{\mathbb{Z}} \cap \ell^\infty(\mathbb{Z}; \mathbb{R})$,

(v) $\boldsymbol{g}(\mu)(t) = \begin{cases} 1, & \sum_{\tau=-\infty}^{t} \mu(\tau) \geq 1, \\ 0, & \text{otherwise.} \end{cases}$

(b) Take
(i) $U = \mathbb{R}$,
(ii) $\mathbb{T} = \mathbb{Z}$,
(iii) $\mathscr{U} = \ell_{\text{loc}}(\mathbb{Z}; \mathbb{R})$,
(iv) $\mathscr{Y} = \ell_{\text{loc}}(\mathbb{Z}; \mathbb{R})$,
(v) $g(\mu)(t) = \mu(kt)$ for some $k \in \mathbb{Z}$.

(c) Take
(i) $U = \mathbb{R}$,
(ii) $\mathbb{T} = \mathbb{Z}$,
(iii) $\mathscr{U} = \ell_{\text{loc}}(\mathbb{Z}; \mathbb{R})$,
(iv) $\mathscr{Y} = \{0, 1\}^{\mathbb{Z}} \cap \ell^\infty(\mathbb{Z}; \mathbb{R})$,
(v) $g(\mu)(t) = \begin{cases} 1, & \mu(t) \in \mathbb{Q}, \\ 0, & \text{otherwise.} \end{cases}$

(d) Take
(i) $U = \mathbb{R}$,
(ii) $\mathbb{T} = \mathbb{Z}$,
(iii) $\mathscr{U} = \ell^1(\mathbb{R}; \mathbb{R})$,
(iv) $\mathscr{Y} = \ell^\infty(\mathbb{R}; \mathbb{R})$,
(v) $g(\mu)(t) = \sum_{\tau=-\infty}^{t} \mathrm{e}^{-|\tau|} \mu(\tau)$.

6.4.2 Let $U, Y \subseteq \mathbb{R}$ be open sets, let $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ be a discrete time-domain, let $n \in \mathbb{Z}_{>0}$, and let

$$F \colon \mathbb{T} \times Y \times \mathrm{L}^{\leq n}_{\text{sym}}(\mathbb{R}; \mathbb{R}) \times U \times \mathrm{L}^{\leq n-1}_{\text{sym}}(\mathbb{R}; \mathbb{R}) \to \mathbb{R}.$$

(See the beginning of Section 3.3.3 for notation.) Suppose that, for each

$$(t, y, y^{(+,1)}, \dots, y^{(+,n-1)}, u, u^{(+,1)}, \dots, u^{(+,n-1)}) \in \mathbb{T} \times Y \times \mathrm{L}^{\leq n-1}_{\text{sym}}(\mathbb{R}; \mathbb{R}) \times U \times \mathrm{L}^{\leq n-1}_{\text{sym}}(\mathbb{R}; \mathbb{R}),$$

we can solve the equation

$$F(t, y, y^{(+,1)}, \dots, y^{(+,n-1)}, y^{(+,n)}, u, u^{(+,1)}, \dots, u^{(+,n-1)}) = 0$$

uniquely for $y^{(+,n)}$ and denote the unique solution by

$$y^{(+,n)} = \widehat{F}(t, y, y^{(+,1)}, \dots, y^{(+,n-1)}, u, u^{(+,1)}, \dots, u^{(+,n-1)}).$$

Suppose that $\widehat{F}$ is continuous and consider the difference equation

$$\eta(t + n\Delta) = \widehat{F}(t, \eta(t), \eta(t + \Delta), \dots, \eta(t + (n-1)\Delta), \mu(t), \mu(t + \Delta), \dots, \mu(t + (n-1)\Delta)).$$

Answer the following questions.

(a) Show that this determines a general time system as per Definition 2.2.9. Clearly identify the spaces of input and output signals.

(b) Argue that a natural choice of states for this system is

$$\xi_j(t) = \eta(t + j\Delta), \qquad j \in \{0, 1, \dots, n-1\}.$$

(c) Derive a discrete-time state space system for which the input/output relation is the same as the general time system from part (a) and for which the states are as in part (b).

6.4.3 For a finite collection of real data, $\boldsymbol{X} = \{x_1, \dots, x_n\}$, the ***mean*** is

$$\operatorname{mean}(\boldsymbol{X}) = \frac{1}{n} \sum_{j=1}^{n} x_j$$

and the ***standard deviation*** is

$$\operatorname{stddev}(\boldsymbol{X}) = \left(\frac{1}{n} \sum_{j=1}^{n} (x_j - \operatorname{mean}(\boldsymbol{X}))^2 \right)^{1/2}.$$

Suppose that, given $\mu \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}; \mathbb{R})$, we define functions

$$\operatorname{mean}(\mu), \operatorname{stddev}(\mu) \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}; \mathbb{R})$$

by

$$\operatorname{mean}(\mu)(n) = \operatorname{mean}(\{\mu(0), \dots, \mu(n)\}), \quad \operatorname{stddev}(\mu)(n) = \operatorname{stddev}(\{\mu(0), \dots, \mu(n)\}).$$

Answer the following questions.

(a) Make this record of "running mean" and "running standard deviation" into a discrete-time input/output system $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$.

(b) Is the system causal? strongly causal? stationary? strongly stationary? memoryless?

Exercises 6.9.8–6.9.11 consider a few linear models for time series analysis, and we refer the reader to the brief discussion preceding Exercise 6.9.8 for some background.

6.4.4 We consider a nonlinear model for time series analysis based on the difference equation

$$\begin{aligned} \eta(k\Delta) = b_1 \eta((k-1)\Delta) + \cdots + b_{n-1} \eta((k-(n-1))\Delta) + b_n \eta((k-n)\Delta) \\ + a_1 \iota((k-1)\Delta)^2 + \cdots + a_m \iota((k-m)\Delta)^2, \qquad k \in \mathbb{Z}_{\geq n}, \end{aligned}$$

for signals $\eta, \iota \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}(\Delta); \mathbb{R})$ and for $a_1, \dots, a_m \in \mathbb{R}$. This is the ***generalised autoregressive conditional heteroscedasticity*** model of order (n, m), denoted GARCH(n, m). Note, for example, that GARCH$(0, 0)$ is simply a white noise process. The coefficients $b_1, \dots, b_n, a_1, \dots, a_m$ are chosen to fit measured data by matching statistical properties.

You will examine some features of GARCH$(1, 1)$, which is determined by the equation

$$\eta(k\Delta) = b\eta((k-1)\Delta) + a\iota((k-1)\Delta)^2, \qquad k \in \mathbb{Z}_{>0}.$$

For this system, answer the following questions.

(a) Show that the solution to the system of difference equations with innovation ι specified and with initial condition $\eta(0) = y_0$ is

$$\eta(k\Delta) = b^k y_0 + a \sum_{j=0}^{k-1} b^j \iota((k-j-1)\Delta)^2, \qquad k \in \mathbb{Z}_{>0}.$$

(b) Show that, when an initial condition $\eta(0) = y_0$ is specified, the previous equation describes a discrete-time input/output system with input ι and output η.

(c) Show that GARCH(1, 1) can be written as an ARMA(1, 1) process with ι^2 playing the rôle of output and $\iota^2 - \eta$ playing the rôle of the innovations.

Section 6.5

Linearisation of systems

Having presented a class of not necessarily linear systems, both in continuous- and discrete-time, we shall now transition to linear systems. To justify the relevance of this, we start by continuous- and discrete-time linearising systems, just as we did for ordinary differential and ordinary difference equations in Section 5.1. We shall proceed much as we did for linearisation in Section 5.1, considering linearisation about trajectories (controlled trajectories, in this case) and then equilibria (controlled equilibria in this case). We also characterise the linearisation in terms of variations of initial conditions and controls, thus giving some meaning to linearisation as the derivative with respect to initial condition and control.

Do I need to read this section? This section is intended as a bridge from the general constructions of Sections 6.1–6.4. As such, the results give context to the importance of the linear systems to which we shall dedicate most of our attention. •

6.5.1 Linearisation of continuous-time state space systems

We consider first the linearisation of continuous-time state space systems. The constructions here mirror closely the results of Section 5.1.1 for linearisation of ordinary differential equations.

6.5.1.1 Linearisation along controlled trajectories Suppose that we have a continuous-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ and that we have a controlled trajectory (ξ_0, μ_0) for Σ defined on $\mathbb{T}' \subseteq \mathbb{T}$. We wish to understand what happens to controlled trajectories "nearby" this fixed controlled trajectory.

To do this, we suppose that $U \subseteq \mathbb{R}^m$ is open and that f and h are continuously differentiable as functions of (x, u). That is, for $t \in \mathbb{T}$ we denote

$$\begin{aligned} f_t &\colon X \times U \to \mathbb{R}^n \\ &\quad (x, u) \mapsto f(t, x, u), \\ h_t &\colon X \times U \to \mathbb{R}^k \\ &\quad (x, u) \mapsto h(t, x, u), \end{aligned}$$

and we require that f_t and h_t be of class C^1 for each $t \in \mathbb{T}$. We denote

$$\begin{aligned} D_1 f(t, x, u) &= D_1 f_t(x, u), \\ D_2 f(t, x, u) &= D_2 f_t(x, u), \\ D_1 h(t, x, u) &= D_1 h_t(x, u), \\ D_2 h(t, x, u) &= D_2 h_t(x, u), \qquad t \in \mathbb{T}. \end{aligned}$$

the partial derivatives with respect to $\boldsymbol{x}$ and $\boldsymbol{u}$, respectively, with t fixed. Thus

$$\begin{aligned}\boldsymbol{D}_1\boldsymbol{f}&\colon \mathbb{T}\times X\times U\to \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n),\\ \boldsymbol{D}_2\boldsymbol{f}&\colon \mathbb{T}\times X\times U\to \mathrm{L}(\mathbb{R}^m;\mathbb{R}^n),\\ \boldsymbol{D}_1\boldsymbol{h}&\colon \mathbb{T}\times X\times U\to \mathrm{L}(\mathbb{R}^n;\mathbb{R}^k),\\ \boldsymbol{D}_2\boldsymbol{h}&\colon \mathbb{T}\times X\times U\to \mathrm{L}(\mathbb{R}^m;\mathbb{R}^k).\end{aligned}$$

We then suppose that we have a controlled trajectory (ξ_0, μ_0), defined on $\mathbb{T}'$, for Σ for which the deviations $\nu \triangleq \xi - \xi_0$ and $\omega = \mu - \mu_0$ are small. Let us try to understand the behaviour of ν. Naïvely, we can do this as follows:

$$\begin{aligned}\dot\xi(t) &= \frac{\mathrm{d}(\xi_0+\nu)}{\mathrm{d}t}(t) = \boldsymbol{f}(t,\xi_0(t)+\nu(t),\mu_0+\omega(t))\\ &= \boldsymbol{f}(t,\xi_0(t),\mu_0(t)) + \boldsymbol{D}_1\boldsymbol{f}(t,\xi_0(t),\mu_0(t))\cdot\nu(t) + \boldsymbol{D}_2\boldsymbol{f}(t,\xi_0(t),\mu_0(t))\cdot\omega(t)+\cdots.\end{aligned}$$

We will not here try to be precise about what "$\cdots$" might mean, but merely say that the idea of the preceding equation is that we approximate using the constant and first-order terms in the Taylor expansion, and then pray that this gives us something meaningful. Note that, since (ξ_0, μ_0) is a controlled trajectory for Σ, the approximation we arrive at is

$$\dot\nu(t) \approx \boldsymbol{D}_1\boldsymbol{f}(t,\xi_0(t),\mu_0(t))\cdot\nu(t) + \boldsymbol{D}_2\boldsymbol{f}(t,\xi_0(t),\mu_0(t))\cdot\omega(t).$$

We similarly denote

$$\eta_0(t) = \boldsymbol{h}(t,\xi_0(t),\mu_0(t)),\quad \eta(t) = \boldsymbol{h}(t,\xi(t),\mu(t)),$$

and deduce, with $\gamma(t) = \eta(t) - \eta_0(t)$, that we have an approximation

$$\gamma(t) \approx \boldsymbol{D}_1\boldsymbol{h}(t,\xi_0(t),\mu_0(t))\cdot\nu(t) + \boldsymbol{D}_2\boldsymbol{h}(t,\xi_0(t),\mu_0(t))\cdot\omega(t).$$

Meaningful or not, the preceding naïve calculations give rise to the following definition.

6.5.1 Definition (Linearisation of a continuous-time state space system along a controlled trajectory) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \boldsymbol{f}, \boldsymbol{h})$ be a continuous-time state space system, supposing that $U \subseteq \mathbb{R}^m$ is open and that $\boldsymbol{f}_t$ and $\boldsymbol{h}_t$ are of class C^1 for every $t \in \mathbb{T}$. For $(\xi_0, \mu_0) \in \mathrm{Ctraj}(\Sigma)$ with domain $\mathbb{T}'$, the ***linearisation of Σ along (ξ_0, μ_0)*** is the continuous-time state space system

$$\Sigma_{L,(\xi_0,\mu_0)} = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{T}', \mathscr{U}_L, \boldsymbol{f}_{L,(\xi_0,\mu_0)}, \boldsymbol{h}_{L,(\xi_0,\mu_0)}),$$

where

(i) $\boldsymbol{f}_{L,(\xi_0,\mu_0)}(t,v,w) = \boldsymbol{D}_1\boldsymbol{f}(t,\xi_0(t),\mu_0(t))\cdot v + \boldsymbol{D}_2\boldsymbol{f}(t,\xi_0(t),\mu_0(t))\cdot w$,

(ii) $\boldsymbol{h}_{L,(\xi_0,\mu_0)}(t,v,w) = \boldsymbol{D}_1\boldsymbol{h}(t,\xi_0(t),\mu_0(t))\cdot v + \boldsymbol{D}_2\boldsymbol{h}(t,\xi_0(t),\mu_0(t))\cdot w$, and

(iii) $\mathscr{U}_L \subseteq \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}'; \mathbb{R}^m)$. •

Note that a controlled trajectory for the linearisation of Σ along (ξ_0, μ_0) satisfies

$$\dot{\nu}(t) = A(t)(\nu(t)) + B(t)(\omega(t)),$$

where

$$A(t) = D_1 f(t, \xi_0(t), \mu_0(t)), \quad B(t) = D_2 f(t, \xi_0(t), \mu_0(t)).$$

The corresponding controlled outputs satisfy

$$\gamma(t) = C(t)(\nu(t)) + D(t)(\omega(t)),$$

where

$$C(t) = D_1 h(t, \xi_0(t), \mu_0(t)), \quad D(t) = D_2 h(t, \xi_0(t), \mu_0(t)).$$

This is what we shall subsequently refer to as a linear continuous-time state space system.

Note that there is an alternative view of linearisation that can be easily developed, one where linearisation is of the *system*, not just along a controlled trajectory. The construction we make is the following.

6.5.2 Definition (Linearisation of a continuous-time state space system) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a continuous-time state space system, supposing that $U \subseteq \mathbb{R}^m$ is open and that f_t and h_t are of class C^1 for every $t \in \mathbb{T}$. The ***linearisation*** of Σ is the continuous-time state space system

$$\Sigma_L = (X \times \mathbb{R}^n, U \times \mathbb{R}^m, \mathbb{T}, \mathscr{U}_L, f_L, h_L),$$

where

(i) $f_L(t, (x, v), (u, w)) = (f(t, x, u), D_1 f(t, x, u) \cdot v + D_2 f(t, x, u) \cdot w)$,

(ii) $h_L(t, (x, v), (u, w)) = (h(t, x, u), D_1 h(t, x, u) \cdot v + D_2 h(t, x, u) \cdot w)$, and

(iii) $\mathscr{U}_L = \{(\mu, \omega) \mid \mu \in \mathscr{U},\ \omega \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}; \mathbb{R}^m)\}$. •

Controlled trajectories of the linearisation of Σ are then pairs $((\xi, \nu), (\mu, \omega))$ satisfying

$$\begin{aligned}\dot{\xi}(t) &= f(t, \xi(t), \mu(t)),\\ \dot{\nu}(t) &= D_1 f(t, \xi(t), \mu(t)) \cdot \nu(t) + D_2 f(t, \xi(t), \mu(t)) \cdot \omega(t),\end{aligned}$$

while controlled outputs satisfy

$$\begin{aligned}\eta(t) &= h(t, \xi(t), \mu(t)),\\ \gamma(t) &= D_1 h(t, \xi(t), \mu(t)) \cdot \nu(t) + D_2 h(t, \xi(t), \mu(t)) \cdot \omega(t).\end{aligned}$$

Thus we see that the linearisation encodes in its definition the original full system.

6.5.1.2 Linearisation about controlled equilibria

In this section we consider what amounts to a special case of linearisation about a controlled trajectory. The controlled trajectory we consider is a very particular sort of controlled trajectory, as given by the following definition.

6.5.3 Definition (Controlled equilibrium for a continuous-time state space system) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a continuous-time state space system. A pair $(x_0, u_0) \in X \times U$ is a ***controlled equilibrium*** for Σ if $f(t, x_0, u_0) = \mathbf{0}$ for every $t \in \mathbb{T}$. •

The following result gives the relationship between controlled equilibria and controlled trajectories.

6.5.4 Proposition (Controlled equilibria and constant controlled solutions) *Let* $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ *be a continuous-time state space system. Then a pair* $(\mathbf{x}_0, \mathbf{u}_0) \in \mathrm{X} \times \mathrm{U}$ *is a controlled equilibrium if and only if* $(\xi_0, \mu_0) \in \mathrm{Ctraj}(\Sigma)$*, where*

$$\xi_0(\mathrm{t}) = \mathbf{x}_0, \quad \mu_0(\mathrm{t}) = \mathbf{u}_0.$$

Proof First suppose that (x_0, u_0) is a controlled equilibrium. Then $\dot{\xi}_0(t) = \mathbf{0}$ for every $t \in \mathbb{T}$ and $f(t, \xi_0(t), \mu_0(t)) = \mathbf{0}$ and so

$$\dot{\xi}_0(t) = f(t, \xi_0(t), \mu_0(t)), \qquad t \in \mathbb{T},$$

and thus $(\xi_0, \mu_0) \in \mathrm{Ctraj}(\Sigma)$.

Next suppose that $(\xi_0, \mu_0) \in \mathrm{Ctraj}(\Sigma)$. Then

$$\mathbf{0} = \dot{\xi}_0(t) = f(t, \xi_0(t), \mu(t)) = f(t, x_0, u_0), \qquad t \in \mathbb{T},$$

so giving that (x_0, u_0) is a controlled equilibrium. ■

Note that, as a consequence of the preceding simple result, we can linearise about the constant controlled trajectory $t \mapsto (x_0, u_0)$ in the event that (x_0, u_0) is a controlled equilibrium. Let us, however, use some particular language in this case.

6.5.5 Definition (Linearisation of a continuous-time state space system about a controlled equilibrium) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a continuous-time state space system, supposing that $U \subseteq \mathbb{R}^m$ is open and that f_t and h_t are of class C^1 for every $t \in \mathbb{T}$, and let (x_0, u_0) be a controlled equilibrium. The ***linearisation of Σ about $(\mathbf{x_0}, \mathbf{u_0})$*** is the continuous-time state space system

$$\Sigma_{L,(x_0,u_0)} = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{T}, \mathscr{U}_L, f_{L,(x_0,u_0)}, h_{L,(x_0,u_0)}),$$

with

$$\begin{aligned} f_{L,(x_0,u_0)} \colon \mathbb{T} \times \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R}^n \\ (t, v, w) &\mapsto D_1 f(t, x_0, u_0) \cdot v + D_2 f(t, x_0, u_0) \cdot w, \end{aligned}$$

and

$$\begin{aligned} h_{L,(x_0,u_0)} \colon \mathbb{T} \times \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R}^k \\ (t, v, w) &\mapsto D_1 h(t, x_0, u_0) \cdot v + D_2 h(t, x_0, u_0) \cdot w, \end{aligned}$$

•

A controlled trajectory (ν, ω) for $f_{L,(x_0,u_0)}$ satisfies

$$\dot{\nu}(t) = A(t)(\nu(t)) + B(t)(\omega(t)),$$

where

$$A(t) = D_1 f(t, x_0, u_0), \quad B(t) = D_2 f(t, x_0, u_0).$$

The corresponding controlled output (y, ω) is given by

$$y(t) = C(t)(\nu(t)) + D(t)(\omega(t)),$$

where

$$C(t) = D_1 h(t, x_0, u_0), \quad D(t) = D_2 h(t, x_0, u_0).$$

Thus we see that the linearisation about a controlled equilibrium is a linear continuous-time state space system, as we shall see subsequently. What is special here, however, is that the linearisation is autonomous if Σ is autonomous. Thus the linearisation when Σ is autonomous is a linear continuous-time state space system with constant coefficients.

6.5.1.3 The flow of the linearisation In this section, in contrast with the preceding sections, we give a very precise characterisation of linearisation. It has the benefit of being precise, but the drawback of being complicated. However, the constructions we give in this section are of some importance in subjects like optimal control theory. We shall do three things: (1) provide conditions under which the flow of a continuous-time state space system is differentiable in state, initial time, and control, as well as final time with respect to which it is always differentiable; (2) give explicit formulae for the derivatives; (3) give an interpretation of these derivatives in terms of "wiggling" of initial conditions in state and time, and variations of the control.

We shall first investigate thoroughly the properties of the flow of a continuous-time state space system that has more regularity properties than are required for the basic existence and uniqueness theorem, Theorem 6.1.10. Let us suppose that we have a continuous-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$. We then have the controlled trajectory (ξ_0, μ_0) with

$$t \mapsto \xi_0(t) \triangleq \Phi^\Sigma(t, t_0, x_0, \mu_0)$$

defined for $t \in J_\Sigma(t_0, x_0, \mu_0)$. We then define

$$\begin{aligned} A_{(t_0,x_0,\mu_0)} \colon J_\Sigma(t_0, x_0, \mu_0) &\to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n) \\ t &\mapsto D_1 f(t, \Phi^\Sigma(t, t_0, x_0, \mu_0)) \end{aligned}$$

and

$$\begin{aligned} B_{(t_0,x_0,\mu_0)} \colon J_\Sigma(t_0, x_0, \mu_0) &\to \mathrm{L}(\mathbb{R}^m; \mathbb{R}^n) \\ t &\mapsto D_2 f(t, \Phi^\Sigma(t, t_0, x_0, \mu_0)). \end{aligned}$$

Now consider the continuous-time state space system $\Sigma_{L,(t_0,x_0,\mu_0)}$ of Definition 6.5.1. We consider first the following ordinary differential equation, defined for $t \in J_\Sigma(t_0, x_0, \mu_0)$:

$$\frac{\mathrm{d}\Psi}{\mathrm{d}s}(s) = A_{(t_0,x_0,\mu_0)}(s) \circ \Psi(s), \quad \Psi(t) = I_n.$$

We denote the solution at time s by $\boldsymbol{\Phi}_{A_{(t_0,x_0,\mu_0)}}(s,t)$; the associated map

$$\boldsymbol{\Phi}_{A_{(t_0,x_0,\mu_0)}} : J_\Sigma(t_0,\boldsymbol{x}_0,\boldsymbol{\mu}_0) \times J_\Sigma(t_0,\boldsymbol{x}_0,\boldsymbol{\mu}_0) \to \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$$

is the state transition map of Section 5.6.1.2. As such, the solution to the initial value problem

$$\frac{\mathrm{d}\boldsymbol{\nu}}{\mathrm{d}s}(s) = \boldsymbol{A}_{(t_0,x_0,\mu_0)}(s) \cdot \boldsymbol{\nu}(s), \quad \boldsymbol{\nu}(t) = \boldsymbol{v}_0 \tag{6.7}$$

is

$$\boldsymbol{\nu}(s) = \boldsymbol{\Phi}_{A_{(t_0,x_0,\mu_0)}}(s,t) \cdot \boldsymbol{v}_0, \qquad s \in J_\Sigma(t_0,\boldsymbol{x}_0,\boldsymbol{\mu}_0).$$

With the preceding background, we can now state the theorem.

6.5.6 Theorem (Differentiability of flows for continuous-time state space systems) *Let $\Sigma = (\mathrm{X},\mathrm{U},\mathbb{T},\mathscr{U},\mathbf{f},\mathbf{h})$ be a continuous-time state space system and make the following assumptions:*

(i) $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{m}}$ *is open;*

(ii) $\mathscr{U} = \mathsf{L}^\infty_{\mathrm{loc}}((\mathbb{T});\mathrm{U})$*;*

(iii) the map $\mathrm{t} \mapsto \mathbf{f}(\mathrm{t},\mathbf{x},\mathbf{u})$ *is measurable for each* $(\mathbf{x},\mathbf{u}) \in \mathrm{X} \times \mathrm{U}$*;*

(iv) the map $(\mathbf{x},\mathbf{u}) \mapsto \mathbf{f}(\mathrm{t},\mathbf{x},\mathbf{u})$ *is continuously differentiable for each* $\mathrm{t} \in \mathbb{T}$*;*

(v) for each $(\mathrm{t},\mathbf{x},\mathbf{u}) \in \mathbb{T} \times \mathrm{X} \times \mathrm{U}$*, there exist* $\alpha, \mathrm{r}, \rho \in \mathbb{R}_{>0}$ *and*

$$\mathrm{g}_0, \mathrm{g}_1 \in \mathsf{L}^1([\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha]; \mathbb{R}_{\geq 0})$$

such that

$$\|\mathbf{f}(\mathrm{s},\mathbf{y},\mathbf{v})\| \leq \mathrm{g}_0(\mathrm{s}), \quad (\mathrm{s},\mathbf{y},\mathbf{v}) \in ([\mathrm{t}_0 - \alpha, \mathrm{t}_0 + \alpha] \cap \mathbb{T}) \times \mathsf{B}^{\mathrm{n}}(\mathrm{r},\mathbf{x}) \times \mathsf{B}^{\mathrm{m}}(\rho,\mathbf{u}),$$

and

$$\left|\frac{\partial \mathrm{f}_{\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{s},\mathbf{y},\mathbf{v})\right|, \left|\frac{\partial \mathrm{f}_{\mathrm{j}}}{\partial \mathrm{u}_{\mathrm{a}}}(\mathrm{s},\mathbf{y},\mathbf{v})\right| \leq \mathrm{g}_1(\mathrm{t}),$$

$(\mathrm{s},\mathbf{y},\mathbf{v}) \in ([\mathrm{t}_0-\rho, \mathrm{t}_0+\rho]\cap\mathbb{T})\times\mathsf{B}^{\mathrm{n}}(\mathrm{r},\mathbf{x})\times\mathsf{B}^{\mathrm{m}}(\rho,\mathbf{u})$, $\mathrm{j},\mathrm{k} \in \{1,\dots,\mathrm{n}\}$, $\mathrm{a} \in \{1,\dots,\mathrm{m}\}$.

Then the following statements hold:

(vi) for $\mathrm{t},\mathrm{t}_0 \in \mathbb{T}$ *with* $\mathrm{t} \geq \mathrm{t}_0$, $\mathrm{D}_\Sigma(\mathrm{t},\mathrm{t}_0)$ *is open in* $\mathrm{X} \times \mathsf{L}^\infty([\mathrm{t}_0,\mathrm{t}];\mathrm{U})$*;*

(vii) $\Phi^\Sigma(\mathrm{t},\mathrm{t}_0)$ *is differentiable at* $(\mathbf{x}_0,\boldsymbol{\mu}_0) \in \mathrm{D}_\Sigma(\mathrm{t},\mathrm{t}_0)$ *and its derivative is given by*

$$\mathbf{D}\Phi^\Sigma_{\mathrm{t},\mathrm{t}_0}(\mathbf{x}_0,\boldsymbol{\mu}_0)\cdot(\mathbf{v},\boldsymbol{\omega}) = \boldsymbol{\Phi}_{\mathbf{A}_{(\mathrm{t}_0,\mathbf{x}_0,\mu_0)}}(\mathrm{t},\mathrm{t}_0)\cdot\mathbf{v} + \int_{\mathrm{t}_0}^{\mathrm{t}} \boldsymbol{\Phi}_{\mathbf{A}_{(\mathrm{t}_0,\mathbf{x}_0,\mu_0)}}(\mathrm{t},\tau)\mathbf{B}_{(\mathrm{t}_0,\mathbf{x}_0,\mu_0)}(\tau)\boldsymbol{\omega}(\tau)\,\mathrm{d}\tau;$$

(viii) the map

$$\begin{aligned} \mathbf{D}\Phi^{\Sigma}(t,t_0)\colon \mathrm{D}_{\Sigma}(t,t_0) &\to \mathrm{L}(\mathbb{R}^n \oplus \mathrm{L}^{\infty}([t_0,t];\mathbb{R}^m);\mathbb{R}^n) \\ (x,\mu) &\mapsto \mathbf{D}\Phi^{\Sigma}_{t,t_0}(x,\mu) \end{aligned}$$

is continuous.[5]

Proof In the proof of Theorem 5.1.8 we showed that the hypotheses of that theorem implies those of Theorem 3.2.13. We can similarly show that the hypotheses of the present theorem imply those of Theorem 6.1.14.

(vi) This follows from Theorem 6.1.14.

(vii) By virtue of the proof of Theorem 6.1.14 there exists $r, r', \rho, \alpha \in \mathbb{R}_{>0}$ such that, if $x \in \mathsf{B}^n(r, x_0)$, $\mu \in \mathsf{B}_{[t_0,t]}(\rho, \mu_0)$, and $t \in [t_0 - \alpha, t_0 + \alpha]$, then $\Phi^{\Sigma}(t, t_0, x, \mu)$ is defined and takes values in $\mathsf{B}^n(r', x_0)$. Moreover, we have

$$\Phi^{\Sigma}(t,t_0,x,\mu) = x + \int_{t_0}^{t} f(s,\Phi^{\Sigma}(s,t_0,x,\mu),\mu(s))\,\mathrm{d}s$$

in this case. We note that r', r, and α depend on g_0 and L_0 according to the required inequalities

$$\left|\int_{t_0}^{t} g_0(s)\,\mathrm{d}s\right| < \frac{r'}{2}, \quad \left|\int_{t_0}^{t} L_0(s)\,\mathrm{d}s\right| < \lambda$$

for some $\lambda \in (0,1)$.

Note that, by Proposition 5.2.2, the linear ordinary differential equation associated to the initial value problem (6.7) possesses unique solutions on $(t_0 - \alpha, t_0 + \alpha)$, cf. the corresponding conclusion in the proof of Theorem 5.1.8.

Now we show that, for each $t \in (t_0 - \alpha, t_0 + \alpha)$, Φ^{Σ}_{t,t_0} is differentiable at (x, μ) sufficiently close to (x_0, μ_0). As we argued in the proof of Theorem 6.1.14, we can assume without loss of generality that there is a compact set $L \subseteq U$ such that $\mu(t) \in L$ for almost every $t \in (t_0 - \alpha, t_0 + \alpha)$ and for every $\mu \in \mathsf{B}_{[t_0-\alpha,t_0+\alpha]}(\rho, U)$. Similarly, we can assume that there is a compact set $K \subseteq X$ such that

$$\Phi^{\Sigma}(t,t_0,x,\mu) \in K, \qquad (t,x,\mu) \in (t_0 - \alpha, t_0 + \alpha) \times \mathsf{B}^n(r,x_0) \times \mathsf{B}_{[t_0-\alpha,t_0+\alpha]}(\rho,U).$$

For this reason, we shall assume, without loss of generality and for simplicity, that f_t is uniformly continuous. By multiplying f by an infinitely differentiable function of (x,u) equal to 1 on $K \times L$, we can assume that $X = \mathbb{R}^n$ and $U = \mathbb{R}^m$. By the Fundamental Theorem of Calculus, for $(x,u) \in X \times U$, we have

$$\int_0^1 (D_1 f(t, x + sh, u + sw) \cdot h + D_2 f(t, x + sh, u + sw) \cdot w)\,\mathrm{d}s = f(t, x + h, u + w) - f(t, x, u).$$

[5] Note that we have not discussed the differentiability of mappings with open subsets of Banach spaces as their domain. However, if one thinks about things for a moment, one can see that the definitions of derivative in Section II-1.4.1 are immediately adapted to the setting of Banach spaces. Moreover, we shall quickly consider this general situation in the next situation.

Therefore,

$$\begin{aligned} f(t,x+h,u+w) - f(t,x,u) - D_1 f(t,x,u)\cdot h - D_2 f(t,x,u)\cdot w \\ = \int_0^1 ((\underbrace{D_1 f(t,x+sh,u+sw) - D_1 f(t,x,u)}_{A_1(t,s,x,u,h,w)})\cdot h \\ + \underbrace{(D_2 f(t,x+sh,u+sw) - D_2 f(t,x,u))}_{A_2(t,s,x,u,h,w)}\cdot w)\,\mathrm{d}s \end{aligned} \tag{6.8}$$

Define

$$M_t(h,w) = \sup\left\{ \int_0^1 |||A_1(t,s,x,u,h,w)||| + |||A_2(t,s,x,u,h,w)|||\,\mathrm{d}s \;\middle|\; (x,u)\in X\times U \right\},$$

and note that M_t is continuous (similarly to the argument in the corresponding part of the proof of Theorem 5.1.8) and that $M_t(\mathbf{0}) = 0$. For $x \in \mathsf{B}^n(r,x_0)$ and h small, consider the initial value problems

$$\dot\xi_0(t) = f(t,\xi_0(t),\mu_0(t)), \quad \xi_0(t_0) = x,$$

and

$$\dot\xi_1(t) = f(t,\xi_1(t),\mu(t)), \quad \xi_1(t_0) = x+h.$$

Denote

$$\delta(t) = \xi_1(t) - \xi_0(t), \quad \omega(t) = \mu(t) - \mu_0(t).$$

We then have

$$\begin{aligned} \dot\delta(t) &= f(t,\xi_0(t)+\delta(t),\mu_0(t)+\omega(t)) - f(t,\xi_0(t),\mu_0(t)) \\ &= \underbrace{D_1 f(t,\xi_0(t),\mu_0(t))}_{A_{(t_0,x,\mu_0)}(t)}\cdot\delta(t) + \underbrace{D_2 f(t,\xi_0(t),\mu_0(t))}_{B_{(t_0,x,\mu_0)}(t)}\cdot\omega(t) \\ &\quad + \underbrace{\int_0^1 (D_1 f(t,\xi_0(t)+s\delta(t),\mu_0(t)+s\omega(t)) - D_1 f(t,\xi_0(t),\mu_0(t)))\cdot\delta(t)\,\mathrm{d}s}_{e_1(t)} \\ &\quad + \underbrace{\int_0^1 (D_1 f(t,\xi_0(t)+s\delta(t),\mu_0(t)+s\omega(t)) - D_2 f(t,\xi_0(t),\mu_0(t)))\cdot\omega(t)\,\mathrm{d}s}_{e_2(t)}, \end{aligned}$$

using (5.1). Note that

$$\begin{aligned} \|e_1(t)\| &\le \int_0^1 \|D_1 f(t,\xi_0(t)+s\delta(t),\mu_0+s\omega(t)) - D_1 f(t,\xi_0(t),\mu_0(t))\cdot\delta(t)\|\,\mathrm{d}s \\ &\le \int_0^1 |||D_1 f(t,\xi_0(t)+s\delta(t)) - D_1 f(t,\xi_0(t))|||\,\|\delta(t)\|\,\mathrm{d}s \\ &\le \|\delta(t)\| M_t(\delta(t),\omega(t)). \end{aligned}$$

In a similar manner,

$$\|\boldsymbol{e}_2(t)\| \leq \|\boldsymbol{\omega}\|_{[t_0-\alpha,t_0+\alpha],\infty} M_t(\boldsymbol{\delta}(t), \boldsymbol{\omega}(t)).$$

Let $\boldsymbol{\nu}$ be the solution to the initial value problem

$$\dot{\boldsymbol{\nu}}(t) = \boldsymbol{A}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}(t) \cdot \boldsymbol{\nu}(t) + \boldsymbol{B}_{(t_0,\boldsymbol{x}_0,\boldsymbol{\mu}_0)}(t) \cdot \boldsymbol{\omega}(t), \quad \boldsymbol{\nu}(t_0) = \boldsymbol{h}.$$

Now, for fixed $t \in (t_0 - \alpha, t_0 + \alpha)$, we have

$$\boldsymbol{\delta}(t) = \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}}(t, t_0) \cdot \boldsymbol{h} + \int_{t_0}^{t} \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}}(t, \tau) \cdot (\boldsymbol{B}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}(\tau)\boldsymbol{\omega}(\tau) + \boldsymbol{e}_1(\tau) + \boldsymbol{e}_2(\tau)) \, \mathrm{d}\tau,$$

by Corollary 5.3.3, noting that $\boldsymbol{\delta}(t_0) = \boldsymbol{h}$. Thus

$$\boldsymbol{\delta}(t) = \boldsymbol{\nu}(t) + \int_{t_0}^{t} \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}}(t, \tau) \cdot (\boldsymbol{e}_1(\tau) + \boldsymbol{e}_2(\tau)) \, \mathrm{d}\tau,$$

again by Corollary 5.3.3 and noting that $\boldsymbol{\nu}(t_0) = \boldsymbol{h}$. Thus

$$\begin{aligned}
\|\boldsymbol{\delta}(t) - \boldsymbol{\nu}(t)\| &\leq \int_{t_0}^{t} |||\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}}(t, \tau)||| \, (\|\boldsymbol{e}_1(\tau)\| + \|\boldsymbol{e}_2(\tau)\|) \, \mathrm{d}\tau \\
&\leq (t - t_0)|||\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}}(t, \cdot)|||_\infty (\|\boldsymbol{e}_1\|_\infty + \|\boldsymbol{e}_2\|_\infty) \\
&\leq (t - t_0)|||\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}}(t, \cdot)|||_\infty (\|\boldsymbol{\delta}\|_{[t_0-\alpha,t_0+\alpha],\infty} + \|\boldsymbol{\omega}\|_{[t_0-\alpha,t_0+\alpha],\infty}) M_t(\boldsymbol{\delta}(t)),
\end{aligned}$$

where the ∞-norm is over the interval $[t_0, t]$. From the continuity of solutions with respect to control and initial condition as proved in Theorem 6.1.14, we have

$$\|\boldsymbol{\delta}\|_{[t_0-\alpha,t_0+\alpha],\infty} \leq C(\|\boldsymbol{h}\| + \|\boldsymbol{\omega}\|_{[t_0-\alpha,t_0+\alpha],\infty})$$

for some $C \in \mathbb{R}_{>0}$, cf. Lemma 1 from the proof of Theorem 3.2.13. Therefore,

$$\|\boldsymbol{\delta}(t) - \boldsymbol{\nu}(t)\| \leq C'(\|\boldsymbol{h}\| + \|\boldsymbol{\omega}\|_{[t_0-\alpha,t_0+\alpha],\infty}) M_t(\boldsymbol{\delta}(t)),$$

where $C' = C\alpha|||\boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x},\boldsymbol{\mu}_0)}}(\cdot, t_0)|||_\infty$. Restoring the pre-abbreviation notation, and taking limits as $\boldsymbol{h} \to \boldsymbol{0}$ and $\boldsymbol{\omega} \to 0$, we obtain this part of the result, in the same manner as the proof of Theorem 5.1.8(iv).

(viii) Next we show that Φ^{Σ}_{t,t_0} is *continuously* differentiable. This can be carried out by an adaptation of the corresponding part of the proof of Theorem 5.1.8, and we leave the tedious details to the reader.

To complete this part of the proof, we need to prove the statement globally. This can be carried out just as was the global part of the proof of Theorem 5.1.8. ■

The rôle of the output map in linearisation is more straightforward, e.g., Proposition 6.5.9 below.

6.5.2 Linearisation of continuous-time input/output systems

Let us next consider the linearisation of continuous-time input/output systems. In this situation, there is not very much one can say about linearisation, other than the fairly obvious thing. So here we point out this obvious thing, and indicate that it is consistent with the linearisation of state space systems in the preceding section.

We will need to be able to differentiate mappings between normed vector spaces. The definition is an obvious adaptation of the usual definition of a derivative between Euclidean spaces.

6.5.7 Definition (Derivative and differentiable map) Let $(\mathsf{X}, \|\cdot\|_{\mathsf{X}})$ and $(\mathsf{Y}, \|\cdot\|_{\mathsf{Y}})$ be normed $\mathbb{R}$-vector spaces. Let $U \subseteq \mathsf{X}$ be an open subset and let $f\colon U \to \mathsf{Y}$ be a map.

(i) The map f is ***differentiable at*** $\mathbf{x_0} \in \mathbf{U}$ if there exists a linear map $\mathsf{L}_{f,x_0}\colon \mathsf{X} \to \mathsf{Y}$ such that
$$\lim_{x \to x_0} \frac{\|f(x) - f(x_0) - \mathsf{L}_{f,x_0}(x - x_0)\|_{\mathsf{Y}}}{\|x - x_0\|_{\mathsf{X}}} = 0.$$

(ii) If f is differentiable at x_0, then the linear map L_{f,x_0}[6] is denoted by $\boldsymbol{D}f(x_0)$ and is called the ***derivative*** of f at x_0.

(iii) If f is differentiable at each point $x \in U$, then f is ***differentiable***.

(iv) If f is differentiable and if the map $x \mapsto \boldsymbol{D}f(x)$ is continuous (using the induced norm of Theorem III-3.5.14 for $\mathrm{L}(\mathsf{X}; \mathsf{Y})$) then f is ***continuously differentiable***, or of ***class*** $\mathbf{C^1}$. •

With this somewhat more abstract notion of differentiability and derivative at hand, we can define the notion of linearisation for continuous-time input/output systems.

6.5.8 Definition (Linearisation of input/output system about behaviour) Let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a continuous-time input/output system and let $(\mu_0, \eta_0) \in \mathscr{B}(\Sigma)$ with $\mu_0 \in \mathscr{U}(\mathbb{S})$.

(i) The system Σ is ***linearisable*** at (μ_0, η_0) if, for every compact sub-time-domain $\mathbb{K} \subseteq \mathbb{S}$,

  (a) $\mathscr{U}(\mathbb{K})$ is open and

  (b) $g_{\mathbb{K}}$ is differentiable at $\mu_0|\mathbb{K}$.

(ii) If Σ is linearisable at (μ_0, η_0), then its ***linearisation*** is the continuous-time input/output system
$$\Sigma_L = (\mathbb{R}^m, \mathbb{S}, \mathsf{L}^\infty(\mathbb{S}; \mathbb{R}^m), \mathsf{L}^\infty(\mathbb{S}; \mathbb{R}^k), g_L)$$
where g_L is defined by requiring that
$$g_L(\omega|\mathbb{K}) = \boldsymbol{D}g_{\mathbb{K}}(\mu_0|\mathbb{K}) \cdot (\omega|\mathbb{K})$$

[6] One can show, just as in Proposition II-1.4.1, that the linear map L_{f,x_0} is unique if it exists.

for every compact sub-time-domain $\mathbb{K} \subseteq \mathbb{S}$. •

We can relate the linearisation of a continuous-time state space system to the linearisation of its associated continuous-time input/output system. We adopt the notation from the statement of Theorem 6.5.6.

6.5.9 Proposition (Linearisation of input/output systems determined from state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a continuous-time state space system satisfying the hypotheses of Theorem 6.5.6. Assume that Σ is proper, output autonomous, and that $\mathbf{h}$ is a continuously differentiable mapping from X to $\mathbb{R}^k$. Let $(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_0) \in \mathrm{D}_\Sigma$ and define $\boldsymbol{\eta}_0 \colon [t_0, t] \to \mathbb{R}^k$ by*

$$\boldsymbol{\eta}_0(t) = \mathbf{h}(\Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_0)).$$

Consider the continuous-time input/output system

$$\Sigma_{i/o}(t_0, \mathbf{x}_0) = (\mathrm{U}, [t_0, t], \mathscr{U}, \mathscr{Y}, \mathbf{g})$$

as in Theorem 6.2.10. Then $\Sigma_{i/o}(t_0, \mathbf{x}_0)$ is linearisable at $(\boldsymbol{\mu}_0, \boldsymbol{\eta}_0)$ and its linearisation is

$$\mathbf{D}\mathbf{g}_{[t_0,t]}(\boldsymbol{\mu}_0) \cdot \boldsymbol{\omega} = \int_{t_0}^{t} \mathbf{D}\mathbf{h}(\Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_0)) \circ \boldsymbol{\Phi}_{\mathbf{A}_{(t_0,\mathbf{x}_0,\boldsymbol{\mu}_0)}}(t, \tau)\mathbf{B}_{(t_0,\mathbf{x}_0,\boldsymbol{\mu}_0)}(\tau)\boldsymbol{\omega}(\tau)\,\mathrm{d}\tau.$$

Proof We have

$$g_{[t_0,t]}(\boldsymbol{\mu})(\tau) = h(\Phi^\Sigma(\tau, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}))$$

for $(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}) \in D_\Sigma$. Therefore, by the Chain Rule (which holds, with the same proof, for mappings between open subsets of normed vector spaces),

$$\boldsymbol{D}g_{[t_0,t]}(\boldsymbol{\mu}_0) \cdot \boldsymbol{\omega} = \boldsymbol{D}h(\Phi^\Sigma(t, t_0, \boldsymbol{x}_0, \boldsymbol{\mu}_0)) \circ \boldsymbol{D}_2\Phi^\Sigma_{t,t_0}(\boldsymbol{x}_0, \boldsymbol{\mu}_0) \cdot \boldsymbol{\omega}.$$

From Theorem 6.5.6(vii) we have

$$\boldsymbol{D}_2\Phi^\Sigma_{t,t_0}(\boldsymbol{x}_0, \boldsymbol{\mu}_0) \cdot \boldsymbol{\omega} = \int_{t_0}^{t} \boldsymbol{\Phi}_{\boldsymbol{A}_{(t_0,\boldsymbol{x}_0,\boldsymbol{\mu}_0)}}(t, \tau)\boldsymbol{B}_{(t_0,\boldsymbol{x}_0,\boldsymbol{\mu}_0)}(\tau)\boldsymbol{\omega}(\tau)\,\mathrm{d}\tau.$$

From this the result follows. ■

6.5.3 Linearisation of discrete-time state space systems

We now repeat what we have done in the preceding two sections for discrete-time systems. As expected, there is a strong resemblance with the results from Section 5.1.2 on linearisation of ordinary difference equations.

6.5.3.1 Linearisation along controlled trajectories Suppose that we have a discrete-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ and a controlled trajectory (ξ_0, μ_0) defined on a sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}$. We wish to understand what happens to solutions "nearby" this fixed controlled trajectory (ξ_0, μ_0).

To do this, we suppose that $U \subseteq \mathbb{R}^m$ is open and that $\boldsymbol{f}$ and $\boldsymbol{h}$ are continuously differentiable as functions of $(\boldsymbol{x}, \boldsymbol{u})$. That is, for $t \in \mathbb{T}$ we denote

$$\begin{aligned} \boldsymbol{f}_t &\colon X \times U \to \mathbb{R}^n \\ &\quad (\boldsymbol{x}, \boldsymbol{u}) \mapsto \boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{u}), \\ \boldsymbol{h}_t &\colon X \times U \to \mathbb{R}^k \\ &\quad (\boldsymbol{x}, \boldsymbol{u}) \mapsto \boldsymbol{h}(t, \boldsymbol{x}, \boldsymbol{u}), \end{aligned}$$

and we require that $\boldsymbol{f}_t$ and $\boldsymbol{h}_t$ be of class C^1 for each $t \in \mathbb{T}$. We denote

$$\begin{aligned} \boldsymbol{D}_1\boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{u}) &= \boldsymbol{D}_1\boldsymbol{f}_t(\boldsymbol{x}, \boldsymbol{u}), \\ \boldsymbol{D}_2\boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{u}) &= \boldsymbol{D}_2\boldsymbol{f}_t(\boldsymbol{x}, \boldsymbol{u}), \\ \boldsymbol{D}_1\boldsymbol{h}(t, \boldsymbol{x}, \boldsymbol{u}) &= \boldsymbol{D}_1\boldsymbol{h}_t(\boldsymbol{x}, \boldsymbol{u}), \\ \boldsymbol{D}_2\boldsymbol{h}(t, \boldsymbol{x}, \boldsymbol{u}) &= \boldsymbol{D}_2\boldsymbol{h}_t(\boldsymbol{x}, \boldsymbol{u}), \qquad t \in \mathbb{T}. \end{aligned}$$

the partial derivatives with respect to $\boldsymbol{x}$ and $\boldsymbol{u}$, respectively, with t fixed. Thus

$$\begin{aligned} \boldsymbol{D}_1\boldsymbol{f} &\colon \mathbb{T} \times X \times U \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n), \\ \boldsymbol{D}_2\boldsymbol{f} &\colon \mathbb{T} \times X \times U \to \mathrm{L}(\mathbb{R}^m; \mathbb{R}^n), \\ \boldsymbol{D}_1\boldsymbol{h} &\colon \mathbb{T} \times X \times U \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^k), \\ \boldsymbol{D}_2\boldsymbol{h} &\colon \mathbb{T} \times X \times U \to \mathrm{L}(\mathbb{R}^m; \mathbb{R}^k). \end{aligned}$$

We then suppose that we have a controlled trajectory (ξ_0, μ_0), defined on $\mathbb{T}'$, for Σ for which the deviations $\boldsymbol{\nu} \triangleq \boldsymbol{\xi} - \boldsymbol{\xi}_0$ and $\boldsymbol{\omega} = \boldsymbol{\mu} - \boldsymbol{\mu}_0$ are small. Let us try to understand the behaviour of $\boldsymbol{\nu}$. Naïvely, we can do this as follows:

$$\begin{aligned} \boldsymbol{\xi}(t + \Delta) &= (\boldsymbol{\xi}_0 + \boldsymbol{\nu})(t + \Delta) = \boldsymbol{f}(t, \boldsymbol{\xi}_0(t) + \boldsymbol{\nu}(t), \boldsymbol{\mu}_0 + \boldsymbol{\omega}(t)) \\ &= \boldsymbol{f}(t, \boldsymbol{\xi}_0(t), \boldsymbol{\mu}_0(t)) + \boldsymbol{D}_1\boldsymbol{f}(t, \boldsymbol{\xi}_0(t), \boldsymbol{\mu}_0(t)) \cdot \boldsymbol{\nu}(t) + \boldsymbol{D}_2\boldsymbol{f}(t, \boldsymbol{\xi}_0(t), \boldsymbol{\mu}_0(t)) \cdot \boldsymbol{\omega}(t) + \cdots . \end{aligned}$$

We will not here try to be precise about what "$\cdots$" might mean, but merely say that the idea of the preceding equation is that we approximate using the constant and first-order terms in the Taylor expansion, and then pray that this gives us something meaningful. Note that, since $(\boldsymbol{\xi}_0, \boldsymbol{\mu}_0)$ is a controlled trajectory for Σ, the approximation we arrive at is

$$\boldsymbol{\nu}(t + \Delta) \approx \boldsymbol{D}_1\boldsymbol{f}(t, \boldsymbol{\xi}_0(t), \boldsymbol{\mu}_0(t)) \cdot \boldsymbol{\nu}(t) + \boldsymbol{D}_2\boldsymbol{f}(t, \boldsymbol{\xi}_0(t), \boldsymbol{\mu}_0(t)) \cdot \boldsymbol{\omega}(t).$$

We similarly denote

$$\boldsymbol{\eta}_0(t) = \boldsymbol{h}(t, \boldsymbol{\xi}_0(t), \boldsymbol{\mu}_0(t)), \quad \boldsymbol{\eta}(t) = \boldsymbol{h}(t, \boldsymbol{\xi}(t), \boldsymbol{\mu}(t)),$$

and deduce, with $\boldsymbol{\gamma}(t) = \boldsymbol{\eta}(t) - \boldsymbol{\eta}_0(t)$, that we have an approximation

$$\boldsymbol{\gamma}(t) \approx \boldsymbol{D}_1\boldsymbol{h}(t, \boldsymbol{\xi}_0(t), \boldsymbol{\mu}_0(t)) \cdot \boldsymbol{\nu}(t) + \boldsymbol{D}_2\boldsymbol{h}(t, \boldsymbol{\xi}_0(t), \boldsymbol{\mu}_0(t)) \cdot \boldsymbol{\omega}(t).$$

Meaningful or not, the preceding naïve calculations give rise to the following definition.

6.5.10 Definition (Linearisation of a discrete-time state space system along a controlled trajectory) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a discrete-time state space system, supposing that $U \subseteq \mathbb{R}^m$ is open and that f_t and h_t are of class C^1 for every $t \in \mathbb{T}$. For $(\xi_0, \mu_0) \in \mathrm{Ctraj}(\Sigma)$ with domain $\mathbb{T}'$, the ***linearisation of Σ along (ξ_0, μ_0)*** is the discrete-time state space system

$$\Sigma_{L,(\xi_0,\mu_0)} = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{T}', \mathscr{U}_L, f_{L,(\xi_0,\mu_0)}, h_{L,(\xi_0,\mu_0)}),$$

where

(i) $f_{L,(\xi_0,\mu_0)}(t, v, w) = D_1 f(t, \xi_0(t), \mu_0(t)) \cdot v + D_2 f(t, \xi_0(t), \mu_0(t)) \cdot w$,

(ii) $h_{L,(\xi_0,\mu_0)}(t, v, w) = D_1 h(t, \xi_0(t), \mu_0(t)) \cdot v + D_2 h(t, \xi_0(t), \mu_0(t)) \cdot w$, and

(iii) $\mathscr{U}_L \subseteq \ell_{\mathrm{loc}}(\mathbb{T}'; \mathbb{R}^m)$. •

Note that a controlled trajectory for the linearisation of Σ along (ξ_0, μ_0) satisfies

$$\nu(t + \Delta) = A(t)(\nu(t)) + B(t)(\omega(t)),$$

where

$$A(t) = D_1 f(t, \xi_0(t), \mu_0(t)), \quad B(t) = D_2 f(t, \xi_0(t), \mu_0(t)).$$

The corresponding controlled outputs satisfy

$$\gamma(t) = C(t)(\nu(t)) + D(t)(\omega(t)),$$

where

$$C(t) = D_1 h(t, \xi_0(t), \mu_0(t)), \quad D(t) = D_2 h(t, \xi_0(t), \mu_0(t)).$$

This is what we shall subsequently refer to as a linear discrete-time state space system.

Note that there is an alternative view of linearisation that can be easily developed, one where linearisation is of the *system*, not just along a controlled trajectory. The construction we make is the following.

6.5.11 Definition (Linearisation of a discrete-time state space system) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a discrete-time state space system, supposing that $U \subseteq \mathbb{R}^m$ is open and that f_t and h_t are of class C^1 for every $t \in \mathbb{T}$. The ***linearisation*** of Σ is the discrete-time state space system

$$\Sigma_L = (X \times \mathbb{R}^n, U \times \mathbb{R}^m, \mathbb{T}, \mathscr{U}_L, f_L, h_L),$$

where

(i) $f_L(t, (x, v), (u, w)) = (f(t, x, u), D_1 f(t, x, u) \cdot v + D_2 f(t, x, u) \cdot w)$,

(ii) $h_L(t, (x, v), (u, w)) = (h(t, x, u), D_1 h(t, x, u) \cdot v + D_2 h(t, x, u) \cdot w)$, and

(iii) $\mathscr{U}_L = \{(\mu, \omega) \mid \mu \in \mathscr{U},\ \omega \in \ell_{\mathrm{loc}}(\mathbb{T}; \mathbb{R}^m)\}$. •

Controlled trajectories of the linearisation of Σ are then pairs $((\xi, \nu), (\mu, \omega))$ satisfying

$$\begin{aligned}\xi(t + \Delta) &= f(t, \xi(t), \mu(t)),\\ \nu(t + \Delta) &= D_1 f(t, \xi(t), \mu(t)) \cdot \nu(t) + D_2 f(t, \xi(t), \mu(t)) \cdot \omega(t),\end{aligned}$$

while controlled outputs satisfy

$$\begin{aligned}\eta(t) &= h(t, \xi(t), \mu(t)),\\ \gamma(t) &= D_1 h(t, \xi(t), \mu(t)) \cdot \nu(t) + D_2 h(t, \xi(t), \mu(t)) \cdot \omega(t).\end{aligned}$$

Thus we see that the linearisation encodes in its definition the original full system.

6.5.3.2 Linearisation about controlled equilibria In this section we consider what amounts to a special case of linearisation about a controlled trajectory. The controlled trajectory we consider is a very particular sort of controlled trajectory, as given by the following definition.

6.5.12 Definition (Controlled equilibrium for a discrete-time state space system) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a discrete-time state space system. A pair $(x_0, u_0) \in X \times U$ is a ***controlled equilibrium*** for Σ if $f(t, x_0, u_0) = x_0$ for every $t \in \mathbb{T}$. •

The following result gives the relationship between controlled equilibria and controlled trajectories.

6.5.13 Proposition (Controlled equilibria and constant controlled solutions) *Let* $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ *be a discrete-time state space system. Then a pair* $(\mathbf{x}_0, \mathbf{u}_0) \in \mathrm{X} \times \mathrm{U}$ *is a controlled equilibrium if and only if* $(\xi_0, \mu_0) \in \mathrm{Ctraj}(\Sigma)$*, where*

$$\xi_0(\mathrm{t}) = \mathbf{x}_0, \quad \mu_0(\mathrm{t}) = \mathbf{u}_0.$$

Proof First suppose that (x_0, u_0) is a controlled equilibrium. Then $\xi_0(t + \Delta) = x_0$ for every $t \in \mathbb{T}$ and $f(t, \xi_0(t), \mu_0(t)) = x_0$ and so

$$\xi_0(t + \Delta) = f(t, \xi_0(t), \mu_0(t)), \qquad t \in \mathbb{T},$$

and thus $(\xi_0, \mu_0) \in \mathrm{Ctraj}(\Sigma)$.

Next suppose that $(\xi_0, \mu_0) \in \mathrm{Ctraj}(\Sigma)$. Then

$$x_0 = \xi_0(t + \Delta) = f(t, \xi_0(t), \mu(t)) = f(t, x_0, u_0), \qquad t \in \mathbb{T},$$

so giving that (x_0, u_0) is a controlled equilibrium. ■

Note that, as a consequence of the preceding simple result, we can linearise about the constant controlled trajectory $t \mapsto (x_0, u_0)$ in the event that (x_0, u_0) is a controlled equilibrium. Let us, however, use some particular language in this case.

6.5.14 Definition (Linearisation of a discrete-time state space system about a controlled equilibrium) Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$ be a discrete-time state space system, supposing that $U \subseteq \mathbb{R}^m$ is open and that f_t and h_t are of class C^1 for every $t \in \mathbb{T}$, and let (x_0, u_0) be a controlled equilibrium. The ***linearisation of Σ about $(\mathbf{x_0}, \mathbf{u_0})$*** is the discrete-time state space system

$$\Sigma_{L,(x_0,u_0)} = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{T}, \mathscr{U}_L, f_{L,(x_0,u_0)}, h_{L,(x_0,u_0)}),$$

with

$$\begin{aligned} f_{L,(x_0,u_0)} \colon \mathbb{T} \times \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R}^n \\ (t, v, w) &\mapsto D_1 f(t, x_0, u_0) \cdot v + D_2 f(t, x_0, u_0) \cdot w, \end{aligned}$$

and

$$\begin{aligned} h_{L,(x_0,u_0)} \colon \mathbb{T} \times \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R}^k \\ (t, v, w) &\mapsto D_1 h(t, x_0, u_0) \cdot v + D_2 h(t, x_0, u_0) \cdot w, \end{aligned}$$

•

A controlled trajectory (ν, ω) for $f_{L,(x_0,u_0)}$ satisfies

$$\nu(t + \Delta) = A(t)(\nu(t)) + B(t)(\omega(t)),$$

where

$$A(t) = D_1 f(t, x_0, u_0), \quad B(t) = D_2 f(t, x_0, u_0).$$

The corresponding controlled output (γ, ω) is given by

$$\gamma(t) = C(t)(\nu(t)) + D(t)(\omega(t)),$$

where

$$C(t) = D_1 h(t, x_0, u_0), \quad D(t) = D_2 h(t, x_0, u_0).$$

Thus we see that the linearisation about a controlled equilibrium is a linear discrete-time state space system, as we shall see subsequently. What is special here, however, is that the linearisation is autonomous if Σ is autonomous. Thus the linearisation when Σ is autonomous is a linear discrete-time state space system with constant coefficients.

6.5.3.3 The flow of the linearisation In this section, in contrast with the preceding sections, we give a very precise characterisation of linearisation. It has the benefit of being precise, but the drawback of being complicated. However, the constructions we give in this section are of some importance in subjects like optimal control theory. We shall do three things: (1) provide conditions under which the flow of a discrete-time state space system is differentiable in state, initial time, and control, as well as final time with respect to which it is always differentiable; (2) give explicit formulae for the derivatives; (3) give an interpretation of these derivatives in terms of "wiggling" of initial conditions in state and time, and variations of the control.

We shall first investigate thoroughly the properties of the flow of a discrete-time state space system that has more regularity properties than are required for the basic existence and uniqueness theorem, Theorem 6.3.9. Let us suppose that we have a discrete-time state space system $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, f, h)$. We then have the controlled trajectory (ξ_0, μ_0) with

$$t \mapsto \xi_0(t) \triangleq \Phi^\Sigma(t, t_0, x_0, \mu_0)$$

defined for $t \in J_\Sigma(t_0, x_0, \mu_0)$. We then define

$$\begin{aligned} A_{(t_0,x_0,\mu_0)} \colon J_\Sigma(t_0, x_0, \mu_0) &\to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n) \\ t &\mapsto D_1 f(t, \Phi^\Sigma(t, t_0, x_0, \mu_0)) \end{aligned}$$

and

$$\begin{aligned} B_{(t_0,x_0,\mu_0)} \colon J_\Sigma(t_0, x_0, \mu_0) &\to \mathrm{L}(\mathbb{R}^m; \mathbb{R}^n) \\ t &\mapsto D_2 f(t, \Phi^\Sigma(t, t_0, x_0, \mu_0)). \end{aligned}$$

Now consider the discrete-time state space system $\Sigma_{L,(t_0,x_0,\mu_0)}$ of Definition 6.5.10. We consider first the following ordinary difference equation, defined for $t \in J_\Sigma(t_0, x_0, \mu_0)$:

$$\Psi(s + \Delta) = A_{(t_0,x_0,\mu_0)}(s) \circ \Psi(s), \quad \Psi(t) = I_n.$$

As a linear ordinary difference equation, this initial value problem has solutions defined for all $s \in J_\Sigma(t_0, x_0, \mu_0) \cap \mathbb{T}_{\geq t}$. Moreover, we denote the solution at time s by $\Phi_{A_{(t_0,x_0,\mu_0)}}(s, t)$; the associated map

$$\Phi_{A_{(t_0,x_0,\mu_0)}} \colon P_\Sigma(t_0, x_0, \mu_0) \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$$

is the state transition map of Section 5.6.1.2. Here we denote

$$P_\Sigma(t_0, x_0, \mu_0) = \{(s, t) \in J_\Sigma(t_0, x_0, \mu_0) \cap \mathbb{T}_{\geq t_0} \times J_\Sigma(t_0, x_0, \mu_0) \cap \mathbb{T}_{\geq t_0} \mid s \geq t\}.$$

In particular, we shall use the fact that the solution to the initial value problem

$$v(s + h) = A_{(t_0,x_0,\mu_0)}(s) \cdot v(s), \quad v(t) = v_0$$

is

$$v(s) = \Phi_{A_{(t_0,x_0,\mu_0)}}(s, t) \cdot v_0, \qquad s \in J_\Sigma(t_0, x_0, \mu_0) \cap \mathbb{T}_{\geq t}.$$

With the preceding background, we can now state the theorem.

6.5.15 Theorem (Differentiability of flows for discrete-time state space systems) *Let $\Sigma = (\mathrm{X}, \mathrm{U}, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a discrete-time state space system and make the following assumptions:*

(i) $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{m}}$ *is open;*

(ii) $\mathscr{U} = \ell^\infty_{\mathrm{loc}}((\mathbb{T}); \mathrm{U})$*;*

(iii) $\mathbf{f}_{\mathrm{t}}$ *is of class* C^1 *for each* $\mathrm{t} \in \mathbb{T}$*.*

Then the following statements hold:

(iv) *for* $t, t_0 \in \mathbb{T}$ *with* $t \geq t_0$, $D_\Sigma(t, t_0)$ *is open in* $X \times \ell^\infty([t_0, t]; U)$;

(v) $\Phi^\Sigma(t, t_0)$ *is differentiable at* $(x_0, \mu_0) \in D_\Sigma(t, t_0)$ *and its derivative is given by*

$$\begin{aligned}&\mathbf{D}\Phi^\Sigma_{t,t_0}(x_0, \mu_0) \cdot (v, \omega)\\ &= \mathbf{\Phi}_{\mathbf{A}_{(t_0,x_0,\mu_0)}}(t, t_0) \cdot v + \sum_{j=0}^{(t-t_0-\Delta)/\Delta} \mathbf{\Phi}_{\mathbf{A}_{(t_0,x_0,\mu_0)}}(t-\Delta, t_0+j\Delta)(\mathbf{B}_{(t_0,x_0,\mu_0)}(t_0+j\Delta)\omega(t_0+j\Delta));\end{aligned}$$

(vi) *the map*

$$\begin{aligned}\mathbf{D}\Phi^\Sigma(t, t_0) \colon D_\Sigma(t, t_0) &\to L(\mathbb{R}^n \oplus \ell^\infty([t_0, t]; \mathbb{R}^m); \mathbb{R}^n)\\ (x, \mu) &\mapsto \mathbf{D}\Phi^\Sigma_{t,t_0}(x, \mu)\end{aligned}$$

is continuous.

Proof The proof uses easy variants of the arguments used in the proof of Theorem 6.5.6, making use of Corollary 5.7.2 in place of Corollary 5.3.3, noting that $\ell^\infty_{\mathrm{loc}}([t_0, t]; \mathbb{R}^m)$ is finite-dimensional, and using the Chain Rule as in the proof of Theorem 5.1.22. ■

We refer to Proposition 6.5.17 below for the contribution to the linearisation of the output map h.

6.5.4 Linearisation of discrete-time input/output systems

As with the linearisation of continuous-time input/output systems, linearisation is quite straightforward for discrete-time input/output systems.

6.5.16 Definition (Linearisation of input/output system about behaviour) Let $\Sigma = (U, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$ be a discrete-time input/output system and let $(\mu_0, \eta_0) \in \mathscr{B}(\Sigma)$ with $\mu_0 \in \mathscr{U}(\mathbb{S})$.

(i) The system Σ is ***linearisable*** at (μ_0, η_0) if, for every finite sub-time-domain $\mathbb{K} \subseteq \mathbb{S}$,

  (a) $\mathscr{U}(\mathbb{K})$ is open and

  (b) $g_{\mathbb{K}}$ is differentiable at $\mu_0|\mathbb{K}$.

(ii) If Σ is linearisable at (μ_0, η_0), then its ***linearisation*** is the discrete-time input/output system

$$\Sigma_L = (\mathbb{R}^m, \mathbb{S}, \ell^\infty(\mathbb{S}; \mathbb{R}^m), \ell^\infty(\mathbb{S}; \mathbb{R}^k), g_L)$$

where g_L is defined by requiring that

$$g_L(\omega|\mathbb{K}) = Dg_{\mathbb{K}}(\mu_0|\mathbb{K}) \cdot (\omega|\mathbb{K})$$

for every finite sub-time-domain $\mathbb{K} \subseteq \mathbb{S}$. •

We can relate the linearisation of a discrete-time state space system to the linearisation of its associated discrete-time input/output system. We adopt the notation from the statement of Theorem 6.5.15.

6.5.17 Proposition (Linearisation of input/output systems determined from state space systems) *Let $\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathbf{f}, \mathbf{h})$ be a discrete-time state space system satisfying the hypotheses of Theorem 6.5.15. Assume that Σ is proper, output autonomous, and that $\mathbf{h}$ is a continuously differentiable mapping from X to $\mathbb{R}^k$. Let $(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_0) \in D_\Sigma$ and define $\boldsymbol{\eta}_0 \colon [t_0, t] \to \mathbb{R}^k$ by*

$$\boldsymbol{\eta}_0(t) = \mathbf{h}(\Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_0)).$$

Consider the discrete-time input/output system

$$\Sigma_{i/o}(t_0, \mathbf{x}_0) = (U, [t_0, t], \mathscr{U}, \mathscr{Y}, \mathbf{g})$$

as in Theorem 6.4.10. Then $\Sigma_{i/o}(t_0, \mathbf{x}_0)$ is linearisable at $(\boldsymbol{\mu}_0, \boldsymbol{\eta}_0)$ and its linearisation is

$$\begin{aligned}\mathbf{Dg}_{[t_0,t]}(\boldsymbol{\mu}_0) \cdot \boldsymbol{\omega} = \sum_{j=0}^{(t-t_0-\Delta)/\Delta} &\mathbf{Dh}(\Phi^\Sigma(t, t_0, \mathbf{x}_0, \boldsymbol{\mu}_0)) \\ &\circ \boldsymbol{\Phi}_{\mathbf{A}_{(t_0,\mathbf{x}_0,\boldsymbol{\mu}_0)}}(t - \Delta, t_0 + j\Delta)(\mathbf{B}_{(t_0,\mathbf{x}_0,\boldsymbol{\mu}_0)}(t_0 + j\Delta)\boldsymbol{\omega}(t_0 + j\Delta)).\end{aligned}$$

Proof The result follows in the same manner as does Proposition 6.5.9. ■

Exercises

6.5.1 Consider a linear continuous-time state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

with constant coefficients as in Section 6.6.2. Answer the following questions.

(a) Use (6.10) to write an explicit formula for the outputs for Σ.
(b) Determine the linearisation Σ_L of Σ as in Definition 6.5.2.
(c) Write an explicit formula for the outputs for Σ_L.
(d) What can you say about the linearisation about the controlled equilibrium $(0, 0)$, i.e., about the zero control and the resulting zero trajectory?

6.5.2 For a control-affine continuous-time state space system

$$\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H}),$$

do the following.

(a) Determine the form of the linearisation of the system Σ_L from Definition 6.5.2.
(b) Is there anything special about the controlled equilibria for a control-affine continuous-time state space system as compared to a general continuous-time state space system?
Hint: *What if the zero control is in $\mathscr{U}$?*

6.5.3 For the circuit with a diode from Exercise 6.1.8, do the following.

(a) Determine the linearisation as in Definition 6.5.2.
(b) Characterise the controlled equilibria for the system.
(c) Determine the linearisation about controlled equilibria as in Definition 6.5.5.

6.5.4 For the forced pendulum from Exercise 6.1.9, do the following.
(a) Determine the linearisation as in Definition 6.5.2.
(b) Characterise the controlled equilibria for the system.
(c) Determine the linearisation about controlled equilibria as in Definition 6.5.5.

6.5.5 For the running mean and standard deviation from Exercise 6.2.5, do the following.
(a) Determine the linearisation as in Definition 6.5.2.
(b) Characterise the controlled equilibria for the system.
(c) Determine the linearisation about controlled equilibria as in Definition 6.5.5.

6.5.6 Consider a linear discrete-time state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

with constant coefficients as in Section 6.8.2. Answer the following questions.
(a) Use (6.13) to write an explicit formula for the outputs for Σ.
(b) Determine the linearisation Σ_L of Σ as in Definition 6.5.11.
(c) Write an explicit formula for the outputs for Σ_L.
(d) What can you say about the linearisation about the controlled equilibrium $(0, 0)$, i.e., about the zero control and the resulting zero trajectory?

6.5.7 For a control-affine discrete-time state space system

$$\Sigma = (X, U, \mathbb{T}, \mathscr{U}, \mathscr{F}, \mathscr{H}),$$

do the following.
(a) Determine the form of the linearisation of the system Σ_L from Definition 6.5.11.
(b) Is there anything special about the controlled equilibria for a control-affine discrete-time state space system as compared to a general discrete-time state space system?
Hint: *What if the zero control is in $\mathscr{U}$?*

6.5.8 For the running mean and standard deviation from Exercise 6.4.3, do the following.
(a) Determine the linearisation as in Definition 6.5.11.
(b) Characterise the controlled equilibria for the system.
(c) Determine the linearisation about controlled equilibria as in Definition 6.5.14.

Section 6.6

Linear continuous-time state space systems

In this section we begin to study the main objects of interest to us in this and the subsequent few chapters. We consider a particular class of continuous-time state space systems that are linear in both state and control. We mirror what we have done with ordinary differential equations by working with systems that are time-dependent, and then time-independent. One of the special things we shall focus on is the sorts of inputs and outputs one can consider. Linearity will allow us to obtain more particular results than we were able to obtain for not necessarily linear systems. In this section we work with continuous-time systems.

Do I need to read this section? This section is a core section in the volume. •

6.6.1 Systems with time-varying coefficients

Let us begin with the definition, recalling from Section 3.1.3.3 the adaptation to using abstract vector spaces in place of Euclidean spaces for linear systems.

6.6.1 Definition (Linear continuous-time state space system) A ***linear continuous-time state space system*** is a nonuple

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

where

(i) X (the ***state space***), U (the ***input space***), and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathbb{T} \subseteq \mathbb{R}$ is a continuous time-domain,

(iii) $\mathsf{A} \colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$, $\mathsf{B} \colon \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{X})$, $\mathsf{C} \colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{Y})$, and $\mathsf{D} \colon \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$, and

(iv) $\mathscr{U}$ is a collection of mappings $\mu \colon \mathbb{T} \to \mathsf{U}$. •

We note that a linear continuous-time state space system is, in particular, a continuous-time state space system (with the mild adaptation from using Euclidean spaces to using finite-dimensional vector spaces) with dynamics defined by

$$\begin{aligned} f \colon \mathbb{T} \times (\mathsf{X} \oplus \mathsf{U}) &\to \mathsf{X} \\ (t, x, u) &\mapsto \mathsf{A}(t)(x) + \mathsf{B}(t)(u) \end{aligned}$$

and with output map

$$\begin{aligned} h \colon \mathbb{T} \times (\mathsf{X} \oplus \mathsf{U}) &\to \mathsf{Y} \\ (t, x, u) &\mapsto \mathsf{C}(t)(x) + \mathsf{D}(t)(u). \end{aligned}$$

We note that linear continuous-time state space systems are, in fact, control-affine continuous-time state space systems. Therefore, all the notions attached

to continuous-time state space systems can be applied to those that are linear. The system theoretic attributes of Section 6.1.1 apply in exactly the same way for linear continuous-time state space systems; the reader can flesh this out in Exercise 6.6.1. One has the set Ctraj(Σ) of controlled trajectories and the set Cout(Σ) of controlled outputs. In particular, if (ξ, μ) is a controlled trajectory with (η, μ) the corresponding controlled output, then these satisfy the equations

$$\begin{aligned}\dot{\xi}(t) &= \mathsf{A}(t)(\xi(t)) + \mathsf{B}(t)(\mu(t)),\\ \eta(t) &= \mathsf{C}(t)(\xi(t)) + \mathsf{D}(t)(\mu(t)).\end{aligned}$$

Moreover, the existence and uniqueness results from Section 6.1.3 for general control-affine systems can be adapted to linear systems, and these results can be extended to account for linearity as in Proposition 5.3.2. One gets the following result upon doing this.

6.6.2 Theorem (Existence and uniqueness of controlled trajectories for linear continuous-time state space systems) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system, let $\mathbb{T}' \subseteq \mathbb{T}$ *be a sub-time-domain, let* $\mu\colon \mathbb{T}' \to \mathsf{U}$*, and make the following assumptions:*

(i) $\mathsf{A} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}'; \mathrm{L}(\mathsf{X}; \mathsf{X}))$*;*

(ii) $\mathrm{t} \mapsto \mathsf{B}(\mathrm{t})(\mu(\mathrm{t}))$ *is in* $\mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}'; \mathsf{X})$*.*

Then, for any $\mathrm{t}_0 \in \mathbb{T}'$ *and* $\mathrm{x}_0 \in \mathsf{X}$*, there exists a unique solution* $\xi \in \mathsf{AC}_{\mathrm{loc}}(\mathbb{T}'; \mathsf{X})$ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \mathsf{A}(\mathrm{t})(\xi(\mathrm{t})) + \mathsf{B}(\mathrm{t})(\mu(\mathrm{t})), \quad \xi(\mathrm{t}_0) \in \mathrm{x}_0;$$

thus $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma)$*.*

Proof This follows immediately from Proposition 5.3.2. ■

Let us make a few more or less immediate comments about controlled trajectories and controlled outputs.

6.6.3 Remarks (Controlled trajectories and controlled outputs for linear continuous-time state space systems)

1. From Corollary 5.3.3 we have an explicit formula for the controlled trajectory $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma)$ with the initial condition x_0 at t_0:

$$\Phi^{\Sigma}(t, t_0, x_0, \mu) = \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)(x_0) + \int_{t_0}^{t} \Phi^{\mathrm{c}}_{\mathsf{A}}(t, \tau) \circ \mathsf{B}(\tau)(\mu(\tau))\, \mathrm{d}\tau, \qquad t \in \mathrm{dom}(\mu). \quad (6.9)$$

The corresponding controlled output (η, μ) is, of course, given by

$$\eta(t) = \mathsf{C}(t) \circ \Phi^{\Sigma}(t, t_0, x_0, \mu) + \mathsf{D}(t)(\mu(t)), \qquad t \in \mathrm{dom}(\mu).$$

2. We note that, in contrast to general continuous-time state space systems, controlled trajectories always exist on the entire domain of definition of the control. This is one feature that makes working with linear systems less complicated than working with general systems.
3. The condition that $t \mapsto \mathsf{B}(t)(\mu(t))$ be locally integrable on $\mathbb{T}'$ can be generally satisfied in two common ways:
   (a) $\mathsf{B} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}'; \mathrm{L}(\mathsf{U}; \mathsf{X}))$ and $\mu \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}'; \mathsf{U})$;
   (b) $\mathsf{B} \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}'; \mathrm{L}(\mathsf{U}; \mathsf{X}))$ and $\mu \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}'; \mathsf{U})$.
   In both cases, local integrability of the product follows from Exercises III-3.8.8 and IV-1.4.4.
4. We can infer immediately from Proposition 6.1.13 and Theorems 6.1.14 and 6.1.18 the properties of the flow Φ^Σ for a linear continuous-time state space system. In particular, we have continuity of the flow with respect to initial condition, initial time, final time, and control. In fact, these conclusions follow most easily and directly from the formula (6.9). •

We note that the dynamics and the output mapping are linear functions of (x, u). That is to say, the mappings

$$\mathsf{X} \oplus \mathsf{U} \ni (x, u) \mapsto \mathsf{A}(t)(x) + \mathsf{B}(t)(u) \in \mathsf{X},$$
$$\mathsf{X} \oplus \mathsf{U} \ni (x, u) \mapsto \mathsf{C}(t)(x) + \mathsf{D}(t)(u) \in \mathsf{Y}$$

are linear for each $t \in \mathbb{T}$. Moreover, the flow is also linear in the sense of the following result.

6.6.4 Proposition (Linearity of flow for linear continuous-time state space systems) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system, let $\mathbb{T}' \subseteq \mathbb{T}$ be a sub-time-domain, and consider the following two cases:

(i) $\mathsf{B} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{X}))$ *and* $\mathscr{U} = \mathsf{L}^\infty_{\mathrm{loc}}((\mathbb{T}); \mathsf{U})$;

(ii) $\mathsf{B} \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{X}))$ *and* $\mathscr{U} = \mathsf{L}^1_{\mathrm{loc}}((\mathbb{T}); \mathsf{U})$.

Then, for each sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}$, the mapping

$$\mathsf{X} \oplus \mathscr{U}(\mathbb{T}') \ni (\mathrm{x}_0, \mu) \mapsto \Phi^\Sigma(\mathrm{t}, \mathrm{t}_0, \mathrm{x}_0, \mu) \in \mathsf{X}$$

is linear for each $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}'$.

Proof This follows immediately from the formula (6.9) for the flow of a linear continuous-time state space system. ■

Note that we ask that $\mathscr{U}$ be either $\mathsf{L}^\infty_{\mathrm{loc}}((\mathbb{T}); \mathsf{U})$ or $\mathsf{L}^1_{\mathrm{loc}}((\mathbb{T}); \mathsf{U})$, and not a subset of these spaces of partially defined spaces of signals. The reason for this is that we need for $\mathscr{U}(\mathbb{T}')$ to be a vector space in order for linearity to make sense. We could generalise this by requiring that $\mathscr{U}(\mathbb{T}')$ be a subspace of $\mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}'; \mathsf{U})$ or $\mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}'; \mathsf{U})$. This is a point of view we shall adopt in Section 6.7.

6.6.2 Systems with constant coefficients

Now we consider systems with the coefficient linear mappings for the dynamics and the output map are independent of time. There are some simplifications that arise in this case that are worth recording, so we devote this section to this class of system.

6.6.5 Definition (Linear continuous-time state space system with constant coefficients) A *linear continuous-time state space system with constant coefficients* is a nonuple

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

where

(i) X (the ***state space***), U (the ***input space***), and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathbb{T} \subseteq \mathbb{R}$ is a continuous time-domain,

(iii) $\mathsf{A} \in \mathrm{L}(\mathsf{X}; \mathsf{X})$, $\mathsf{B} \in \mathrm{L}(\mathsf{U}; \mathsf{X})$, $\mathsf{C} \in \mathrm{L}(\mathsf{X}; \mathsf{Y})$, and $\mathsf{D} \in \mathrm{L}(\mathsf{U}; \mathsf{Y})$, and

(iv) $\mathscr{U} \subseteq \mathsf{L}^1_{\mathrm{loc}}((\mathbb{T}); \mathsf{U})$. •

In this case, the system is an autonomous continuous-time state space system, and the dynamics and output map are defined, independent of time, as

$$\begin{aligned} f \colon \mathsf{X} \oplus \mathsf{U} &\to \mathsf{X} \\ (x, u) &\mapsto \mathsf{A}(x) + \mathsf{B}(u) \end{aligned}$$

and

$$\begin{aligned} h \colon \mathsf{X} \oplus \mathsf{U} &\to \mathsf{X} \\ (x, u) &\mapsto \mathsf{C}(x) + \mathsf{D}(u), \end{aligned}$$

respectively.

We note that a linear continuous-time state space system with constant coefficients is autonomous, and so is stationary, and strongly stationary if and only if $\mathsf{D} = 0$ (see Exercise 6.6.1). This stationarity is often reflected with some particular terminology.

6.6.6 Terminology What we call a linear continuous-time state space system with constant coefficients is often called a ***linear time-invariant system***, or an ***LTI system***, in short. We shall stick to the more cumbersome terminology in order to maintain internal consistency with other terminology elsewhere in this volume. We do not object, however, to a reader using the terminology "LTI system" in their private life.[7] •

[7] This is in contrast with our nonstandard notation $\mathscr{F}_{\mathrm{CD}}$, $\mathscr{F}_{\mathrm{CC}}$, $\mathscr{F}_{\mathrm{DC}}$, and $\mathscr{F}_{\mathrm{DD}}$, along with our nonstandard terminology, "continuous-discrete Fourier transform," "continuous-continuous Fourier transform," "continuous-discrete Fourier transform," and "discrete-discrete Fourier transform" used in Chapters IV-5, IV-6, and IV-7. While this terminology is *not* widespread, it is clearly superior to what is widespread, and we insist that the reader use and proliferate our nonstandard notation and language.

Note that a controlled trajectory (ξ, μ), with associated controlled output (η, μ), satisfies

$$\dot{\xi}(t) = \mathsf{A}(\xi(t)) + \mathsf{B}(\mu(t)),$$
$$\eta(t) = \mathsf{C}(\xi(t)) + \mathsf{D}(\mu(t)).$$

We have the following slight simplification of the existence and uniqueness theorem for systems with constant coefficients.

6.6.7 Theorem (Existence and uniqueness of controlled trajectories for linear continuous-time state space systems with constant coefficients) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients, let $\mathbb{T}' \subseteq \mathbb{T}$ be a sub-time-domain, and let $\mu \colon \mathbb{T}' \to \mathsf{U}$. Then, for any $\mathrm{t}_0 \in \mathbb{T}'$ and $\mathrm{x}_0 \in \mathsf{X}$, there exists a unique solution $\xi \in \mathsf{AC}_{\mathrm{loc}}(\mathbb{T}'; \mathsf{X})$ to the initial value problem

$$\dot{\xi}(\mathrm{t}) = \mathsf{A}(\xi(\mathrm{t})) + \mathsf{B}(\mu(\mathrm{t})), \quad \xi(\mathrm{t}_0) \in \mathrm{x}_0;$$

thus $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma)$.

Proof This follows immediately from Theorem 6.6.2 since, in this case, we have that $t \mapsto \mathsf{B}(\mu(t))$ is in $\mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}'; \mathsf{X})$ if $\mu \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}'; \mathsf{X})$, making use of Exercise IV-1.4.4. ■

We can simplify, for systems with constant coefficients, some of the discussion concerning flows and controlled outputs.

6.6.8 Remarks (Controlled trajectories and controlled outputs for linear continuous-time state space systems with constant coefficients)

1. From Theorem 5.3.8 we have an explicit formula for the controlled trajectory $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma)$ with the initial condition x_0 at t_0:

$$\Phi^{\Sigma}(t, t_0, x_0, \mu) = \mathrm{e}^{\mathsf{A}(t-t_0)}(x_0) + \int_{t_0}^{t} \mathrm{e}^{\mathsf{A}(t-\tau)}(\mathsf{B}(\mu(\tau)))\, \mathrm{d}\tau, \qquad t \in \mathrm{dom}(\mu). \tag{6.10}$$

   The corresponding controlled output (η, μ) is, of course, given by

$$\eta(t) = \mathsf{C}(t) \circ \Phi^{\Sigma}(t, t_0, x_0, \mu) + \mathsf{D}(t)(\mu(t)), \qquad t \in \mathrm{dom}(\mu).$$

2. We can infer immediately from Proposition 6.1.13 and Theorem 6.1.18 the properties of the flow Φ^{Σ} for a linear continuous-time state space system with constant coefficients. In particular, we have continuity of the flow with respect to initial condition, initial time, final time, and control. In fact, these conclusions follow most easily and directly from the formula (6.10). •

The situation concerning linearity is similar for systems with constant coefficients to systems with time-varying coefficients. First we note that the dynamics and the output mapping are linear functions of (x, u). That is to say, the mappings

$$\mathsf{X} \oplus \mathsf{U} \ni (x, u) \mapsto \mathsf{A}(x) + \mathsf{B}(u) \in \mathsf{X},$$
$$\mathsf{X} \oplus \mathsf{U} \ni (x, u) \mapsto \mathsf{C}(x) + \mathsf{D}(u) \in \mathsf{Y}$$

are linear. Moreover, the flow is also linear in the sense of the following result.

6.6.9 Proposition (Linearity of flow for linear continuous-time state space systems with constant coefficients) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients, and let $\mathbb{T}' \subseteq \mathbb{T}$ be a sub-time-domain. Then, for each sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}$, the mapping

$$\mathsf{X} \oplus \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}'; \mathsf{U}) \ni (\mathrm{x}_0, \mu) \mapsto \Phi^{\Sigma}(\mathrm{t}, \mathrm{t}_0, \mathrm{x}_0, \mu) \in \mathsf{X}$$

is linear for each $\mathrm{t}, \mathrm{t}_0 \in \mathbb{T}'$.

Proof This follows immediately from the formula (6.10) for the flow of a linear continuous-time state space system. ■

6.6.3 The impulse transmission map and the impulse response

In this section we introduce an important player in the theory of linear systems, both in continuous- and discrete-time, and both in the time-varying and constant coefficient cases. Here we work with the continuous-time case, and we refer the reader to Section 5.3.3 for the required background on ordinary differential equations with distributions as right-hand side.

6.6.3.1 The time-varying case

We begin with the definition.

6.6.10 Definition (Impulse transmission map for linear continuous-time state space systems) Let

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system.

(i) The ***proper impulse transmission map*** for Σ is the function

$$\mathrm{pitm}_{\Sigma} \colon \mathbb{T} \times \mathbb{T} \to \mathsf{L}(\mathsf{U}; \mathsf{Y})$$

defined by

$$\mathrm{pitm}_{\Sigma}(t, \tau) = \mathbf{1}_{\geq 0}(t - \tau)\mathsf{C}(t) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}(t, \tau) \circ \mathsf{B}(\tau).$$

Now suppose that $\mathbb{T} = \mathbb{R}$ and that $\mathsf{D} \in \mathsf{C}^0(\mathbb{R}; \mathsf{L}(\mathsf{U}; \mathsf{Y}))$.

(ii) The ***impulse transmission map*** for Σ at $t_0 \in \mathbb{R}$ is the distribution $\mathrm{itm}_\Sigma \in \mathscr{D}'(\mathbb{R} \times \mathbb{R}; \mathsf{L}(\mathsf{U}; \mathsf{Y}))$ given by

$$\mathrm{itm}_{\Sigma,t_0} = \theta_{\mathrm{pitm}_{\Sigma,t_0}} + \mathsf{D}(t_0) \otimes (\tau_{t_0}^* \delta),$$

where $\mathrm{pitm}_{\Sigma,t_0}(t) = \mathrm{pitm}_\Sigma(t, t_0)$. •

In Theorem 5.3.11 we showed how the continuous-time state transition map arose as the solution to a distributional differential equation. Here we use this interpretation to arrive at a distributional interpretation for the impulse transmission map.

6.6.11 Theorem (A distributional interpretation of the proper impulse transmission map) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with $\mathbb{T} = \mathbb{R}$ *and let* $\mathsf{t}_0 \in \mathbb{R}$. *Then, for* $\mathsf{u} \in \mathsf{U}$,

$$\mathrm{pitm}_\Sigma(\mathsf{t}, \mathsf{t}_0)(\mathsf{u}) = \mathsf{C}(\mathsf{t}) \circ \xi_{\mathsf{t}_0}(\mathsf{t}),$$

where $\xi_{\mathsf{t}_0} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{X})$ *is the locally integrable function associated to a solution of the distributional differential equation*

$$\theta_{\mathsf{t}_0}^{(1)} = \mathsf{A} \circ \theta_{\mathsf{t}_0} + \mathsf{B}(\mathsf{t}_0)(\mathsf{u} \otimes (\tau_{\mathsf{t}_0}^* \delta)),$$

where $\mathsf{A} \circ \theta_{\mathsf{t}_0}$ *is as given in Remark IV-3.2.53–3 and where* $\mathsf{B}(\mathsf{t}_0)(\mathsf{u} \otimes (\tau_{\mathsf{t}_0}^* \delta))$ *is as given in Remark IV-3.2.53–1.*

Proof As in Theorem 5.3.11, let $\Theta_{t_0} \in \mathscr{D}'(\mathbb{R}; \mathsf{L}(\mathsf{X}; \mathsf{X}))$ be the regular distribution associated to the function

$$t \mapsto \Xi_{t_0}(t) \triangleq \mathsf{1}_{\geq 0}(t - t_0)\Phi_{\mathsf{A}}^{\mathrm{c}}(t, t_0).$$

Let $\theta_{t_0} = \Theta_{t_0}(\mathsf{B}(t_0)(u))$. We have

$$\xi_{t_0}(t) = \mathsf{1}_{\geq 0}(t - t_0)\Phi_{\mathsf{A}}^{\mathrm{c}}(t, t_0) \circ \mathsf{B}(t_0)(u) = \Xi_{t_0}(t)(\mathsf{B}(t_0)(u)).$$

Therefore,

$$\theta_{t_0} = \Theta_{t_0}(\mathsf{B}(t_0)(u)) = \theta_{\Xi_{t_0}}(\mathsf{B}(t_0)(u)) = \theta_{\Xi_{t_0}(\mathsf{B}(t_0)(u))} = \theta_{\xi_{t_0}},$$

and so θ_{t_0} is the distribution associated with the locally integrable function ξ_{t_0}.

By Theorem 5.3.11, Θ_{t_0} is a solution to the distributional differential equation

$$\Theta_{t_0}^{(1)} = \mathsf{A} \circ \Theta_{t_0} + \mathrm{id}_{\mathsf{X}} \otimes (\tau_{t_0}^* \delta).$$

Therefore, by evaluating both sides of this distributional differential equation at $\mathsf{B}(t_0)(u)$, θ_{t_0} is a solution to the distributional differential equation

$$\theta_{t_0} = \mathsf{A} \circ \Theta_{t_0}(\mathsf{B}(t_0)(u)) + (\mathrm{id}_{\mathsf{X}} \otimes (\tau_{t_0}^* \delta))(\mathsf{B}(t_0)(u)) = \mathsf{A} \circ \theta_{t_0} + \mathsf{B}(t_0)(u \otimes (\tau_{t_0}^* \delta)).$$

Thus ξ_{t_0} is the locally integrable function associated to the solution θ_{t_0} as above. It then follows directly that pitm_Σ is as claimed in the statement of the theorem. ■

Let us see how we can interpret the (non-proper) impulse transmission map in a manner analogous to the preceding theorem. As in the proof of the theorem, we again let θ_{t_0} be the given solution to the distributional differential equation

$$\theta^{(1)}_{t_0} = \mathsf{A} \circ \theta_{t_0} + \mathsf{B}(t_0)(u \otimes (\tau^*_{t_0}\delta)).$$

Thus the "input" to this system is, not a locally integrable control, but the distribution $u \otimes (\tau^*_{t_0}\delta)$, i.e., a "pulse" of the control u applied at time t_0. We then have

$$\begin{aligned}\langle \mathsf{C} \circ \theta_{t_0} + \mathsf{D}(u \otimes (\tau^*_{t_0}\delta)); \phi\rangle &= \int_{\mathbb{R}} 1_{\geq 0}(t - t_0)\phi(t)\mathsf{C}(t) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}(t, \tau) \circ \mathsf{B}(\tau)(u)\, d\tau \\ &\quad + \mathsf{D}(u)(\tau^*_{t_0}\delta)(\phi) \\ &= \theta_{\mathrm{pitm}_{\Sigma,t_0}(u)}(\phi) + \mathsf{D}(t_0)(u)\phi(0) \\ &= \langle \mathrm{itm}_{\Sigma,t_0}(u); \phi\rangle.\end{aligned}$$

Note that the product of D with $u \otimes (\tau^*_{t_0}\delta)$ in the first line and the product of $\mathsf{D}(u)$ with $\tau^*_{t_0}\delta$ is to be thought of in terms of Corollary IV-3.7.28. In this (notationally complicated, but conceptually not difficult) manner that one can regard the distribution $\mathrm{itm}_{\Sigma,t_0}(u)$ as the "output" associated with the "input" $u \otimes (\tau^*_{t_0}\delta)$.

The next result follows immediately from the definition of the impulse transmission map and the formula (6.9).

6.6.12 Proposition (Using the impulse transmission map to determine outputs) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system and let $\mu \in \mathscr{U}$. Then the output corresponding to an initial condition x_0 at time t_0 is

$$\eta(\mathrm{t}) = \mathsf{C}(\mathrm{t}) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}(\mathrm{t}, \mathrm{t}_0)(\mathrm{x}_0) + \int_{\mathrm{t}_0}^{\mathrm{t}} \mathrm{pitm}_{\Sigma}(\mathrm{t}, \tau)\mu(\tau)\, \mathrm{d}\tau + \mathsf{D}(\mathrm{t})(\mu(\mathrm{t})), \qquad \mathrm{t} \in \mathrm{dom}(\mu)_{\geq \mathrm{t}_0}.$$

The punchline of the result is that the output is a linear combination of three terms:

$$\underbrace{\mathsf{C}(t) \circ \Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)(x_0)}_{\text{term 1}} + \underbrace{\int_{t_0}^{t} \mathrm{pitm}_{\Sigma}(t, \tau)\mu(\tau)\, \mathrm{d}\tau}_{\text{term 2}} + \underbrace{\mathsf{D}(t)(\mu(t))}_{\text{term 3}}. \tag{6.11}$$

Let us describe these terms, intuitively.

1. The first term is the contribution from a nonzero initial provides the contribution to the output from the nonzero initial condition x_0 at time t_0.
2. The second term has the most complex interpretation. First of all, by Theorem 6.6.11, the integrand $\mathrm{pitm}_{\Sigma}(t, \tau)\mu(\tau)$ is the output obtained from the input $\mu(\tau) \otimes (\tau^*_{\tau}\delta)$, i.e., an impulse of $\mu(\tau)$ at time τ. The second terms can then be thought of as the "sum" of these contributions as τ goes from t_0 to t.

3. The third term simply arises from the direct transmission from input to output determined by D. This can be thought of as the memoryless part of the system; see Example 2.2.31–2.

6.6.3.2 The constant coefficient case The preceding constructions simplify substantially in the case of constant coefficient systems. This is fortunate, since it is this case that we will examine with respect to transform methods in Chapters 7 and 8. Let us record the simplifications.

The definition we make is the following. For systems with constant coefficients, there is no reason to not take the time-domain to be $\mathbb{R}$, and so we do so.

6.6.13 Definition (Impulse response) Let

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{R}$.

(i) The ***proper impulse response*** for Σ is the function

$$\begin{aligned} \mathrm{pir}_\Sigma \colon \mathbb{R} &\to \mathrm{L}(\mathsf{U}; \mathsf{Y}) \\ t &\mapsto 1_{\geq 0}(t) \mathsf{C} \circ \mathrm{e}^{\mathsf{A}t} \circ \mathsf{B}. \end{aligned}$$

(ii) The ***impulse response*** for Σ is the distribution $\mathrm{ir}_\Sigma \in \mathscr{D}'(\mathbb{R}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ given by

$$\mathrm{ir}_\Sigma = \theta_{\mathrm{pir}_\Sigma} + \mathsf{D} \otimes \delta. \qquad \bullet$$

The connection between the impulse response and the impulse transmission map is given by the following result.

6.6.14 Proposition (The impulse response and the impulse transmission map) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{R}$. *Then the proper impulse transmission map is given by*

$$\mathrm{pitm}_\Sigma(\mathrm{t}, \tau) = \mathrm{pir}_\Sigma(\mathrm{t} - \tau), \qquad \mathrm{t}, \tau \in \mathbb{R},\ \mathrm{t} \geq \tau.$$

Proof This follows from the definitions, and the fact that, for A being independent of time, we have $\Phi^{\mathrm{c}}_{\mathsf{A}}(t, \tau) = \mathrm{e}^{\mathsf{A}(t-\tau)}$ by definition. ■

Thus everything we said about the impulse transmission map above can be translated into a statement about the impulse response in the constant coefficient case. However, since there is more that can be said, and what can be said can be said more simply, let us record these translations explicitly.

We begin by giving a distributional interpretation of the impulse response. We point out that we have an additional uniqueness assertion in the following result, represented by replacing occurrences of "a" in Theorem 6.6.11 with "the unique" in the next result.

6.6.15 Theorem (A distributional interpretation of the proper impulse transmission map) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{R}$. *Then*

$$\mathrm{pir}_\Sigma(t) = \mathsf{C} \circ \xi_0(t),$$

where $\xi_0 \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathsf{X})$ *is the locally integrable function associated to the unique solution of the distributional differential equation*

$$\theta_0^{(1)} = \mathsf{A}(\theta_0) + \mathsf{B}(\mathsf{u} \otimes \delta),$$

where $\mathsf{A}(\theta_0)$ *and* $\mathsf{B}(\mathsf{u} \otimes \delta)$ *are as given in Remark IV-3.2.53–1.*

Proof The characterisation of pir_Σ follows from Theorem 6.6.15 with $t_0 = 0$, noting that, when A is independent of time and if ir_ξ is the regular distribution associated to a locally integrable signal ξ,

$$\langle \mathsf{A} \circ \theta_\xi; \phi \rangle = \int_{\mathbb{R}} \phi(t) \mathsf{A} \circ \xi(t) \, \mathrm{d}t = \mathsf{A}(\langle \theta_\xi; \phi \rangle) = \langle \mathsf{A}(\theta_\xi); \phi \rangle$$

for $\phi \in \mathscr{D}(\mathbb{R}; \mathbb{R})$.

As for the uniqueness assertion of the theorem, we note that $\mathsf{B}(u \otimes \delta) \in \mathscr{D}'_+(\mathbb{R}; \mathsf{X})$ and $\theta_0 \in \mathscr{D}'_+(\mathbb{R}; \mathsf{X})$. Thus uniqueness follows from Theorem 5.3.13. ■

We note that the proof of Theorem 5.3.13 relies on properties of the convolution algebra $\mathscr{D}'_+(\mathbb{R}; \mathbb{R})$. We shall have more to say about the rôle of convolution in the theory of continuous-time linear systems in Section 6.7.4.

The theorem tells us that, when $\mathsf{D} = 0$, the proper impulse response if the "output" corresponding to an "input" $u \otimes \delta$, i.e., a pulse of u at time 0. Similarly to the case of the impulse transmission map, it also holds that the (non-proper) impulse response is the output for this same input, when D is not necessarily zero.

Of course, just as in Proposition 6.6.12, we can use the impulse response to characterise outputs for linear continuous-time state space systems with constant coefficients.

6.6.16 Proposition (Using the impulse response to determine outputs) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{R}$ *and let* $\mu \in \mathscr{U}$. *Then the output corresponding to the initial condition* x_0 *at time* t_0 *is*

$$\eta(\mathsf{t}) = \mathsf{C} \circ \mathrm{e}^{\mathsf{A}(\mathsf{t}-\mathsf{t}_0)}(\mathsf{x}_0) + \int_{\mathsf{t}_0}^{\mathsf{t}} \mathrm{pir}_\Sigma(\mathsf{t} - \tau)\mu(\tau) \, \mathrm{d}\tau + \mathsf{D} \circ \mu(\mathsf{t}), \qquad \mathsf{t} \in \mathrm{dom}(\mu)_{\geq \mathsf{t}_0}.$$

Proof This is a direct translation of Proposition 6.6.12 to the constant coefficient case. ■

The interpretation we give for the three terms in the output are the same as we gave after the statement of Proposition 6.6.12. However, in the constant coefficient case, the time $t_0 = 0$ is distinguished, and this gives rise to distinguishing this particular case.

6.6.17 Definition (Zero-state/zero-time response) Let

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{R}$ and let $\mu \in \mathscr{U}$ with $\operatorname{supp}(\mu) = \mathbb{R}_{\geq 0}$. The ***zero-state/zero-time response*** to the input μ is

$$\begin{aligned} \zeta_\mu \colon \mathbb{R} &\to \mathsf{Y} \\ t &\mapsto \int_0^t 1_{\geq 0}(t)\mathrm{pir}_\Sigma(t-\tau)\mu(\tau)\,\mathrm{d}\tau. \end{aligned}$$

•

These sorts of considerations will be explored in detail in Section 6.7.6.2.

Exercises

6.6.1 Consider a linear continuous-time state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}).$$

Answer the following questions.

(a) Show that Σ defines a general input/output system as per Definition 2.1.3. Identify the components of the general input/output system.
(b) Show that Σ is a general time system as per Definition 2.2.9. Identify the components of the general time system.
(c) Show that, as a general time system, Σ is output complete.
(d) Show that, as a general time system, Σ is complete.
(e) Show that Σ is causal and is strongly causal if and only if $\mathsf{D} = 0$.
(f) Show that Σ has a dynamical systems representation as per Definition 2.2.19. Identify components of the dynamical systems representation.
(g) Show that Σ has a state space representation as per Definition 2.2.24. Identify components of the state space representation.

6.6.2 For the following linear continuous-time state space systems

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}),$$

compute the proper impulse transmission map (or proper impulse response, as appropriate) and the impulse transmission map (or impulse response, as appropriate).

(a) Take
(i) $\mathsf{X} = \mathbb{R}$,
(ii) $\mathsf{U} = \mathbb{R}$,
(iii) $\mathsf{Y} = \mathbb{R}$,
(iv) $\mathbb{T} = \mathbb{R}$,
(v) $\mathscr{U} = \mathsf{L}^1_{\text{loc}}(\mathbb{R}; \mathbb{R})$,
(vi) $\boldsymbol{A} = \begin{bmatrix} 0 \end{bmatrix}$,
(vii) $\boldsymbol{B} = \begin{bmatrix} 1 \end{bmatrix}$,
(viii) $\boldsymbol{C} = \begin{bmatrix} 1 \end{bmatrix}$,
(ix) $\boldsymbol{D} = \begin{bmatrix} 1 \end{bmatrix}$,

(b) Take
(i) $\mathsf{X} = \mathbb{R}^2$,
(ii) $\mathsf{U} = \mathbb{R}$,
(iii) $\mathsf{Y} = \mathbb{R}^2$,
(iv) $\mathbb{T} = \mathbb{R}$,
(v) $\mathscr{U} = \mathsf{L}^1_{\text{loc}}(\mathbb{R}; \mathbb{R})$,
(vi) $\boldsymbol{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$,
(vii) $\boldsymbol{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,
(viii) $\boldsymbol{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$,
(ix) $\boldsymbol{D} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,

(c) Take
(i) $\mathsf{X} = \mathbb{R}^3$,
(ii) $\mathsf{U} = \mathbb{R}^2$,
(iii) $\mathsf{Y} = \mathbb{R}^2$,
(iv) $\mathbb{T} = \mathbb{R}$,
(v) $\mathscr{U} = \mathsf{L}^1_{\text{loc}}(\mathbb{R}; \mathbb{R})$,
(vi) $\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$,
(vii) $\boldsymbol{B} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$,
(viii) $\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & -2 & 0 \end{bmatrix}$,
(ix) $\boldsymbol{D} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$,

(d) Take
(i) $\mathsf{X} = \mathbb{R}^2$,
(ii) $\mathsf{U} = \mathbb{R}$,
(iii) $\mathsf{Y} = \mathbb{R}$,
(iv) $\mathbb{T} = \mathbb{R}_{>0}$,
(v) $\mathscr{U} = \mathsf{L}^1_{\text{loc}}(\mathbb{R}_{>0}; \mathbb{R})$,
(vi) $\boldsymbol{A} = \begin{bmatrix} t^{-1} & -1 \\ t^{-2} & 2t^{-1} \end{bmatrix}$,
(vii) $\boldsymbol{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$,
(viii) $\boldsymbol{C} = \begin{bmatrix} 1 & 0 \end{bmatrix}$,
(ix) $\boldsymbol{D} = \begin{bmatrix} 0 \end{bmatrix}$,

Hint: *Refer to Example 5.2.10.*

6.6.3 Consider the following differential equation for functions $\eta, \mu \colon \mathbb{R} \to \mathbb{F}$:

$$\frac{\mathrm{d}^n \eta}{\mathrm{d}t^n}(t) + p_{n-1} \frac{\mathrm{d}^{n-1} \eta}{\mathrm{d}t^{n-1}}(t) + \cdots + p_1 \frac{\mathrm{d}\eta}{\mathrm{d}t}(t) + p_0 \eta(t) = c_{n-1} \frac{\mathrm{d}^{n-1} \mu}{\mathrm{d}t^{n-1}}(t) + c_{n-2} \frac{\mathrm{d}^{n-2} \mu}{\mathrm{d}t^{n-2}}(t) + \cdots + c_1 \frac{\mathrm{d}\mu}{\mathrm{d}t}(t) + c_0 \mu(t).$$

Answer the following questions.

(a) Show that this determines a general time system as per Definition 2.2.9. Clearly identify the spaces of input and output signals.

(b) Argue that a natural choice of states for this system is

$$\xi_j(t) = \frac{\mathrm{d}^j \eta}{\mathrm{d}t^j}(t), \qquad j \in \{0, 1, \ldots, n\}.$$

(c) Argue that a somewhat less natural, but still valid, choice of states is $\xi_n(t) = \eta(t)$ and

$$\xi_{n-j}(t) = \sum_{k=0}^{j} p_{n-j+k} \frac{\mathrm{d}^k \eta}{\mathrm{d}t^k}(t) - \sum_{k=0}^{j-1} c_{n-j+k} \frac{\mathrm{d}^k \mu}{\mathrm{d}t^k}(t), \qquad j \in \{1, \ldots, n-1\}.$$

(d) Derive a linear continuous-time state space system with constant coefficients for which the input/output relation is the same as the general time system from part (a) and for which the states are as in part (c).

6.6.4 Consider the mass-spring system in Figure 6.7 with identical masses m and

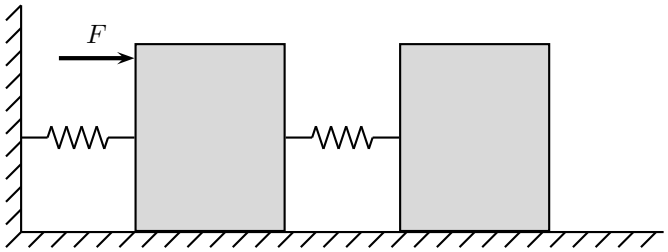


Figure 6.7 Mass-spring system

linear springs with spring constant k. Suppose the leftmost mass is subject to a force F. Answer the following questions.

(a) Write the equations of motion in the form of a linear continuous-time state space system, clearly identifying the nine system components.

(b) Is the system one with constant coefficients?

(c) Find the eigenvalues and eigenvectors of the system linear map A. Give a physical interpretation of both.

6.6.5 For the circuit shown in Figure 6.8, answer the following questions.

(a) Write the governing equations in the form of a linear continuous-time state space system, clearly identifying the nine system components.

(b) Is the system one with constant coefficients?

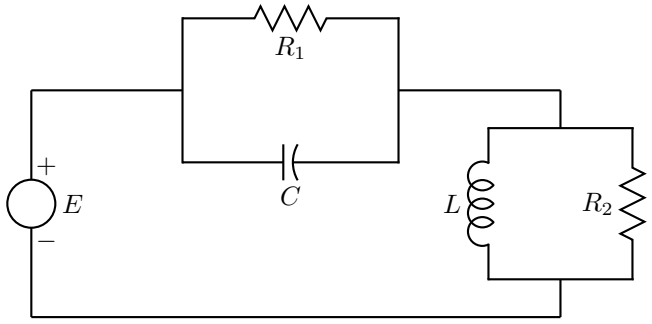


Figure 6.8 A linear circuit

Section 6.7

Linear continuous-time input/output systems

We now undertake the systematic development of a class of input/output systems having the property of linearity. We begin with a general construction of what attributes such systems should have (basically, they should be input/output systems that are linear). We then focus on particular classes of input/output systems, those defined by integration in some way. These systems capture, as special cases, the input/output behaviour of linear state space systems. We close our presentation in this section by proving this relationship.

Do I need to read this section? As with the preceding section, the material in this section is to be regarded as a core part of the material in this volume. •

6.7.1 General definitions

We being by considering a general setting for linear continuous-time input/output systems. The essential definition, which follows, is basically the Definition 6.2.3 for continuous-time input/output systems, with the addition of linearity. This requires linearity for both the spaces of input and output signals, and of the system mappings.

6.7.1 Definition (Linear continuous-time input/output system) A *linear continuous-time input/output system* is a sextuple $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$, where

(i) U (the ***input space***) and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathbb{T} \subseteq \mathbb{R}$ is an interval (the ***time-domain***),

(iii) $\mathscr{U} \subseteq \mathsf{U}^{(\mathbb{T})}$ is a space of partially defined signals with topology (the ***input signals***) such that, for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, $\mathscr{U}(\mathbb{S})$ is a subspace of $\mathsf{U}^{\mathbb{S}}$,

(iv) $\mathscr{Y} \subseteq \mathsf{Y}^{(\mathbb{T})}$ is a space of partially defined signals with topology (the ***output signals***) such that, for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, $\mathscr{Y}(\mathbb{S})$ is a subspace of $\mathsf{Y}^{\mathbb{S}}$, and

(v) $g \colon \mathscr{U} \to \mathscr{Y}$ has the following properties:

  (a) for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, the restriction of g to $\mathscr{U}(\mathbb{S})$, denoted by $g_{\mathbb{S}}$, takes values in $\mathscr{Y}(\mathbb{S})$;

  (b) if $\mathbb{S}, \mathbb{S}' \subseteq \mathbb{T}$ are sub-time-domains with $\mathbb{S}' \subseteq \mathbb{S}$, then $g_{\mathbb{S}}|\mathscr{U}(\mathbb{S}') = g_{\mathbb{S}'}$;

  (c) $g_{\mathbb{S}}$ is linear and continuous for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$.

Moreover,

(vi) a pair (μ, η) with $\mu \in \mathscr{U}(\mathbb{S})$ and $\eta = g_{\mathbb{S}}(\mu)$ is a ***behaviour*** for Σ, and we denote by $\mathscr{B}(\Sigma)$ the set of behaviours. •

Of course, linear continuous-time input/output systems are continuous-time input/output systems, accepting the mild generalisation from using general finite-dimensional vector spaces in place of Euclidean spaces. Thus all of the comments made in Section 6.2.2 about the connections between continuous-time input/output systems and the general classes of systems from Chapter 2 are applicable to linear continuous-time input/output systems. In addition, linear continuous-time input/output systems are linear time systems as per Definition 2.2.43. Note that, due to the fact that we work with spaces of input and output signals that are comprised of partially defined signals, it is not generally the case that a linear continuous-time input/output system is a linear general input/output system as per Definition 2.1.12. However, were we to restrict to the case of signals only defined on the entire time-domain $\mathbb{T}$, i.e., to not allow partially defined signals, then such a linear continuous-time input/output system would indeed be a linear general input/output system as per Definition 2.1.12.

6.7.2 Integral kernel systems

We now consider a special class of linear continuous-time input/output systems. As we shall assert precisely in Section 6.7.6, the class of systems we consider generalise aspects of the input/output behaviour of a linear continuous-time state space system.

The initial ingredient to the constructions we make is contained in the following definition.

6.7.2 Definition (Integral kernel, integral operator) Let $\mathbb{T} \subseteq \mathbb{R}$ be a continuous time-domain, and let U and Y be finite-dimensional $\mathbb{R}$-vector spaces.

(i) An ***integral kernel*** from U to Y on $\mathbb{T}$ is a mapping

$$\mathsf{K} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y}).$$

For $t, \tau \in \mathbb{T}$, we shall denote

$$\mathsf{K}_t \colon \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y}) \qquad \mathsf{K}^\tau \colon \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$$
$$\tau \mapsto \mathsf{K}(t, \tau) \qquad t \mapsto \mathsf{K}(t, \tau).$$

Let $\mathscr{U} \subseteq \mathsf{U}^{\mathbb{T}}$ be a subspace.

(ii) An integral kernel K is ***compatible*** with $\mathscr{U}$ if, for every $\mu \in \mathscr{U}$ and for almost every $t \in \mathbb{T}$, $\mathsf{K}_t(\mu) \in \mathrm{L}^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$.

(iii) If K is compatible with $\mathscr{U}$, the ***integral operator*** defined by K is the mapping

$$g_{\mathsf{K}} \colon \mathscr{U} \to \mathsf{Y}^{\mathbb{T}}$$

defined by

$$g_{\mathsf{K}}(\mu)(t) = \int_{\mathbb{T}} \mathsf{K}(t,\tau)(\mu(\tau))\,\mathrm{d}\tau, \qquad \text{a.e. } t \in \mathbb{T}. \qquad \bullet$$

As yet, we do not have the structure of a continuous-time input/output system, since the domain and codomain of g_{K} do not have useful structure (other than their vector space structure). We need to provide conditions on K that ensure that the integral operator g_{K} takes signals from a nice domain into a nice codomain. The following definition captures the properties we want.

6.7.3 Definition (Integral kernel system) An ***integral kernel system*** is a sextuple $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathsf{K})$ where

(i) U (the ***input space***) and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathbb{T} \subseteq \mathbb{R}$ is a continuous time-domain (the ***time-domain***),

(iii) $\mathscr{U} \subseteq \mathsf{U}^{\mathbb{T}}$ is a subspace of signals with topology (the ***input signals***),

(iv) $\mathscr{Y} \subseteq \mathsf{Y}^{\mathbb{T}}$ is a subspace of signals with topology (the ***output signals***),

(v) K is an integral kernel compatible with $\mathscr{U}$, and

(vi) the integral operator g_{K} is a continuous linear mapping from $\mathscr{U}$ to $\mathscr{Y}$. $\bullet$

Of course, the definition gives us no insight into which K's, $\mathscr{U}$'s, and $\mathscr{Y}$'s might possibly comprise a continuous-time kernel system. In order to obtain characterisations which give such systems, we have to prove something, and the following result gives some cases that work.

6.7.4 Theorem (Some integral kernel systems) *Let $\mathbb{T} \subseteq \mathbb{R}$ be a continuous time-domain, let U and Y be finite-dimensional $\mathbb{R}$-vector spaces, and let $\mathrm{p} \in [1, \infty]$. In the following cases, the integral kernel $\mathsf{K}\colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$, the input space $\mathscr{U}$, and the output space $\mathscr{Y}$ are such that*

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathsf{K})$$

is an integral kernel system:

(i) (a) $\mathsf{K}_{\mathrm{t}} \in \mathrm{L}^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_1$ *is in* $\mathrm{L}^\infty(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,
(b) $\mathscr{U} \subseteq \mathrm{L}^\infty(\mathbb{T}; \mathsf{U})$, *and*
(c) $\mathscr{Y} = \mathrm{L}^\infty(\mathbb{T}; \mathsf{Y})$;

(ii) (a) $\mathsf{K}_{\mathrm{t}} \in \mathrm{L}^\infty(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_\infty$ *is in* $\mathrm{L}^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,
(b) $\mathscr{U} \subseteq \mathrm{L}^1(\mathbb{T}; \mathsf{U})$, *and*
(c) $\mathscr{Y} = \mathrm{L}^1(\mathbb{T}; \mathsf{Y})$;

(iii) (a) $\mathsf{K}_{\mathrm{t}} \in \mathrm{L}^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_1$ *is in* $\mathrm{L}^\infty(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,
(b) $\mathsf{K}_{\mathrm{t}} \in \mathrm{L}^\infty(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_\infty$ *is in* $\mathrm{L}^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,
(c) $\mathscr{U} \subseteq \mathrm{L}^{\mathrm{p}}(\mathbb{T}; \mathsf{U})$, *and*
(d) $\mathscr{Y} = \mathrm{L}^{\mathrm{p}}(\mathbb{T}; \mathsf{Y})$.

Proof (i) First of all, for $t \in \mathbb{T}$,

$$\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)\mu(\tau)\| \, \mathrm{d}\tau \leq \|\mu\|_\infty \int_{\mathbb{T}} \|\mathsf{K}_t\| \, \mathrm{d}\tau < \infty,$$

giving the compatibility of K with $\mathscr{U}$ in this case. Also by Exercise III-3.8.8,

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)\|_\infty &= \operatorname{ess\,sup}\left\{ \left\| \int_{\mathbb{T}} \mathsf{K}(t,\tau)(\mu(\tau)) \, \mathrm{d}\tau \right\| \;\middle|\; t \in \mathbb{T} \right\} \\
&\leq \operatorname{ess\,sup}\left\{ \int_{\mathbb{T}} \left\| \mathsf{K}(t,\tau)(\mu(\tau)) \right\| \, \mathrm{d}\tau \;\middle|\; t \in \mathbb{T} \right\} \\
&\leq \operatorname{ess\,sup}\left\{ \|\mu\|_\infty \int_{\mathbb{T}} \|\mathsf{K}(t,\tau)\| \, \mathrm{d}\tau \;\middle|\; t \in \mathbb{T} \right\} \\
&\leq \underbrace{\operatorname{ess\,sup}\{\|\mathsf{K}_t\|_1 \mid t \in \mathbb{T}\}}_{C_\infty} \|\mu\|_\infty.
\end{aligned}$$

Thus $\|g_{\mathsf{K}}(\mu)\|_\infty \leq C_\infty \|\mu\|_\infty$, giving continuity of g_{K} by Theorem III-3.5.8.

(ii) We first have, for $t \in \mathbb{T}$,

$$\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)\mu(\tau)\| \, \mathrm{d}\tau \leq \|\mathsf{K}_t\|_\infty \int_{\mathbb{T}} \|\mu(\tau)\| \, \mathrm{d}\tau < \infty,$$

giving the compatibility of K with $\mathscr{U}$. We also have

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)\|_1 &= \int_{\mathbb{T}} \left\| \int_{\mathbb{T}} \mathsf{K}(t,\tau)(\mu(\tau)) \, \mathrm{d}\tau \right\| \mathrm{d}t \\
&\leq \int_{\mathbb{T}} \left(\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)(\mu(\tau))\| \, \mathrm{d}\tau \right) \mathrm{d}t \\
&= \int_{\mathbb{T}} \left(\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)(\mu(\tau))\| \, \mathrm{d}t \right) \mathrm{d}\tau \\
&\leq \int_{\mathbb{T}} \left(\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)\| \, \mathrm{d}t \right) \|\mu(\tau)\| \, \mathrm{d}\tau \\
&\leq \underbrace{\left(\int_{\mathbb{T}} \|\mathsf{K}_t\|_\infty \, \mathrm{d}t \right)}_{C_1} \left(\int_{\mathbb{T}} \|\mu(\tau)\| \, \mathrm{d}\tau \right)
\end{aligned}$$

using Fubini's Theorem. Thus $\|g_{\mathsf{K}}(\mu)\|_1 \leq C_1 \|\mu\|_1$, and we get this part of the theorem by Theorem III-3.5.8.

(iii) Clearly we can restrict ourselves to $p \in (1,\infty)$. Thus we take $p' \in (1,\infty)$ to be the conjugate index for which $\frac{1}{p} + \frac{1}{p'} = 1$. Let C_∞ and C_1 be as defined in the first two parts of the proof.

To determine the compatibility of K with μ, for $\mu \in \mathsf{L}^p(\mathbb{T};\mathsf{U})$, write

$$\mu_0(t) = \begin{cases} \mu(t), & \|\mu(t)\| \leq 1, \\ 0, \text{otherwise}, \end{cases}$$

and $\mu_1 = \mu - \mu_0$. Note that $\mu_0 \in \mathsf{L}^\infty(\mathbb{T}; \mathsf{U})$ and $\mu_1 \in \mathsf{L}^p(\mathbb{T}; \mathsf{U})$. Moreover,

$$\|\mu_1(t)\| \leq \|\mu_1(t)\|^p, \qquad t \in \mathbb{T},$$

and so $\mu_1 \in \mathsf{L}^1(\mathbb{T}; \mathsf{U})$. One can then combine the compatibility conclusions from the first two parts of the proof to conclude that K is compatible with $\mathsf{L}^p(\mathbb{T}; \mathsf{U})$.

We now compute, using Hölder's inequality in the form of Lemma III-3.8.54,

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)(t)\| &\leq \int_{\mathbb{T}} \|\mathsf{K}(t,\tau)(\mu(\tau))\| \, \mathrm{d}\tau \\
&\leq \int_{\mathbb{T}} (\|\mathsf{K}(t,\tau)\|^{1/p} \|\mu(\tau)\|) \|\mathsf{K}(t,\tau)\|^{1/p'} \, \mathrm{d}\tau \\
&\leq \left(\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)\| \|\mu(\tau)\|^p \, \mathrm{d}\tau \right)^{1/p} \left(\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)\| \, \mathrm{d}\tau \right)^{1/p'} \\
&\leq C_\infty^{1/p'} \left(\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)\| \|\mu(\tau)\|^p \, \mathrm{d}\tau \right)^{1/p}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)\|_p^p &\leq C_\infty^{p/p'} \int_{\mathbb{T}} \left(\int_{\mathbb{T}} \|\mathsf{K}(t,\tau)\| \|\mu(\tau)\|^p \, \mathrm{d}t \right) \mathrm{d}\tau \\
&\leq C_\infty^{p/p'} \int_{\mathbb{T}} \left(\int_{\mathbb{T}} \|\mathsf{K}_t\|_\infty \|\mu(\tau)\|^p \, \mathrm{d}t \right) \mathrm{d}\tau \\
&\leq C_\infty^{p/p'} \left(\int_{\mathbb{T}} \|\mu(\tau)\|^p \, \mathrm{d}\tau \right) \left(\int_{\mathbb{T}} \|\mathsf{K}_t\|_\infty \, \mathrm{d}t \right) \\
&\leq C_\infty^{p/p'} C_1 \|\mu\|_p^p.
\end{aligned}$$

Thus we have

$$\|g_{\mathsf{K}}(\mu)\|_p \leq C_1^{1/p} C_\infty^{1/p'} \|\mu\|_p,$$

giving the result, again using Theorem III-3.5.8. ■

The preceding result, while interesting, is limited in scope. Indeed, it has nothing to say about systems that take $\mathsf{L}^p_{\mathrm{loc}}(\mathbb{T}; \mathsf{U})$ to $\mathsf{L}^q_{\mathrm{loc}}(\mathbb{T}; \mathsf{Y})$. The restriction in Theorem 6.7.4 to input and output spaces that are L^p-spaces has more to do with the stability of the systems than with their general system theoretic attributes. However, to overcome these limitations in systematic way requires putting some general restrictions on the kernel K and/or the input signals $\mathscr{U}$. One nice class of kernels are those that give rise to causal systems. Let us define the class of kernels in this case.

6.7.5 Definition (Causal integral kernel) Let $\mathbb{T}$ be a continuous time-domain, and let U and Y be finite-dimensional $\mathbb{R}$-vector spaces. An integral kernel

$$\mathsf{K} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$$

is ***causal*** if $\mathsf{K}(t,\tau) = 0$ for $\tau > t$. •

Let us relate this notion of a causal integral kernel to a causal system.

6.7.6 Lemma (Causal integral kernels give rise to strongly causal integral kernel systems) *Let $\mathbb{T}$ be a continuous time-domain, let U and Y be finite-dimensional $\mathbb{R}$-vector spaces, and let $\mathscr{U}$ be a set of input signals. If K is a causal integral kernel compatible with $\mathscr{U}$, then the continuous-time input/output system $\mathrm{g}_{\mathsf{K}} \colon \mathscr{U} \to \mathsf{Y}^{\mathbb{T}}$ is strongly causal.*

Proof Let $\mu_1, \mu_2 \in \mathscr{U}$ satisfy $\operatorname{dom}(\mu_1) = \operatorname{dom}(\mu_2)$ and let $t \in \operatorname{dom}(\mu_1) = \operatorname{dom}(\mu_2)$. Suppose that $(\mu_1)_{\mathbb{T}_{<t} \cap \operatorname{dom}(\mu_1)} = (\mu_2)_{\mathbb{T}_{<t} \cap \operatorname{dom}(\mu_2)}$. Then

$$\begin{aligned} g_{\mathsf{K}}(\mu_1)(t) &= \int_{\mathbb{T}} \mathsf{K}(t,\tau)(\mu_1(\tau))\,\mathrm{d}\tau = \int_{\mathbb{T}_{<t}} \mathsf{K}(t,\tau)(\mu_1(\tau))\,\mathrm{d}\tau \\ &= \int_{\mathbb{T}_{<t}} \mathsf{K}(t,\tau)(\mu_2(\tau))\,\mathrm{d}\tau = \int_{\mathbb{T}} \mathsf{K}(t,\tau)(\mu_2(\tau))\,\mathrm{d}\tau = g_{\mathsf{K}}(\mu_2)(t). \end{aligned}$$

This is the desired strong causality. ■

One might like to have the causality of the kernel as being necessary for the strong causality of the associated integral operator. However, to state a general such theorem requires having some relationship between the kernel and the set of inputs that will just be confusing. The basic idea, however, is clear. If the integral kernel is *not* causal, then there will be some $t \in \mathbb{T}$ and an interval $\mathbb{S} \subseteq \mathbb{T}_{>t}$ such that

$$\int_{\mathbb{S}} \|\mathsf{K}(t,\tau)\|\,\mathrm{d}\tau \neq 0.$$

Generally speaking, one can expect there to be an input μ for which

$$\int_{\mathbb{S}} \mathsf{K}(t,\tau)(\mu(\tau))\,\mathrm{d}\tau \neq 0.$$

If one can additionally ask that $\operatorname{supp}(\mu) \subseteq \mathbb{S}$, then we would have $g_{\mathsf{K}}(\mu)(t) \neq 0$, even though μ is zero up to time t. This would preclude causality.

With the above considerations at hand, let us consider situations where a causal integral kernel defines an integral kernel system. As we see, the condition of causality of the integral kernel, as well as the causality of the set of input signals as in Definition IV-1.1.16, ensures causality of the system.

6.7.7 Theorem (Integral kernel systems with causal kernels and causal inputs) *Let $\mathbb{T} \subseteq \mathbb{R}$ be a continuous time-domain, let U and Y be finite-dimensional $\mathbb{R}$-vector spaces, and let $\mathrm{p} \in [1, \infty]$. In the following cases, the causal integral kernel $\mathsf{K} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$, the input space $\mathscr{U}$, and the output space $\mathscr{Y}$ are such that*

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathsf{K})$$

is an integral kernel system:

(i) *(a)* $\mathsf{K}_{\mathrm{t}} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_{\mathbb{K},1}$ *is in* $\mathrm{L}^\infty_{\mathrm{loc}}(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for every compact interval* $\mathbb{K} \subseteq \mathbb{T}$,

(b) $\mathscr{U} \subseteq \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T};\mathsf{U})$ *and there exists* $\mathsf{t}_0 \in \mathbb{T}$ *such that* $\operatorname{supp}(\mu) \subseteq \mathbb{T}_{\geq \mathsf{t}_0}$ *for every* $\mu \in \mathscr{U}$, *and*

(c) $\mathscr{Y} = \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T};\mathsf{Y})$;

(ii) (a) $\mathsf{K}_\mathsf{t} \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{U};\mathsf{Y}))$ *for each* $\mathsf{t} \in \mathbb{T}$ *and* $\mathsf{t} \mapsto \|\mathsf{K}_\mathsf{t}\|_{\mathbb{K},\infty}$ *is in* $\mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{U};\mathsf{Y}))$ *for every compact interval* $\mathbb{K} \subseteq \mathbb{T}$,

(b) $\mathscr{U} \subseteq \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathsf{U})$ *and there exists* $\mathsf{t}_0 \in \mathbb{T}$ *such that* $\operatorname{supp}(\mu) \subseteq \mathbb{T}_{\geq \mathsf{t}_0}$ *for every* $\mu \in \mathscr{U}$, *and*

(c) $\mathscr{Y} = \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathsf{Y})$;

(iii) (a) $\mathsf{K}_\mathsf{t} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{U};\mathsf{Y}))$ *for each* $\mathsf{t} \in \mathbb{T}$ *and* $\mathsf{t} \mapsto \|\mathsf{K}_\mathsf{t}\|_{\mathbb{K},1}$ *is in* $\mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{U};\mathsf{Y}))$ *for every compact interval* $\mathbb{K} \subseteq \mathbb{T}$,

(b) $\mathsf{K}_\mathsf{t} \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{U};\mathsf{Y}))$ *for each* $\mathsf{t} \in \mathbb{T}$ *and* $\mathsf{t} \mapsto \|\mathsf{K}_\mathsf{t}\|_{\mathbb{K},\infty}$ *is in* $\mathsf{L}^1_{\mathrm{loc}}(\mathbb{T};\mathrm{L}(\mathsf{U};\mathsf{Y}))$ *for every compact interval* $\mathbb{K} \subseteq \mathbb{T}$,

(c) $\mathscr{U} \subseteq \mathsf{L}^\mathsf{p}_{\mathrm{loc}}(\mathbb{T};\mathsf{U})$ *and there exists* $\mathsf{t}_0 \in \mathbb{T}$ *such that* $\operatorname{supp}(\mu) \subseteq \mathbb{T}_{\geq \mathsf{t}_0}$ *for every* $\mu \in \mathscr{U}$, *and*

(d) $\mathscr{Y} = \mathsf{L}^\mathsf{p}_{\mathrm{loc}}(\mathbb{T};\mathsf{Y})$.

Proof (i) Let us first show that K is compatible with $\mathscr{U}$. Let $\mu \in \mathscr{U}$ and let $t \in \mathbb{T}$. Then

$$\mathsf{K}(t,\tau)\mu(\tau) = 0, \qquad \tau < t_0,\ \tau > t.$$

Thus $\operatorname{supp}(\mathsf{K}_t(\mu)) \subseteq [t_0, t]$. Thus our hypotheses and Exercise III-3.8.8 ensure that $\mathsf{K}_t(\mu) \in \mathsf{L}^1(\mathbb{T};\mathsf{Y})$.

Let $\mathbb{K} \subseteq \mathbb{T}$ be a compact interval and let $\mathbb{L} \subseteq \mathbb{T}$ be a compact interval such that $\mathbb{K}_{\geq t_0} \subseteq \mathbb{L}$. Then, for $\mu \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T};\mathsf{U})$, we have

$$\begin{aligned}
\|g_\mathsf{K}(\mu)\|_{\mathbb{K},\infty} &\leq \operatorname{ess\,sup}\left\{ \int_\mathbb{T} \|\mathsf{K}(t,\tau)(\mu(\tau))\| \,\mathrm{d}\tau \;\middle|\; t \in \mathbb{K} \right\} \\
&= \operatorname{ess\,sup}\left\{ \int_{t_0}^t \|\mathsf{K}(t,\tau)(\mu(\tau))\| \,\mathrm{d}\tau \;\middle|\; t \in \mathbb{K}_{\geq t_0} \right\} \\
&\leq \|\mu\|_{\mathbb{L},\infty} \operatorname{ess\,sup}\left\{ \int_{t_0}^t \|\mathsf{K}(t,\tau)\| \,\mathrm{d}\tau \;\middle|\; t \in \mathbb{K}_{\geq t_0} \right\} \\
&\leq \underbrace{\operatorname{ess\,sup}\{\|\mathsf{K}_t\|_{\mathbb{K},1} \mid t \in \mathbb{K}\}}_{C_{\mathbb{K},\infty}} \|\mu\|_{\mathbb{L},\infty}.
\end{aligned}$$

Thus $\|g_\mathsf{K}(\mu)\|_{\mathbb{K},\infty} \leq C_{\mathbb{K},\infty}\|\mu\|_{\mathbb{L},\infty}$, and this gives the result.

(ii) One prove compatibility of K with $\mathscr{U}$ similarly to part (i).

Now let $\mathbb{K} \subseteq \mathbb{T}$ be compact and let $\mathbb{L} \subseteq \mathbb{T}$ be a compact interval for which $\mathbb{K}_{\geq t_0} \subseteq \mathbb{L}$.

Then, for $\mu \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T};\mathsf{U})$, we compute

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)\|_{\mathbb{K},1} &= \int_{\mathbb{K}} \left\| \int_{\mathbb{T}} \mathsf{K}(t,\tau)(\mu(\tau))\,\mathrm{d}\tau \right\| \mathrm{d}t \\
&\leq \int_{\mathbb{K}} \left(\int_{t_0}^{t} \|\mathsf{K}(t,\tau)(\mu(\tau))\|\,\mathrm{d}\tau \right) \mathrm{d}t \\
&= \int_{t_0}^{t} \left(\int_{\mathbb{K}} \|\mathsf{K}(t,\tau)(\mu(\tau))\|\,\mathrm{d}t \right) \mathrm{d}\tau \\
&\leq \int_{t_0}^{t} \left(\int_{\mathbb{K}} \|\mathsf{K}(t,\tau)\|\,\mathrm{d}t \right) \|\mu(\tau)\|\,\mathrm{d}\tau \\
&\leq \underbrace{\left(\int_{\mathbb{K}} \|\mathsf{K}_t\|_{\mathbb{K},\infty}\,\mathrm{d}t \right)}_{C_{\mathbb{K},1}} \left(\int_{\mathbb{L}} \|\mu(\tau)\|\,\mathrm{d}\tau \right).
\end{aligned}$$

That is, $\|g_{\mathsf{K}}(\mu)\|_{\mathbb{K},1} \leq C_{\mathbb{K},1}\|\mu\|_{\mathbb{L},1}$, giving this part of the result.

(iii) We can take $p \in (1,\infty)$. The compatibility of K with $\mathscr{U}$ is proved by combining the argument from part (i) above and part (iii) from Theorem 6.7.4.

Now let $\mathbb{K} \subseteq \mathbb{T}$ be a compact interval and take a compact interval $\mathbb{L} \subseteq \mathbb{T}$ such that $\mathbb{K}_{\geq t_0} \subseteq \mathbb{L}$. Let $C_{\mathbb{K},\infty}$ and $C_{\mathbb{K},1}$ be as defined in the first two parts of the proof. Then, for $\mu \in \mathsf{L}^p_{\mathrm{loc}}(\mathbb{T};\mathsf{U})$ and $t \in \mathbb{K}_{\geq t_0}$, we compute

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)(t)\| &\leq \int_{t_0}^{t} \|\mathsf{K}(t,\tau)(\mu(\tau))\|\,\mathrm{d}\tau \\
&\leq \int_{t_0}^{t} (\|\mathsf{K}(t,\tau)\|^{1/p}\|\mu(\tau)\|)\|\mathsf{K}(t,\tau)\|^{1/p'}\,\mathrm{d}\tau \\
&\leq \left(\int_{t_0}^{t} \|\mathsf{K}(t,\tau)\|\|\mu(\tau)\|^p\,\mathrm{d}\tau \right)^{1/p} \left(\int_{t_0}^{t} \|\mathsf{K}(t,\tau)\|\,\mathrm{d}\tau \right)^{1/p'} \\
&\leq C^{1/p'}_{\mathbb{K},\infty} \left(\int_{t_0}^{t} \|\mathsf{K}(t,\tau)\|\|\mu(\tau)\|^p\,\mathrm{d}\tau \right)^{1/p}.
\end{aligned}$$

We also have $g_{\mathsf{K}}(\mu)(t) = 0$ for $t < t_0$. Therefore, using Fubini's Theorem,

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)\|^p_{\mathbb{K},p} &\leq C^{p/p'}_{\mathbb{K},\infty} \int_{t_0}^{t} \left(\int_{\mathbb{K}} \|\mathsf{K}(t,\tau)\|\|\mu(\tau)\|^p\,\mathrm{d}t \right) \mathrm{d}\tau \\
&\leq C^{p/p'}_{\mathbb{K},\infty} \int_{t_0}^{t} \left(\int_{\mathbb{K}} \|\mathsf{K}_t\|_{\mathbb{K},\infty}\|\mu(\tau)\|^p\,\mathrm{d}t \right) \mathrm{d}\tau \\
&\leq C^{p/p'}_{\mathbb{K},\infty} \left(\int_{t_0}^{t} \|\mu(\tau)\|^p\,\mathrm{d}\tau \right) \left(\int_{\mathbb{K}} \|\mathsf{K}_t\|_{\mathbb{K},\infty}\,\mathrm{d}t \right) \\
&\leq C^{p/p'}_{\mathbb{K},\infty} C_{\mathbb{K},1}\|\mu\|^p_{\mathbb{L},p}.
\end{aligned}$$

Thus we have

$$\|g_{\mathsf{K}}(\mu)\|_{\mathbb{K},p} \leq C^{1/p}_{\mathbb{K},1} C^{1/p'}_{\mathbb{K},\infty}\|\mu\|_{\mathbb{L},p},$$

which is the desired result. ■

6.7.3 Integral kernel systems with distribution kernels

6.7.4 Continuous-time convolution systems

In this section we consider a special class of integral kernel systems. These arise from requiring stationarity of the integral kernel system, and the following result captures the manner in which stationarity arises. We focus on systems with time-domain $\mathbb{T} = \mathbb{R}$, since this can be done without loss of generality in any case.

6.7.8 Proposition (Stationary integral kernel systems) *Let* U *and* Y *be finite-dimensional* $\mathbb{R}$*-vector spaces and let* $\mathsf{K}\colon \mathbb{R} \times \mathbb{R} \to \mathrm{L}(\mathsf{U};\mathsf{Y})$ *be an integral kernel compatible with a set* $\mathscr{U}$ *of input signals. Suppose that* $\mathscr{U}$ *is translation invariant, i.e., that* $\tau_{\mathrm{a}}^*\mu \in \mathscr{U}$ *for every* $\mathrm{a} \in \mathbb{R}$ *and* $\mu \in \mathscr{U}$*. Denote by*

$$\Sigma_{\mathsf{K}} = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathsf{Y}^{\mathbb{R}}, \mathbb{R}, \mathrm{g}_{\mathsf{K}})$$

the continuous-time input/output system. Then the following statements hold:

(i) if

(a) $\mathsf{K} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^2; \mathrm{L}(\mathsf{U};\mathsf{Y}))$*,*

(b) $\mathscr{U}$ *has the property that, if* $\mathrm{f} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$ *satisfies*

$$\int_{\mathbb{R}} \mathrm{f}(\mathrm{t})\mu(\mathrm{t})\,\mathrm{dt} = 0, \qquad \mu \in \mathscr{U},$$

then $\mathrm{f} = 0$*, and*

(c) Σ_{K} *is stationary,*

then there exists $\mathsf{k} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathrm{L}(\mathsf{U};\mathsf{Y}))$ *such that* $\mathsf{K}(\mathrm{t},\tau) = \mathsf{k}(\mathrm{t}-\tau)$ *for almost every* $(\mathrm{t},\tau) \in \mathbb{R}^2$*;*

(ii) if there exists $\mathsf{k} \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathrm{L}(\mathsf{U};\mathsf{Y}))$ *such that* $\mathsf{K}(\mathrm{t},\tau) = \mathsf{k}(\mathrm{t}-\tau)$ *for almost every* $(\mathrm{t},\tau) \in \mathbb{R}^2$*, then* Σ_{K} *is strongly stationary.*

Proof (i) The hypotheses ensure that, for every $a \in \mathbb{R}$ and for every behaviour (μ,η) for Σ_{K}, $(\tau_a^*\mu, \tau_a^*\eta)$ is also a behaviour. Note that this gives, for every behaviour (μ,η),

$$\eta(t) = \int_{\mathbb{R}} \mathsf{K}(t,\tau)(\mu(\tau))\,\mathrm{d}\tau$$

and

$$\eta(t-a) = \int_{\mathbb{R}} \mathsf{K}(t,\tau)(\mu(\tau-a))\,\mathrm{d}\tau$$

for almost every $t \in \mathbb{R}$. By a change of variable, the second of these equations becomes

$$\eta(t) = \int_{\mathbb{R}} \mathsf{K}(t+a,\tau+a)(\mu(\tau))\,\mathrm{d}\tau.$$

Thus we have

$$\int_{\mathbb{R}} (\mathsf{K}(t,\tau) - \mathsf{K}(t+a,\tau+a))(\mu(\tau))\,\mathrm{d}\tau = 0$$

for almost every $t \in \mathbb{R}$. Thus

$$\mathsf{K}(t,\tau) = \mathsf{K}(t+a,\tau+a), \qquad a \in \mathbb{R}, \text{ a.e. } (t,\tau) \in \mathbb{R}^2.$$

Thus let $Z \subseteq \mathbb{R}^2$ be such that

$$\mathsf{K}(t,\tau) = \mathsf{K}(t+a,\tau+a), \qquad a \in \mathbb{R}, \ (t,\tau) \in \mathbb{R}^2 \setminus Z.$$

Therefore, for $(t,\tau) \in \mathbb{R}^2 \setminus Z$ we have, by taking $a = -\tau$, $\mathsf{K}(t,\tau) = \mathsf{K}(t-\tau,0)$. Therefore, taking $\mathsf{k}(t) = \mathsf{K}(t,0)$, we get the result.

(ii) We leave this to the reader as Exercise 6.7.2. ■

With this result in mind, we make the following definitions.

6.7.9 Definition (Convolution kernel, convolution operator defined by convolution kernel) Let U and Y be finite-dimensional $\mathbb{R}$-vector spaces.

(i) A ***continuous-time convolution kernel*** from U to Y is a mapping

$$\mathsf{k} \colon \mathbb{R} \to \mathrm{L}(\mathsf{U};\mathsf{Y}).$$

Let $\mathscr{U} \subseteq \mathsf{U}^{\mathbb{R}}$ be a subspace.

(ii) A continuous-time convolution kernel k is ***compatible*** with $\mathscr{U}$ if, for every $\mu \in \mathscr{U}$ and for almost every $t \in \mathbb{R}$, $\tau \mapsto \mathsf{k}(t-\tau) \circ \mu(\tau) \in \mathrm{L}^1(\mathbb{R};\mathrm{L}(\mathsf{U};\mathsf{Y}))$.

(iii) If k is compatible with $\mathscr{U}$, the ***continuous-time convolution operator*** defined by k is the mapping

$$g_{\mathsf{k}} \colon \mathscr{U} \to \mathsf{Y}^{\mathbb{R}}$$

defined by

$$g_{\mathsf{k}}(\mu)(t) = \int_{\mathbb{R}} \mathsf{k}(t-\tau)(\mu(\tau))\,\mathrm{d}\tau, \qquad \text{a.e. } t \in \mathbb{R}. \qquad \bullet$$

6.7.10 Definition (Continuous-time convolution system) A ***continuous-time convolution system*** is a quintuple $\Sigma = (\mathsf{U},\mathsf{Y},\mathscr{U},\mathscr{Y},\mathsf{k})$ where

(i) U (the ***input space***) and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathscr{U} \subseteq \mathsf{U}^{\mathbb{R}}$ is a subspace of signals with topology (the ***input signals***),

(iii) $\mathscr{Y} \subseteq \mathsf{Y}^{\mathbb{R}}$ is a subspace of signals with topology (the ***output signals***),

(iv) $\mathsf{k} \in$ is a continuous-time convolution kernel compatible with $\mathscr{U}$, and

(v) the continuous-time convolution operator g_{k} is a continuous linear mapping from $\mathscr{U}$ to $\mathscr{Y}$. $\bullet$

Of course, a continuous-time convolution system

$$\Sigma = (\mathsf{U},\mathsf{Y},\mathscr{U},\mathscr{Y},\mathsf{k})$$

is an integral kernel system

$$\Sigma' = (\mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathscr{Y}, \mathsf{K})$$

with $\mathsf{K}(t, \tau) = \mathsf{k}(t - \tau)$. Moreover, in Sections IV-4.2.1 and IV-4.2.2, we gave results about convolvable pairs of signals defined on $\mathbb{R}$ that we can use here to give some specific instances of continuous-time convolution systems. We refer the reader to the above listed sections for precise results as reproducing these would be an unnecessary distraction.

As with Theorem 6.7.4, the results from Sections IV-4.2.1 and IV-4.2.2 are quite restrictive in that they apply only to signals that are integrable in some sense, and this is a quite limited class of signals. This can be rectified, both mathematically and practically, by restricting to causal kernels and inputs. Based on Definition 6.7.5, we make the following definition.

6.7.11 Definition (Causal continuous-time convolution kernel) Let U and Y be finite-dimensional $\mathbb{R}$-vector spaces. A continuous-time convolution kernel

$$\mathsf{k} \colon \mathbb{R} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$$

is ***causal*** if $\mathsf{k}(t) = 0$ for $t \in \mathbb{R}_{<0}$. •

We can then extend the applicability of the results from Sections IV-4.2.1 and IV-4.2.2 to causal convolution kernels, and with spaces of input and output signals that are only appropriately locally integrable. In this respect, we refer the reader to Sections IV-4.2.3, and IV-4.2.4 that provide some classes of convolution systems with causal kernels. Rather than reproduce all of the results from these sections in our specific setting here, let us simply indicate the steps one must take to adapt the results.

Let U and Y be finite-dimensional $\mathbb{R}$-vector spaces and let k be a causal convolution kernel residing in an appropriate space of locally integrable signals. Suppose that $\mathscr{U}$ is a subset of an appropriate space of locally integrable signals and that $t_0 \in \mathbb{R}$ is such that $\mu(t) = 0$ for all $\mu \in \mathscr{U}$ and $t < t_0$. Let $\mathbb{K} \subseteq \mathbb{R}$ be a compact interval satisfying

$$\sup \mathbb{K} \geq \inf \operatorname{supp}(\mu),$$

noting that, when this is not true, then $\mathsf{k} * \mu | \mathbb{K} = 0$. Then, letting $\mathbb{L}$ satisfy

$$\inf \mathbb{L} \leq \min\{0, t_0\},$$
$$\sup \mathbb{L} \geq \max\{\sup \mathbb{K}, \sup \mathbb{K} - t_0\},$$

we can use the appropriate variant of, for example, Theorem IV-4.2.19, to give continuity of the input/output g_{k}.

6.7.12 Remark (The "punchline" for continuous-time convolution systems) The technicalities of the results in this section may obscure the simple reasons why continuous-time convolution systems are important. Let us summarise these reasons.

1. Among the integral kernel systems, convolution systems are distinguished by being the stationary systems. This is the content of Proposition 6.7.8.
2. Causality for continuous-time convolution systems is easily characterised by the requirement that the convolution kernel vanish for negative time. Thus causal continuous-time convolution systems give a large and interesting class of causal stationary continuous-time linear systems.

6.7.5 Continuous-time convolution systems with distribution kernels

Schwartz kernel theorem

6.7.6 Linear continuous-time state space systems as linear continuous-time input/output systems

A merely mildly astute reader will have noticed that integral kernel system arise in the input/output relations for linear continuous-time state space systems, and that this integral kernel description simplifies to a convolution system in the case where the state space system has constant coefficients. In this section we make these connections precise, which is comparatively easy given the work that we have done already.

6.7.6.1 The time-varying case Let us get straight to the point and state the main results of this section.

6.7.13 Theorem (Integral kernel systems from linear continuous-time state space systems) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system and let $p \in [1, \infty]$. *Let* $t_0 \in \mathbb{T}$ *and let*

$$\mathscr{U} \subseteq \{\mu \in \mathsf{L}^p_{\mathrm{loc}}(\mathbb{T}; \mathsf{U}) \mid \mu(t) = 0,\ t < t_0\},$$
$$\mathscr{Y} = \{\eta \in \mathsf{L}^p_{\mathrm{loc}}(\mathbb{T}; \mathsf{Y}) \mid \eta(t) = 0,\ t < t_0\}.$$

Then

$$\Sigma_{\mathrm{i/o}}(t_0) = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathrm{pitm}_\Sigma)$$

is a causal integral kernel system in the following cases:

(i) *(a)* $\mathsf{B} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{U}; \mathsf{X}))$ *and* $\mathsf{C} \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{X}; \mathsf{Y}))$ *and*
(b) $p = \infty$*;*

(ii) *(a)* $\mathsf{B} \in \mathsf{L}^\infty_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{U}; \mathsf{X}))$ *and* $\mathsf{C} \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{T}; \mathsf{L}(\mathsf{X}; \mathsf{Y}))$ *and*

(b) $p = 1$;

(iii) *(a)* $B \in L^1_{\mathrm{loc}}(\mathbb{T}; L(U; X))$ *and* $C \in L^\infty_{\mathrm{loc}}(\mathbb{T}; L(X; Y))$,

(b) $B \in L^\infty_{\mathrm{loc}}(\mathbb{T}; L(U; X))$ *and* $C \in L^1_{\mathrm{loc}}(\mathbb{T}; L(X; Y))$, *and*

(c) $p \in [1, \infty]$.

Proof We shall show that the integral kernel pitm_Σ satisfies the following conditions from Theorem 6.7.7:

1. (part (i)) $\mathrm{pitm}_{\Sigma,t} \in L^1_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ for each $t \in \mathbb{T}$ and $t \mapsto \|\mathrm{pitm}_{\Sigma,t}\|_{\mathbb{K},1}$ is in $L^\infty_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ for every compact interval $\mathbb{K} \subseteq \mathbb{T}$;
2. (part (ii)) $\mathrm{pitm}_{\Sigma,t} \in L^\infty_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ for each $t \in \mathbb{T}$ and $t \mapsto \|\mathrm{pitm}_{\Sigma,t}\|_{\mathbb{K},\infty}$ is in $L^1_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ for every compact interval $\mathbb{K} \subseteq \mathbb{T}$;
3. (part (iii))
   (a) $\mathrm{pitm}_{\Sigma,t} \in L^1_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ for each $t \in \mathbb{T}$ and $t \mapsto \|\mathrm{pitm}_{\Sigma,t}\|_{\mathbb{K},1}$ is in $L^\infty_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ for every compact interval $\mathbb{K} \subseteq \mathbb{T}$ and
   (b) $\mathrm{pitm}_{\Sigma,t} \in L^\infty_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ for each $t \in \mathbb{T}$ and $t \mapsto \|\mathrm{pitm}_{\Sigma,t}\|_{\mathbb{K},\infty}$ is in $L^1_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ for every compact interval $\mathbb{K} \subseteq \mathbb{T}$.

We recall from Theorem 3.2.13(viii) that $(t, \tau) \mapsto \Phi^c_A(t, \tau)$ is continuous. Thus $\tau \mapsto \Phi^c_A(t, \tau)$ is in $L^r_{\mathrm{loc}}(\mathbb{T}; L(X; X))$ for every $t \in \mathbb{T}$ and every $r \in [1, \infty]$. Let $q_1, q_2 \in [1, \infty]$. Then

$$\tau \mapsto \mathrm{pitm}_\Sigma(t, \tau) = 1_{\geq 0}(t - \tau) C(t) \circ \Phi^c_A(t, \tau) \circ B(\tau)$$

is in $L^{q_1}_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ if and only if $B \in L^{q_1}_{\mathrm{loc}}(\mathbb{T}; L(U; X))$. Now let $\mathbb{K}, \mathbb{L} \subseteq \mathbb{T}$ be compact and denote

$$M = \sup\{\|\Phi^c_A(t, \tau)\| \mid (t, \tau) \in \mathbb{L} \times \mathbb{K}\},$$

noting that $M < \infty$ since Φ^c_A is continuous. Suppose first that $q_1, q_2 \in [1, \infty)$ and compute

$$\begin{aligned}\int_{\mathbb{L}} \|\mathrm{pitm}_{\Sigma,t}\|^{q_2}_{\mathbb{K},q_1} \, \mathrm{d}t &\leq \int_{\mathbb{L}} \left\| \left(\int_{\mathbb{K}} \left\| C(t) \circ \Phi^c_A(t, \tau) \circ B(\tau) \right\|^{q_1} \, \mathrm{d}\tau \right)^{1/q_1} \right\|^{q_2} \, \mathrm{d}t \\ &\leq M^{q_2} \|B\|^{q_2}_{\mathbb{K},q_1} \int_{\mathbb{L}} \|C(t)\|^{q_2} \, \mathrm{d}t.\end{aligned}$$

Thus we conclude in this case that

$$\tau \mapsto \mathrm{pitm}_\Sigma(t, \tau) = 1_{\geq 0}(t - \tau) C(t) \circ \Phi^c_A(t, \tau) \circ B(\tau)$$

is in $L^{q_1}_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ and

$$t \mapsto \|\mathrm{pitm}_{\Sigma,t}\|_{\mathbb{K},q_1}$$

is in $L^{q_2}_{\mathrm{loc}}(\mathbb{T}; L(U; Y))$ if and only if $B \in L^{q_1}_{\mathrm{loc}}(\mathbb{T}; L(U; X))$ and $C \in L^{q_2}_{\mathrm{loc}}(\mathbb{T}; L(X; Y))$. Using similarly styled arguments, one shows that this conclusion is valid for $q_1, q_2 \in [1, \infty]$.

The theorem now follows from Theorem 6.7.7 by considering the pairs $(q_1, q_2) \in \{(1, \infty), (\infty, 1), (p, p)\}$. Note that causality follow by the definitions. ∎

We see from our decomposition (6.11) of an arbitrary output of a linear continuous-time state space system that the map sending an input μ to the second term in this decomposition is precisely the input/output system $\Sigma_{\mathrm{i/o}}(t_0)$.

Distribution version of this theorem.

6.7.6.2 The constant coefficient case In this case, the main result is the following.

6.7.14 Theorem (Continuous-time convolution systems from linear continuous-time state space systems with constant coefficients) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients and let $\mathrm{p}, \mathrm{q}, \mathrm{r} \in [1, \infty]$ *satisfy one of the following two criterion: (1)* $\mathrm{p} = \mathrm{q} = \mathrm{r} = 1$*; (2)* $\frac{1}{\mathrm{p}} - \frac{1}{\mathrm{q}} = 1 - \frac{1}{\mathrm{r}}$*. Let*

$$\begin{aligned}\mathscr{U} &\subseteq \{\mu \in \mathsf{L}^{\mathrm{p}}_{\mathrm{loc}}(\mathbb{R}; \mathsf{U}) \mid \ \mu(\mathrm{t}) = 0, \ \mathrm{t} < 0\},\\ \mathscr{Y} &= \{\eta \in \mathsf{L}^{\mathrm{q}}_{\mathrm{loc}}(\mathbb{R}; \mathsf{Y}) \mid \ \eta(\mathrm{t}) = 0, \ \mathrm{t} < 0\}.\end{aligned}$$

Then

$$\Sigma_{\mathrm{i/o}} = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathrm{pir}_{\Sigma})$$

is a causal continuous-time convolution system.

Proof We have

$$\mathrm{pir}_{\Sigma}(t) = 1_{\geq 0}(t)\mathsf{C} \circ \mathrm{e}^{\mathsf{A}t} \circ \mathsf{B},$$

and so pir_{Σ} is continuous by Theorem 5.2.6(i) and Theorem 5.2.20(ix). Thus $\mathrm{pir}_{\Sigma} \in \mathsf{L}^{r}_{\mathrm{loc}}(\mathbb{R}; \mathsf{L}(\mathsf{U}; \mathsf{Y}))$ for every $r \in [1, \infty]$. The result now follows from Corollary IV-4.2.14 and Theorem IV-4.2.19. ■

Here we see that the map sending an input to its zero-state/zero-input response, as in Definition 6.6.17, defines the input/output system $\Sigma_{\mathrm{i/o}}$.

Distribution version of this theorem.

6.7.7 Linear continuous-time differential input/output systems

Exercises

6.7.1 For each of the listed attributes, give an example of a linear continuous-time input/output system with that attribute. You are not allowed to choose a system of the sort considered in either of Sections 6.7.2 and 6.7.4.

Here are the attributes:

(a) causal;
(b) not causal;
(c) stationary;
(d) not stationary;
(e) memoryless;
(f) not memoryless.

6.7.2 Complete the proof of Proposition 6.7.8.

6.7.3 For $a \in \mathbb{R}_{\geq 0}$, consider the function

$$\begin{aligned} d_a \colon \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathbb{F}) &\to \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathbb{F}) \\ \mu &\mapsto \tau_a^* \mu. \end{aligned}$$

Answer the following questions.

(a) Show that d_a is a linear continuous-time input/output system.
(b) Determine its system theoretic properties, i.e., is it causal? strongly causal? finitely observable? stationary? strongly stationary?
(c) Let $t_0 \in \mathbb{R}$. Show that, to determine $d_a(\mu)(t)$ for all $t \geq t_0$, you must know $\mu(t)$ for $t \geq t_0 - a$.
(d) Argue that the state space for the system starting from t_0 is $\mathsf{L}^1([t_0 - a, t_0]; \mathbb{F})$.
(e) Can this system be converted into a continuous-time state space system?
(f) Compare the situation in this discrete-time case with the continuous-time case presented in Exercise 6.9.3.

6.7.4 We consider a simple RLC circuit as depicted in Figure 6.9. The voltage

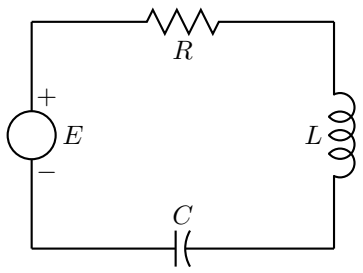


Figure 6.9 An RLC circuit

supplied by the source is the input with values denoted by v and if the current in the circuit is the output whose values are i.

Answer the following questions.

(a) Provide a differential equation that models the current $t \mapsto \imath(t)$ given the voltage $t \mapsto v(t)$ supplied by the source.
(b) Show that the mapping $v \mapsto \imath$ defines a continuous-time input/output system.
(c) Determine its system theoretic properties, i.e., is it causal? strongly causal? finitely observable? stationary? strongly stationary? memoryless?
(d) What is the state space for the system?
(e) What is the control set for the system?
(f) What is the time-domain for the system?
(g) What is a reasonable choice for the space $\mathscr{U}$ of input signals?

(h) If the source provides a constant voltage V_0, what is the current I_0 in the circuit as $t \to \infty$?

6.7.5 We consider the filter of Butterworth [1930]. Answer the following questions.

(a) For $n \in \mathbb{Z}_{>0}$, show that the polynomial

$$P_n = X^{2n} + (-1)^n$$

has $2n$ roots, exactly half of which have negative real part.

Let $\lambda_1, \dots, \lambda_n$ be the roots of P_n with negative real part.

(b) Show that the polynomial $Q_n = \prod_{j=1}^{n}(X - \lambda_j)$ has real coefficients.

The ***Butterworth filter*** of order n is determined by the ordinary differential equation

$$D_n(\eta) = \mu$$

for $\eta, \mu \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$, where D is the differential operator with constant coefficients whose symbol is Q_n; see Section 4.2.2.2 for notation.

(c) Explicitly determine the differential equations defining the Butterworth filters of orders $n \in \{1, 2, 3, 4, 5\}$.

6.7.6 A ***continuous-time sliding averager*** takes an input signal $\mu \in \mathsf{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathbb{R})$ and returns the signal

$$\eta(t) = \frac{1}{T_+ + T_-} \int_{t-T_-}^{t+T_+} \mu(\tau)\, \mathrm{d}\tau, \qquad t \in \mathbb{R},$$

for $T_-, T_+ \in \mathbb{R}_{\geq 0}$ with $T_+ + T_- \in \mathbb{R}_{>0}$. Answer the following questions.

(a) Show that this is a continuous-time convolution system and determine the convolution kernel.

(b) Is the system causal? strongly causal? stationary? strongly stationary? memoryless?

(c) Compute the output associated with the input $\mu = 1_{\geq 0}$.

(d) Let $T_+ = T_- = 1$ and compute the output associated with the input $\mu(t) = \sin(\pi t)$.

6.7.7 Consider a continuous-time convolution system $\Sigma = (\mathbb{R}, \mathbb{R}, \mathbb{R}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ with $\mathsf{k} \in \mathsf{L}^1(\mathbb{R}; \mathbb{R})$. Define the ***step response*** of the system to be the output associated with the input $\mu = 1_{\geq 0}$: $1_\Sigma = \mathsf{k} * 1_{\geq 0}$ (evidently we are assuming that $1_{\geq 0} \in \mathscr{U}$). Show that

$$1_\Sigma(t) = \int_{-\infty}^{t} \mathsf{k}(\tau)\, \mathrm{d}\tau, \quad \mathsf{k}(t) = \frac{\mathrm{d}1_\Sigma}{\mathrm{d}t}(t), \qquad t \in \mathbb{R}.$$

6.7.8 Using the step response from Exercise 6.7.7, suppose that you know that a continuous-time convolution system $\Sigma = (\mathbb{R}, \mathbb{R}, \mathbb{R}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ has the step response

$$1_\Sigma(t) = \begin{cases} 1, & t \in [a, b), \\ 0, & \text{otherwise} \end{cases}$$

for some $a, b \in \mathbb{R}$ satisfying $a < b$.

(a) Determine the convolution kernel.

(b) What is $g_{\mathsf{k}}(\mu)(t)$ for a continuous input μ?

(c) For which values of a and b is the system causal?

6.7.9 Suppose that you know that a continuous-time convolution system $\Sigma = (\mathbb{R}, \mathbb{R}, \mathbb{R}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ has the input/output pair shown in Figure 6.10. Determine

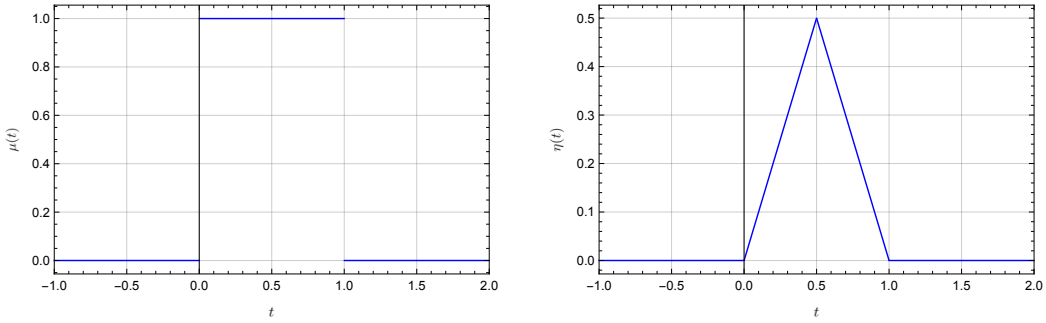


Figure 6.10 An input (left) and corresponding output (right) for a continuous-time convolution system

the convolution kernel for the system.

Hint: *Use linearity, stationarity, and Exercise 6.7.7.*

6.7.10 Consider the RLC circuit in Figure 6.11. Answer the following questions.

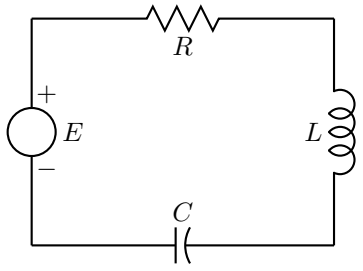


Figure 6.11 An RLC circuit

(a) Derive a scalar second-order differential equation for the current through the circuit given the voltage at the source as input. Express the equations as those for an initially at rest input/output system.

(b) Derive a scalar second-order differential equation for the voltage across the inductor given the voltage at the source as input. Express the equations as those for an initially at rest input/output system.

Section 6.8

Linear discrete-time state space systems

We now carry out for discrete-time systems the constructions of Section 6.6 for continuous-time systems, considering a particular class of discrete-time state space systems that are linear in both state and control. We mirror what we have done with ordinary difference equations by working with systems that are time-dependent, and then time-independent. Linearity will allow us to obtain more particular results than we were able to obtain for not necessarily linear systems.

Do I need to read this section? As with the preceding two sections, the material in this section is to be regarded as a core part of the material in this volume. •

6.8.1 Systems with time-varying coefficients

Let us begin with the definition, recalling from Section 3.3.3.3 the adaptation to using abstract vector spaces in place of Euclidean spaces for linear systems.

6.8.1 Definition (Linear discrete-time state space system) A *linear discrete-time state space system* is a nonuple

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

where

(i) X (the *state space*), U (the *input space*), and Y (the *output space*) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ is a discrete time-domain,

(iii) $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{X})$, $\mathsf{B}\colon \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{X})$, $\mathsf{C}\colon \mathbb{T} \to \mathrm{L}(\mathsf{X}; \mathsf{Y})$, and $\mathsf{D}\colon \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$, and

(iv) $\mathscr{U}$ is a collection of mappings $\mu\colon \mathbb{T} \to \mathsf{U}$. •

We note that a linear discrete-time state space system is, in particular, a discrete-time state space system (with the mild adaptation from using Euclidean spaces to using finite-dimensional vector spaces) with dynamics defined by

$$\begin{aligned} f\colon \mathbb{T} \times (\mathsf{X} \oplus \mathsf{U}) &\to \mathsf{X} \\ (t, x, u) &\mapsto \mathsf{A}(t)(x) + \mathsf{B}(t)(u) \end{aligned}$$

and with output map

$$\begin{aligned} h\colon \mathbb{T} \times (\mathsf{X} \oplus \mathsf{U}) &\to \mathsf{Y} \\ (t, x, u) &\mapsto \mathsf{C}(t)(x) + \mathsf{D}(t)(u). \end{aligned}$$

We note that linear discrete-time state space systems are, in fact, control-affine discrete-time state space systems. Therefore, all the notions attached to discrete-time state space systems can be applied to those that are linear. The system theoretic

attributes of Section 6.3.1 apply in exactly the same way for linear discrete-time state space systems; the reader can flesh this out in Exercise 6.8.1. One has the set Ctraj(Σ) of controlled trajectories and the set Cout(Σ) of controlled outputs. In particular, if (ξ, μ) is a controlled trajectory with (η, μ) the corresponding controlled output, then these satisfy the equations

$$\begin{aligned}\xi(t + \Delta) &= \mathsf{A}(t)(\xi(t)) + \mathsf{B}(t)(\mu(t)),\\ \eta(t) &= \mathsf{C}(t)(\xi(t)) + \mathsf{D}(t)(\mu(t)).\end{aligned}$$

Moreover, the existence and uniqueness result from Theorem 6.3.9 for general discrete-time state space systems can be adapted to linear systems. One gets the following result upon doing this.

6.8.2 Theorem (Existence and uniqueness of controlled trajectories for linear discrete-time state space systems) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system, let $\mathbb{T}' \subseteq \mathbb{T}$ be a sub-time-domain, let $\mu\colon \mathbb{T}' \to \mathsf{U}$. Then, for any $\mathrm{t}_0 \in \mathbb{T}'$ and $\mathrm{x}_0 \in \mathsf{X}$, there exists a unique solution $\xi\colon \mathbb{T}' \to \mathsf{X}$ to the initial value problem

$$\xi(\mathrm{t} + \Delta) = \mathsf{A}(\mathrm{t})(\xi(\mathrm{t})) + \mathsf{B}(\mathrm{t})(\mu(\mathrm{t})), \quad \xi(\mathrm{t}_0) \in \mathrm{x}_0;$$

thus $(\xi, \mu) \in$ Ctraj(Σ).

Proof This follows immediately from Theorem 6.3.9. ∎

Let us make a few more or less immediate comments about controlled trajectories and controlled outputs.

6.8.3 Remarks (Controlled trajectories and controlled outputs for linear discrete-time state space systems)

1. From Corollary 5.7.2 we have an explicit formula for the controlled trajectory $(\xi, \mu) \in$ Ctraj(Σ) with the initial condition x_0 at t_0:

$$\begin{aligned}\Phi^{\Sigma}(t, t_0, x_0, \mu) &= \Phi^{\mathrm{d}}_{\mathsf{A},t_0}(t)(x_0)\\ &\quad + \sum_{j=0}^{(t-t_0-\Delta)/\Delta} \Phi^{\mathrm{d}}_{\mathsf{A},t_0+(j+1)\Delta}(t) \circ \mathsf{B}(t_0 + j\Delta)(\mu(t_0 + j\Delta)), \qquad t \in \mathrm{dom}(\mu). \quad (6.12)\end{aligned}$$

   The corresponding controlled output (η, μ) is, of course, given by

$$\eta(t) = \mathsf{C}(t) \circ \Phi^{\Sigma}(t, t_0, x_0, \mu) + \mathsf{D}(t)(\mu(t)), \qquad t \in \mathrm{dom}(\mu).$$

2. We note that, in contrast to general discrete-time state space systems, controlled trajectories always exist on the entire domain of definition of the control. This is one feature that makes working with linear systems less complicated than working with general systems.

3. In contrast to linear continuous-time state space systems, we do not have to fuss with the exact way in which the system components A and B depend on time, and on how this time dependence interacts with the nature of time dependence of the control.
4. We can infer immediately from Proposition 6.3.12 and Theorem 6.3.13 the properties of the flow Φ^{Σ} for a linear discrete-time state space system. In particular, we have continuity of the flow with respect to initial condition, initial time, final time, and control. In fact, these conclusions follow most easily and directly from the formula (6.12). •

We note that the dynamics and the output mapping are linear functions of (x, u). That is to say, the mappings

$$\mathsf{X} \oplus \mathsf{U} \ni (x, u) \mapsto \mathsf{A}(t)(x) + \mathsf{B}(t)(u) \in \mathsf{X},$$
$$\mathsf{X} \oplus \mathsf{U} \ni (x, u) \mapsto \mathsf{C}(t)(x) + \mathsf{D}(t)(u) \in \mathsf{Y}$$

are linear for each $t \in \mathbb{T}$. Moreover, the flow is also linear in the sense of the following result.

6.8.4 Proposition (Linearity of flow for linear discrete-time state space systems) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system and let $\mathbb{T}' \subseteq \mathbb{T}$ be a sub-time-domain. Then, for each sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}$, the mapping

$$\mathsf{X} \oplus \mathscr{U}(\mathbb{T}') \ni (\mathsf{x}_0, \mu) \mapsto \Phi^{\Sigma}(\mathsf{t}, \mathsf{t}_0, \mathsf{x}_0, \mu) \in \mathsf{X}$$

is linear for each $\mathsf{t}, \mathsf{t}_0 \in \mathbb{T}'$.

Proof This follows immediately from the formula (6.12) for the flow of a linear discrete-time state space system. ■

6.8.2 Systems with constant coefficients

Now we consider systems with the coefficient linear mappings for the dynamics and the output map are independent of time. There are some simplifications that arise in this case that are worth recording, so we devote this section to this class of system.

6.8.5 Definition (Linear discrete-time state space system with constant coefficients) A ***linear discrete-time state space system with constant coefficients*** is a nonuple

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

where

(i) X (the ***state space***), U (the ***input space***), and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ is a discrete time-domain,
(iii) $\mathsf{A} \in \mathrm{L}(\mathsf{X};\mathsf{X})$, $\mathsf{B} \in \mathrm{L}(\mathsf{U};\mathsf{X})$, $\mathsf{C} \in \mathrm{L}(\mathsf{X};\mathsf{Y})$, and $\mathsf{D} \in \mathrm{L}(\mathsf{U};\mathsf{Y})$, and
(iv) $\mathscr{U} \subseteq \ell_{\mathrm{loc}}((\mathbb{T});\mathsf{U})$. •

In this case, the system is an autonomous discrete-time state space system, and the dynamics and output map are defined, independent of time, as

$$\begin{aligned} f\colon \mathsf{X} \oplus \mathsf{U} &\to \mathsf{X} \\ (x,u) &\mapsto \mathsf{A}(x) + \mathsf{B}(u) \end{aligned}$$

and

$$\begin{aligned} h\colon \mathsf{X} \oplus \mathsf{U} &\to \mathsf{X} \\ (x,u) &\mapsto \mathsf{C}(x) + \mathsf{D}(u), \end{aligned}$$

respectively.

We note that a linear discrete-time state space system with constant coefficients is autonomous, and so is stationary, and strongly stationary if and only if $\mathsf{D} = 0$ (see Exercise 6.8.1). This stationarity is often reflected with some particular terminology.

6.8.6 Terminology What we call a linear discrete-time state space system with constant coefficients is often called a ***linear time-invariant system***, or an ***LTI system***, in short. As with continuous-time systems, we shall stick to the more cumbersome terminology in order to maintain internal consistency with other terminology elsewhere in this volume, but do not object to a reader using the terminology "LTI system" in their private life. •

Note that a controlled trajectory (ξ, μ), with associated controlled output (η, μ), satisfies

$$\begin{aligned} \xi(t+\Delta) &= \mathsf{A}(\xi(t)) + \mathsf{B}(\mu(t)), \\ \eta(t) &= \mathsf{C}(\xi(t)) + \mathsf{D}(\mu(t)). \end{aligned}$$

We have the following slight simplification of the existence and uniqueness theorem for systems with constant coefficients.

6.8.7 Theorem (Existence and uniqueness of controlled trajectories for linear discrete-time state space systems with constant coefficients) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients, let $\mathbb{T}' \subseteq \mathbb{T}$ be a sub-time-domain, and let $\mu\colon \mathbb{T}' \to \mathsf{U}$. Then, for any $\mathrm{t}_0 \in \mathbb{T}'$ and $\mathrm{x}_0 \in \mathsf{X}$, there exists a unique solution $\xi\colon \mathbb{T}' \to \mathsf{X}$ to the initial value problem

$$\xi(\mathrm{t}+\Delta) = \mathsf{A}(\xi(\mathrm{t})) + \mathsf{B}(\mu(\mathrm{t})), \quad \xi(\mathrm{t}_0) \in \mathrm{x}_0;$$

thus $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma)$.

Proof This follows immediately from Theorem 6.8.2. ■

We can simplify, for systems with constant coefficients, some of the discussion concerning flows and controlled outputs.

6.8.8 Remarks (Controlled trajectories and controlled outputs for linear discrete-time state space systems with constant coefficients)

1. From Theorem 5.7.7 we have an explicit formula for the controlled trajectory $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma)$ with the initial condition x_0 at t_0:

$$\Phi^{\Sigma}(t, t_0, x_0, \mu) = \mathsf{P}_{\mathsf{A}}\left(\tfrac{t-t_0}{\Delta}\right)(x_0) + \sum_{j=0}^{(t-t_0-\Delta)/\Delta} \mathsf{P}_{\mathsf{A}}\left(\tfrac{t-t_0-(j+1)\Delta}{\Delta}\right)(\mathsf{B}(\mu(t_0 + j\Delta))), \qquad t \in \mathrm{dom}(\mu). \quad (6.13)$$

   The corresponding controlled output (η, μ) is, of course, given by

$$\eta(t) = \mathsf{C} \circ \Phi^{\Sigma}(t, t_0, x_0, \mu) + \mathsf{D} \circ \mu(t), \qquad t \in \mathrm{dom}(\mu).$$

2. We can infer immediately from Proposition 6.3.12 and Theorem 6.3.13 the properties of the flow Φ^{Σ} for a linear discrete-time state space system with constant coefficients. In particular, we have continuity of the flow with respect to initial condition, initial time, final time, and control. In fact, these conclusions follow most easily and directly from the formula (6.13). •

The situation concerning linearity is similar for systems with constant coefficients to systems with time-varying coefficients. First we note that the dynamics and the output mapping are linear functions of (x, u). That is to say, the mappings

$$\mathsf{X} \oplus \mathsf{U} \ni (x, u) \mapsto \mathsf{A}(x) + \mathsf{B}(u) \in \mathsf{X},$$
$$\mathsf{X} \oplus \mathsf{U} \ni (x, u) \mapsto \mathsf{C}(x) + \mathsf{D}(u) \in \mathsf{Y}$$

are linear. Moreover, the flow is also linear in the sense of the following result.

6.8.9 Proposition (Linearity of flow for linear discrete-time state space systems with constant coefficients) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients, and let $\mathbb{T}' \subseteq \mathbb{T}$ be a sub-time-domain. Then, for each sub-time-domain $\mathbb{T}' \subseteq \mathbb{T}$, the mapping

$$\mathsf{X} \oplus \ell_{\mathrm{loc}}(\mathbb{T}'; \mathsf{U}) \ni (\mathsf{x}_0, \mu) \mapsto \Phi^{\Sigma}(\mathsf{t}, \mathsf{t}_0, \mathsf{x}_0, \mu) \in \mathsf{X}$$

is linear for each $\mathsf{t}, \mathsf{t}_0 \in \mathbb{T}'$.

Proof This follows immediately from the formula (6.13) for the flow of a linear discrete-time state space system. ■

6.8.3 The impulse transmission map and the impulse response

We now consider the discrete-time versions of the impulse transmission map and the impulse response.

6.8.3.1 The time-varying case We begin with the definition, recalling from Example IV-1.1.9–5 the definition of the pulse signal P. We shall, with a mild abuse of notation, denote

$$\tau^*_{t_0}\mathsf{P}(t) = \begin{cases} 1, & t = t_0, \\ 0, & t \neq t_0, \end{cases}$$

for $t, t_0 \in \mathbb{T}$. Let us also define

6.8.10 Definition (Impulse transmission map for linear discrete-time state space systems) Let

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system.

(i) The ***proper impulse transmission map*** for Σ at t_0 is the function

$$\mathrm{pitm}_{\Sigma,t_0} \colon \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$$

defined by

$$\mathrm{pitm}_{\Sigma,t_0}(t) = 1_{\geq 0}(t - (t_0 + \Delta))\mathsf{C}(t) \circ \Phi^{\mathrm{d}}_{\mathsf{A},t_0+\Delta}(t) \circ \mathsf{B}(t_0).$$

(ii) The ***impulse transmission map*** for Σ at $t_0 \in \mathbb{R}$ is the function

$$\mathrm{itm}_{\Sigma,t_0} \colon \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$$

defined by

$$\mathrm{itm}_{\Sigma,t_0}(t) = \mathrm{pitm}_{\Sigma,t_0}(t) + \tau^*_{t_0}\mathsf{P}(t)\mathsf{D}(t). \qquad \bullet$$

Let us give a simple, direct characterisation of the proper impulse response. We note that, unlike in the continuous-time case, in the discrete-time case we do not need to work with distributional interpretation. The reason for this is that the pulse gives the effect of a delta-signal, but is a *bona fide* signal in the discrete-time case.

6.8.11 Theorem (An interpretation of the proper impulse transmission map) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system and let $\mathrm{t}_0 \in \mathbb{R}$. *Then*

$$\mathrm{pitm}_{\Sigma,t_0}(\mathrm{t})(\mathrm{u}) = 1_{\geq 0}(\mathrm{t} - (\mathrm{t}_0 + \Delta))\mathsf{C}(\mathrm{t}) \circ \xi_{\mathrm{t}_0}(\mathrm{t}),$$

where $\xi_{\mathrm{t}_0} \colon \mathbb{T} \to \mathsf{X}$ *is the solution of the initial value problem*

$$\xi_{\mathrm{t}_0}(\mathrm{t} + \Delta) = \mathsf{A}(\mathrm{t}) + \mathsf{B}(\mathrm{t})(\tau^*_{\mathrm{t}_0}\mathsf{P}(\mathrm{t})\mathrm{u}), \quad \xi_{\mathrm{t}_0}(\mathrm{t}_0) = 0.$$

Proof Let $\mu = (\tau_{t_0}^* \mathsf{P})u$. By (6.12), the solution to the initial value problem in the statement of the theorem is

$$\Phi^{\Sigma}(t, t_0, 0, \mu) = \sum_{j=0}^{(t-t_0-\Delta)/\Delta} \Phi^{\mathrm{d}}_{\mathsf{A}, t_0+(j+1)\Delta}(t) \circ \mathsf{B}(t_0 + j\Delta)(\mu(t_0 + j\Delta)) = \Phi^{\mathrm{d}}_{\mathsf{A}, t_0+\Delta}(t) \circ \mathsf{B}(t_0)(u).$$

The theorem then follows immediately from definitions and conventions for summation. ■

It is also easy, since we do not have to work with distributions, to see that the (non-proper) impulse transmission map has in interpretation analogous to the preceding theorem. Indeed, if ξ_{t_0} is the solution to the initial value problem

$$\xi_{t_0}(t + \Delta) = \mathsf{A}(t) + \mathsf{B}(t)(\tau_{t_0}^* \mathsf{P}(t)u), \quad \xi_{t_0}(t_0) = 0,$$

i.e., the state response to the input $(\tau_{t_0}^* \mathsf{P})u$, then the corresponding output is

$$\eta(t) = \mathsf{C}(t)\xi_{t_0}(t) + \mathsf{D}(t)(\tau_{t_0}^* \mathsf{P}(t)u) = \mathrm{pitm}_{\Sigma, t_0}(t)(u) + \mathsf{D}(t_0)(u) = \mathrm{itm}_{\Sigma, t_0}(t)(u).$$

Thus the initial transmission map for Σ at t_0 is the output for zero initial condition at t_0 corresponding to the pulse input of u at time t_0.

The next result follows immediately from the definition of the impulse transmission map and the formula (6.9).

6.8.12 Proposition (Using the impulse transmission map to determine outputs) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system and let $\mu \in \mathscr{U}$. Then the output corresponding to an initial condition x_0 at time t_0 is

$$\eta(\mathrm{t}) = \Phi^{\mathrm{d}}_{\mathsf{A}, \mathrm{t}_0}(\mathrm{t})(\mathrm{x}_0) + \sum_{\mathrm{j}=0}^{(\mathrm{t}-\mathrm{t}_0-\Delta)/\Delta} \mathrm{pitm}_{\Sigma, \mathrm{t}_0+\mathrm{j}\Delta}(\mathrm{t})(\mu(\mathrm{t}_0+\mathrm{j}\Delta)) + \mathsf{D}(\mathrm{t})(\mu(\mathrm{t})), \qquad \mathrm{t} \in \mathrm{dom}(\mu)_{\geq \mathrm{t}_0}.$$

The punchline of the result is that the output is a linear combination of three terms:

$$\underbrace{\mathsf{C}(t) \circ \Phi^{\mathrm{d}}_{\mathsf{A}}(t, t_0)(x_0)}_{\text{term 1}} + \underbrace{\sum_{j=0}^{(t-t_0-\Delta)/\Delta} \mathrm{pitm}_{\Sigma, t_0+j\Delta}(t)\mu(t_0 + j\Delta)}_{\text{term 2}} + \underbrace{\mathsf{D}(t)(\mu(t))}_{\text{term 3}}. \tag{6.14}$$

Let us describe these terms, intuitively, just as we have already done for continuous-time systems.

1. The first term is the contribution from a nonzero initial provides the contribution to the output from the nonzero initial condition x_0 at time t_0.

2. The second term has the most complex interpretation. First of all, by Theorem 6.8.11, the summand $\mathrm{pitm}_{\Sigma,t_0+j\Delta}(tq)\mu(t_0+j\Delta)$ is the output obtained from the input $(\tau^*_{t_0+j\Delta}\mathsf{P})u$, i.e., a pulse of $\mu(t_0+j\Delta)$ at time $t_0+j\Delta$. The second terms can then be thought of as the sum of these contributions as $t_0+j\Delta$ goes from t_0 to $t-\Delta$.
3. The third term simply arises from the direct transmission from input to output determined by D.

6.8.3.2 The constant coefficient case The preceding constructions simplify substantially in the case of constant coefficient systems. This is fortunate, since it is this case that we will examine with respect to transform methods in Chapters 7 and 8. Let us record the simplifications.

The definition we make is the following. For systems with constant coefficients, there is no reason to not take the time-domain to be $\mathbb{Z}(\Delta)$, and so we do so.

6.8.13 Definition (Impulse response) Let

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{Z}(\Delta)$.

(i) The ***proper impulse response*** for Σ is the function

$$\begin{aligned}\mathrm{pir}_\Sigma \colon \mathbb{Z}(\Delta) &\to \mathrm{L}(\mathsf{U};\mathsf{Y})\\ t &\mapsto 1_{\geq 0}(t-\Delta)\mathsf{C}\circ\mathsf{P}_\mathsf{A}\left(\tfrac{t-\Delta}{\Delta}\right)\circ\mathsf{B}.\end{aligned}$$

(ii) The ***impulse response*** for Σ is the function

$$\begin{aligned}\mathrm{ir}_\Sigma \colon \mathbb{Z}(\Delta) &\to \mathrm{L}(\mathsf{U};\mathsf{Y})\\ t &\mapsto \mathrm{pir}_\Sigma(t) + \mathsf{P}(t)\mathsf{D}.\end{aligned}$$

•

The connection between the impulse response and the impulse transmission map is given by the following result.

6.8.14 Proposition (The impulse response and the impulse transmission map) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{Z}(\Delta)$. *Then the proper impulse transmission map is given by*

$$\mathrm{pitm}_{\Sigma,\tau}(\mathrm{t}) = \mathrm{pir}_\Sigma(\mathrm{t}-\tau), \qquad \mathrm{t},\tau \in \mathbb{Z}(\Delta),\ \mathrm{t}\geq\tau.$$

Proof This follows from the definitions, and the fact that, for A being independent of time, we have $\Phi^{\mathrm{d}}_\mathsf{A}(t,\tau) = \mathsf{P}_\mathsf{A}\left(\frac{t-\tau}{\Delta}\right)$ by definition. ■

Thus everything we said about the impulse transmission map above can be translated into a statement about the impulse response in the constant coefficient case. However, since there is more that can be said, and what can be said can be said more simply, let us record these translations explicitly.

We begin by giving the impulse response as a solution to an initial value problem.

6.8.15 Theorem (An interpretation of the proper impulse response) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{Z}(\Delta)$. *Then*

$$\mathrm{pir}_\Sigma(t)(u) = 1_{\geq 0}(t - \Delta)\mathsf{C} \circ \xi_0(t),$$

where $\xi_0 \colon \mathbb{Z}(\Delta) \to \mathsf{X}$ *is the unique solution of the initial value problem*

$$\xi_0(t + \Delta) = \mathsf{A}(\xi_0(t)) + \mathsf{B}(\mathsf{P}(t)u), \qquad \xi(0) = 0.$$

Proof This follows immediately from Theorem 6.8.11. ∎

The theorem tells us that, when $\mathsf{D} = 0$, the proper impulse response if the "output" corresponding to an "input" $\mathsf{P}u$, i.e., a pulse of u at time 0. Similarly to the case of the impulse transmission map, it also holds that the (non-proper) impulse response is the output for this same input, when D is not necessarily zero.

Of course, just as in Proposition 6.8.12, we can use the impulse response to characterise outputs for linear discrete-time state space systems with constant coefficients.

6.8.16 Proposition (Using the impulse response to determine outputs) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{Z}(\Delta)$ *and let* $\mu \in \mathscr{U}$. *Then the output corresponding to the initial condition* x_0 *at time* t_0 *is*

$$\eta(t) = \mathsf{P}_\mathsf{A}\left(\tfrac{t-t_0}{\Delta}\right)(x_0) + \sum_{j=0}^{(t-t_0-\Delta)/\Delta} \mathrm{pir}_\Sigma(t - t_0 - j\Delta)(\mu(t_0 + j\Delta)) + \mathsf{D} \circ \mu(t), \quad t \in \mathrm{dom}(\mu)_{\geq t_0}.$$

Proof This is a direct translation of Proposition 6.8.12 to the constant coefficient case. ∎

The interpretation we give for the three terms in the output are the same as we gave after the statement of Proposition 6.8.12. However, in the constant coefficient case, the time $t_0 = 0$ is distinguished, and this gives rise to distinguishing this particular case.

6.8.17 Definition (Zero-state/zero-time response) Let

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients and with $\mathbb{T} = \mathbb{Z}(\Delta)$ and let $\mu \in \mathscr{U}$ with $\operatorname{supp}(\mu) = \mathbb{Z}_{\geq 0}(\Delta)_{\geq 0}$. The ***zero-state/zero-time response*** to the input μ is

$$\begin{aligned} \zeta_\mu \colon \mathbb{Z}(\Delta) &\to \mathsf{Y} \\ t &\mapsto \sum_{j=0}^{(t-\Delta)/\Delta} \operatorname{pir}_\Sigma(t - j\Delta)\mu(j\Delta). \end{aligned}$$

•

These sorts of considerations will be considered in detail in Section 6.9.6.2.

Exercises

6.8.1 Consider a linear discrete-time state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}).$$

Answer the following questions.

(a) Show that Σ defines a general input/output system as per Definition 2.1.3. Identify the components of the general input/output system.
(b) Show that Σ is a general time system as per Definition 2.2.9. Identify the components of the general time system.
(c) Show that, as a general time system, Σ is output complete.
(d) Show that, as a general time system, Σ is complete.
(e) Show that Σ is causal and is strongly causal if and only if $\mathsf{D} = 0$.
(f) Show that Σ has a dynamical systems representation as per Definition 2.2.19. Identify components of the dynamical systems representation.
(g) Show that Σ has a state space representation as per Definition 2.2.24. Identify components of the state space representation.

6.8.2 For the following linear discrete-time state space systems

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}),$$

compute the proper impulse transmission map (or proper impulse response, as appropriate) and the impulse transmission map (or impulse response, as appropriate).

(a) Take

(i) $\mathsf{X} = \mathbb{R}$,
(ii) $\mathsf{U} = \mathbb{R}$,
(iii) $\mathsf{Y} = \mathbb{R}$,
(iv) $\mathbb{T} = \mathbb{Z}$,
(v) $\mathscr{U} = \ell_{\text{loc}}(\mathbb{Z}; \mathbb{R})$,
(vi) $\boldsymbol{A} = \begin{bmatrix} 0 \end{bmatrix}$,
(vii) $\boldsymbol{B} = \begin{bmatrix} 1 \end{bmatrix}$,
(viii) $\boldsymbol{C} = \begin{bmatrix} 1 \end{bmatrix}$,
(ix) $\boldsymbol{D} = \begin{bmatrix} 1 \end{bmatrix}$,

(b) Take
(i) $\mathsf{X} = \mathbb{R}^2$,
(ii) $\mathsf{U} = \mathbb{R}$,
(iii) $\mathsf{Y} = \mathbb{R}^2$,
(iv) $\mathbb{T} = \mathbb{Z}$,
(v) $\mathscr{U} = \ell_{\text{loc}}(\mathbb{Z}; \mathbb{R})$,
(vi) $\boldsymbol{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$,
(vii) $\boldsymbol{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,
(viii) $\boldsymbol{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$,
(ix) $\boldsymbol{D} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,

(c) Take
(i) $\mathsf{X} = \mathbb{R}^3$,
(ii) $\mathsf{U} = \mathbb{R}^2$,
(iii) $\mathsf{Y} = \mathbb{R}^2$,
(iv) $\mathbb{T} = \mathbb{Z}$,
(v) $\mathscr{U} = \ell_{\text{loc}}(\mathbb{Z}; \mathbb{R})$,
(vi) $\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$,
(vii) $\boldsymbol{B} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$,
(viii) $\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & -2 & 0 \end{bmatrix}$,
(ix) $\boldsymbol{D} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$,

(d) Take
(i) $\mathsf{X} = \mathbb{R}$,
(ii) $\mathsf{U} = \mathbb{R}$,
(iii) $\mathsf{Y} = \mathbb{R}$,
(iv) $\mathbb{T} = \mathbb{Z}$,
(v) $\mathscr{U} = \ell_{\text{loc}}(\mathbb{Z}; \mathbb{R})$,
(vi) $\boldsymbol{A} = \begin{bmatrix} -a(t) \end{bmatrix}$,
(vii) $\boldsymbol{B} = \begin{bmatrix} 1 \end{bmatrix}$,
(viii) $\boldsymbol{C} = \begin{bmatrix} 1 \end{bmatrix}$,
(ix) $\boldsymbol{D} = \begin{bmatrix} 0 \end{bmatrix}$,

Hint: *Refer to Example 4.6.5.*

6.8.3 Let

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system. Let $\Delta \in \mathbb{R}_{>0}$. Define

$$\begin{aligned}
\mathsf{A}_{\text{disc}}(k\Delta) &= \Phi^{\text{c}}_{\mathsf{A}}((k+1)\Delta, k\Delta),\\
\mathsf{B}_{\text{disc}}(k\Delta) &= \mathsf{B}(k\Delta),\\
\mathsf{C}_{\text{disc}}(k\Delta) &= \mathsf{C}(k\Delta),\\
\mathsf{D}_{\text{disc}}(k\Delta) &= \mathsf{D}(k\Delta),\\
\mathscr{U}_{\text{disc}} &= \{\mu_{\text{disc}} \colon \operatorname{dom}(\mu) \cap \mathbb{Z}(\Delta) \to \mathsf{U} \mid \mu_{\text{disc}}(k\Delta) = \mu(k\Delta),\ \mu \in \mathscr{U}\},
\end{aligned}$$

for $k \in \mathbb{Z}$. Define a linear discrete-time state space system by

$$\Sigma_{\text{disc}} = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{Z}(\Delta), \mathscr{U}_{\text{disc}}, \mathsf{A}_{\text{disc}}, \mathsf{B}_{\text{disc}}, \mathsf{C}_{\text{disc}}, \mathsf{D}_{\text{disc}})$$

and answer the following questions.

(a) Explain what it means for $(\eta_{\text{disc}}, \mu_{\text{disc}}) \in \text{Cout}(\Sigma_{\text{disc}})$ to be a sampled controlled output for some controlled output of Σ.

(b) Give conditions on $\mathscr{U}$ that ensure that every controlled output of $\mathscr{U}_{\text{disc}}$ is a sampled controlled output for some controlled output of Σ.

(c) Show that, if Σ is a linear continuous-time state space system with constant coefficients, then Σ_{disc} is a linear discrete-time state space system with constant coefficients.

(d) Is it true that, if

$$\Sigma'_{\text{disc}} = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{Z}(\Delta), \mathscr{U}'_{\text{disc}}, \mathsf{A}'_{\text{disc}}, \mathsf{B}'_{\text{disc}}, \mathsf{C}'_{\text{disc}}, \mathsf{D}'_{\text{disc}})$$

is a linear discrete-time state space system, then there exists a linear continuous state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

such that $\Sigma'_{\text{disc}} = \Sigma_{\text{disc}}$?

6.8.4 Consider the following difference equation for functions $\eta, \mu \colon \mathbb{Z}(\Delta) \to \mathbb{F}$:

$$\begin{aligned}\eta(t + n\Delta) + p_{n-1}\eta(t + (n-1)\Delta) + \cdots + p_1\eta(t + \Delta) + p_0\eta(t)\\ = c_{n-1}\mu(t + (n-1)\Delta) + c_{n-2}(t + (n-2)\Delta) + \cdots + c_1\mu(t + \Delta) + c_0\mu(t).\end{aligned}$$

Answer the following questions.

(a) Show that this determines a general time system as per Definition 2.2.9. Clearly identify the spaces of input and output signals.

(b) Argue that a natural choice of states for this system is

$$\xi_j(t) = \eta(t + j\Delta), \qquad j \in \{0, 1, \ldots, n-1\}.$$

(c) Argue that a somewhat less natural, but still valid, choice of states is $\xi_n(t) = \eta(t)$ and

$$\xi_{n-j}(t) = \sum_{k=0}^{j} p_{n-j+k}\eta(t + k\Delta) - \sum_{k=0}^{j-1} c_{n-j+k}\mu(t + k\Delta), \qquad j \in \{1, \ldots, n-1\}.$$

(d) Derive a linear discrete-time state space system with constant coefficients for which the input/output relation is the same as the general time system from part (a) and for which the states are as in part (c).

Section 6.9

Linear discrete-time input/output systems

The final class of systems we consider in this chapter are linear discrete-time input/output systems. The structure of this section follows the continuous-time case, with an initial consideration of general systems, and then specialisation to systems defined by kernels, first summation kernels and then convolution kernels.

Do I need to read this section? As with the preceding three sections, the material in this section is to be regarded as a core part of the material in this volume. •

6.9.1 General definitions

We being by considering a general setting for linear discrete-time input/output systems. The essential definition, which follows, is basically the Definition 6.4.3 for discrete-time input/output systems, with the addition of linearity. This requires linearity for both the spaces of input and output signals, and of the system mappings.

6.9.1 Definition (Linear discrete-time input/output system) A ***linear discrete-time input/output system*** is a quintuple $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, g)$, where

(i) U (the ***input space***) and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ is a discrete time-domain (the ***time-domain***),

(iii) $\mathscr{U} \subseteq \mathsf{U}^{(\mathbb{T})}$ is a space of partially defined signals with topology (the ***input signals***) such that, for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, $\mathscr{U}(\mathbb{S})$ is a subspace of $\mathsf{U}^{\mathbb{S}}$,

(iv) $\mathscr{Y} \subseteq \mathsf{Y}^{(\mathbb{T})}$ is a space of partially defined signals with topology (the ***output signals***) such that, for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, $\mathscr{Y}(\mathbb{S})$ is a subspace of $\mathsf{Y}^{\mathbb{S}}$, and

(v) $g \colon \mathscr{U} \to \mathscr{Y}$ has the following properties:

  (a) for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$, the restriction of g to $\mathscr{U}(\mathbb{S})$, denoted by $g_{\mathbb{S}}$, takes values in $\mathscr{Y}(\mathbb{S})$;

  (b) if $\mathbb{S}, \mathbb{S}' \subseteq \mathbb{T}$ are sub-time-domains with $\mathbb{S}' \subseteq \mathbb{S}$, then $g_{\mathbb{S}}|\mathscr{U}(\mathbb{S}') = g_{\mathbb{S}'}$;

  (c) $g_{\mathbb{S}}$ is linear and continuous for every sub-time-domain $\mathbb{S} \subseteq \mathbb{T}$.

Moreover,

(vi) a pair (μ, η) with $\mu \in \mathscr{U}(\mathbb{S})$ and $\eta = g_{\mathbb{S}}(\mu)$ is a ***behaviour*** for Σ, and we denote by $\mathscr{B}(\Sigma)$ the set of behaviours. •

Of course, linear discrete-time input/output systems are discrete-time input/output systems, accepting the mild generalisation from using general finite-dimensional vector spaces in place of Euclidean spaces. Thus all of the comments made in Section 6.4.2 about the connections between discrete-time input/output systems and the general classes of systems from Chapter 2 are applicable to linear discrete-time input/output systems. In addition, linear discrete-time input/output systems are linear time systems as per Definition 2.2.43. Note that, due to the fact that we work with spaces of input and output signals that are comprised of partially defined signals, it is not generally the case that a linear discrete-time input/output system is a linear general input/output system as per Definition 2.1.12. However, were we to restrict to the case of signals only defined on the entire time-domain $\mathbb{T}$, i.e., to not allow partially defined signals, then such a linear discrete-time input/output system would indeed be a linear general input/output system as per Definition 2.1.12.

6.9.2 Summation kernel systems

We now consider a special class of linear discrete-time input/output systems. As we shall assert precisely in Section 6.9.6, the class of systems we consider generalise aspects of the input/output behaviour of a linear discrete-time state space system.

The initial ingredient to the constructions we make is contained in the following definition.

6.9.2 Definition (Summation kernel, summation operator) Let $\mathbb{T} \subseteq \mathbb{R}$ be a discrete time-domain, and let U and Y be finite-dimensional $\mathbb{R}$-vector spaces.

(i) A ***summation kernel*** from U to Y on $\mathbb{T}$ is a mapping

$$\mathsf{K} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y}).$$

For $t, \tau \in \mathbb{T}$, we shall denote

$$\begin{aligned} \mathsf{K}_t \colon \mathbb{T} &\to \mathrm{L}(\mathsf{U}; \mathsf{Y}) & \qquad \mathsf{K}^\tau \colon \mathbb{T} &\to \mathrm{L}(\mathsf{U}; \mathsf{Y}) \\ \tau &\mapsto \mathsf{K}(t, \tau) & \qquad t &\mapsto \mathsf{K}(t, \tau). \end{aligned}$$

Let $\mathscr{U} \subseteq \mathsf{U}^{\mathbb{T}}$ be a subspace.

(ii) A summation kernel K is ***compatible*** with $\mathscr{U}$ if, for every $\mu \in \mathscr{U}$ and for every $t \in \mathbb{T}$, $\mathsf{K}_t(\mu) \in \ell^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$.

(iii) If K is compatible with $\mathscr{U}$, the ***summation operator*** defined by K is the mapping

$$g_{\mathsf{K}} \colon \mathscr{U} \to \mathsf{Y}^{\mathbb{T}}$$

defined by

$$g_{\mathsf{K}}(\mu)(t) = \sum_{\tau \in \mathbb{T}} \mathsf{K}(t, \tau)(\mu(\tau)), \qquad t \in \mathbb{T}. \qquad \bullet$$

As yet, we do not have the structure of a discrete-time input/output system, since the domain and codomain of g_{K} do not have useful structure (other than their vector space structure). We need to provide conditions on K that ensure that the summation operator g_{K} takes signals from a nice domain into a nice codomain. The following definition captures the properties we want.

6.9.3 Definition (Summation kernel system) A ***summation kernel system*** is a sextuple $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathsf{K})$ where

(i) U (the ***input space***) and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ is a discrete time-domain (the ***time-domain***),

(iii) $\mathscr{U} \subseteq \mathsf{U}^{\mathbb{T}}$ is a subspace of signals with topology (the ***input signals***),

(iv) $\mathscr{Y} \subseteq \mathsf{Y}^{\mathbb{T}}$ is a subspace of signals with topology (the ***output signals***),

(v) K is a summation kernel compatible with $\mathscr{U}$, and

(vi) the summation operator g_{K} is a continuous linear mapping from $\mathscr{U}$ to $\mathscr{Y}$. •

Of course, the definition gives us no insight into which K's, $\mathscr{U}$'s, and $\mathscr{Y}$'s might possibly comprise a discrete-time kernel system. In order to obtain characterisations which give such systems, we have to prove something, and the following result gives some cases that work.

6.9.4 Theorem (Some summation kernel systems) *Let $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ be a discrete time-domain, let U and Y be finite-dimensional $\mathbb{R}$-vector spaces, and let $\mathrm{p} \in [1, \infty]$. In the following cases, the summation kernel $\mathsf{K} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$, the input space $\mathscr{U}$, and the output space $\mathscr{Y}$ are such that*

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathsf{K})$$

is a summation kernel system:

(i) (a) $\mathsf{K}_{\mathrm{t}} \in \ell^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_1$ *is in* $\ell^\infty(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,
(b) $\mathscr{U} \subseteq \ell^\infty(\mathbb{T}; \mathsf{U})$, *and*
(c) $\mathscr{Y} = \ell^\infty(\mathbb{T}; \mathsf{Y})$;

(ii) (a) $\mathsf{K}_{\mathrm{t}} \in \ell^\infty(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_\infty$ *is in* $\ell^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,
(b) $\mathscr{U} \subseteq \ell^1(\mathbb{T}; \mathsf{U})$, *and*
(c) $\mathscr{Y} = \ell^1(\mathbb{T}; \mathsf{Y})$;

(iii) (a) $\mathsf{K}_{\mathrm{t}} \in \ell^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_1$ *is in* $\ell^\infty(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,
(b) $\mathsf{K}_{\mathrm{t}} \in \ell^\infty(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$ *for each* $\mathrm{t} \in \mathbb{T}$ *and* $\mathrm{t} \mapsto \|\mathsf{K}_{\mathrm{t}}\|_\infty$ *is in* $\ell^1(\mathbb{T}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,
(c) $\mathscr{U} \subseteq \ell^{\mathrm{p}}(\mathbb{T}; \mathsf{U})$, *and*
(d) $\mathscr{Y} = \ell^{\mathrm{p}}(\mathbb{T}; \mathsf{Y})$.

Proof (i) First of all, for $t \in \mathbb{T}$,

$$\sum_{\tau\in\mathbb{T}} \|\mathsf{K}(t,\tau)\mu(\tau)\| \leq \|\mu\|_\infty \sum_{\tau\in\mathbb{T}} \|\mathsf{K}_t\| < \infty,$$

giving the compatibility of K with $\mathscr{U}$ in this case. Also by Exercise III-3.8.2,

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)\|_\infty &= \sup\left\{ \left\| \sum_{\tau\in\mathbb{T}} \mathsf{K}(t,\tau)(\mu(\tau)) \right\| \;\middle|\; t \in \mathbb{T} \right\} \\
&\leq \sup\left\{ \sum_{\tau\in\mathbb{T}} \left\| \mathsf{K}(t,\tau)(\mu(\tau)) \right\| \;\middle|\; t \in \mathbb{T} \right\} \\
&\leq \sup\left\{ \|\mu\|_\infty \sum_{\tau\in\mathbb{T}} \|\mathsf{K}(t,\tau)\| \;\middle|\; t \in \mathbb{T} \right\} \\
&\leq \underbrace{\sup\{\|\mathsf{K}_t\|_1 \mid t \in \mathbb{T}\}}_{C_\infty} \|\mu\|_\infty.
\end{aligned}$$

Thus $\|g_{\mathsf{K}}(\mu)\|_\infty \leq C_\infty \|\mu\|_\infty$, giving continuity of g_{K} by Theorem III-3.5.8.

(ii) We first have, for $t \in \mathbb{T}$,

$$\sum_{\tau\in\mathbb{T}} \|\mathsf{K}(t,\tau)\mu(\tau)\| \leq \|\mathsf{K}_t\|_\infty \sum_{\tau\in\mathbb{T}} \|\mu(\tau)\| < \infty,$$

giving the compatibility of K with $\mathscr{U}$. We also have

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)\|_1 &= \sum_{t\in\mathbb{T}} \left\| \sum_{\tau\in\mathbb{T}} \mathsf{K}(t,\tau)(\mu(\tau)) \right\| \\
&\leq \sum_{t\in\mathbb{T}} \left(\sum_{\tau\in\mathbb{T}} \|\mathsf{K}(t,\tau)(\mu(\tau))\| \right) \\
&= \sum_{\tau\in\mathbb{T}} \left(\sum_{t\in\mathbb{T}} \|\mathsf{K}(t,\tau)(\mu(\tau))\| \right) \\
&\leq \sum_{\tau\in\mathbb{T}} \left(\sum_{t\in\mathbb{T}} \|\mathsf{K}(t,\tau)\| \right) \|\mu(\tau)\| \\
&\leq \underbrace{\left(\sum_{t\in\mathbb{T}} \|\mathsf{K}_t\|_\infty \right)}_{C_1} \left(\sum_{\tau\in\mathbb{T}} \|\mu(\tau)\| \right)
\end{aligned}$$

using Fubini's Theorem. Thus $\|g_{\mathsf{K}}(\mu)\|_1 \leq C_1 \|\mu\|_1$, and we get this part of the theorem by Theorem III-3.5.8.

(iii) Clearly we can restrict ourselves to $p \in (1,\infty)$. Thus we take $p' \in (1,\infty)$ to be the conjugate index for which $\frac{1}{p} + \frac{1}{p'} = 1$. Let C_∞ and C_1 be as defined in the first two parts of the proof.

To determine the compatibility of K with μ, for $\mu \in \ell^p(\mathbb{T};\mathsf{U})$, write

$$\mu_0(t) = \begin{cases} \mu(t), & \|\mu(t)\| \leq 1, \\ 0, & \text{otherwise}, \end{cases}$$

and $\mu_1 = \mu - \mu_0$. Note that $\mu_0 \in \ell^\infty(\mathbb{T};\mathsf{U})$ and $\mu_1 \in \ell^p(\mathbb{T};\mathsf{U})$. Moreover,

$$\|\mu_1(t)\| \leq \|\mu_1(t)\|^p, \qquad t \in \mathbb{T},$$

and so $\mu_1 \in \ell^1(\mathbb{T};\mathsf{U})$. One can then combine the compatibility conclusions from the first two parts of the proof to conclude that K is compatible with $\ell^p(\mathbb{T};\mathsf{U})$.

We now compute, using Hölder's inequality in the form of Lemma III-3.8.16,

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)(t)\| &\leq \sum_{\tau\in\mathbb{T}} \|\mathsf{K}(t,\tau)(\mu(\tau))\| \\
&\leq \sum_{\tau\in\mathbb{T}} (\|\mathsf{K}(t,\tau)\|^{1/p}\|\mu(\tau)\|)\|\mathsf{K}(t,\tau)\|^{1/p'} \\
&\leq \left(\sum_{\tau\in\mathbb{T}} \|\mathsf{K}(t,\tau)\|\|\mu(\tau)\|^p\right)^{1/p} \left(\sum_{\tau\in\mathbb{T}} \|\mathsf{K}(t,\tau)\|\right)^{1/p'} \\
&\leq C_\infty^{1/p'} \left(\sum_{\tau\in\mathbb{T}} \|\mathsf{K}(t,\tau)\|\|\mu(\tau)\|^p\right)^{1/p}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|g_{\mathsf{K}}(\mu)\|_p^p &\leq C_\infty^{p/p'} \sum_{\tau\in\mathbb{T}} \left(\sum_{t\in\mathbb{T}} \|\mathsf{K}(t,\tau)\|\|\mu(\tau)\|^p\right) \\
&\leq C_\infty^{p/p'} \sum_{\tau\in\mathbb{T}} \left(\sum_{t\in\mathbb{T}} \|\mathsf{K}_t\|_\infty \|\mu(\tau)\|^p\right) \\
&\leq C_\infty^{p/p'} \left(\sum_{\tau\in\mathbb{T}} \|\mu(\tau)\|^p\right)\left(\sum_{t\in\mathbb{T}} \|\mathsf{K}_t\|_\infty\right) \\
&\leq C_\infty^{p/p'} C_1 \|\mu\|_p^p.
\end{aligned}$$

Thus we have

$$\|g_{\mathsf{K}}(\mu)\|_p \leq C_1^{1/p} C_\infty^{1/p'} \|\mu\|_p,$$

giving the result, again using Theorem III-3.5.8. ■

The preceding result, while interesting, is limited in scope. Indeed, it has nothing to say about systems that take $\ell^p_{\text{loc}}(\mathbb{T};\mathsf{U})$ to $\ell^q_{\text{loc}}(\mathbb{T};\mathsf{Y})$. The restriction in Theorem 6.9.4 to input and output spaces that are ℓ^p-spaces has more to do with the stability of the systems than with their general system theoretic attributes. However, to overcome these limitations in systematic way requires putting some general restrictions on the kernel K and/or the input signals $\mathscr{U}$. One nice class of kernels are those that give rise to causal systems. Let us define the class of kernels in this case.

6.9.5 Definition (Causal summation kernel) Let $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ be a discrete time-domain, and let U and Y be finite-dimensional $\mathbb{R}$-vector spaces. A summation kernel

$$\mathsf{K} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$$

is ***causal*** if $\mathsf{K}(t, \tau) = 0$ for $\tau > t$. •

Let us relate this notion of a causal summation kernel to a causal system.

6.9.6 Lemma (Causal summation kernels give rise to causal summation kernel systems) *Let $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ be a discrete time-domain, let U and Y be finite-dimensional $\mathbb{R}$-vector spaces, and let $\mathscr{U}$ be a set of input signals. If K is a causal summation kernel compatible with $\mathscr{U}$, then the discrete-time input/output system $\mathrm{g}_{\mathsf{K}} \colon \mathscr{U} \to \mathsf{Y}^{\mathbb{T}}$ is causal.*

Proof Let $\mu_1, \mu_2 \in \mathscr{U}$ satisfy $\mathrm{dom}(\mu_1) = \mathrm{dom}(\mu_2)$ and let $t \in \mathrm{dom}(\mu_1) = \mathrm{dom}(\mu_2)$. Suppose that $(\mu_1)_{\mathbb{T}_{\leq t} \cap \mathrm{dom}(\mu_1)} = (\mu_2)_{\mathbb{T}_{\leq t} \cap \mathrm{dom}(\mu_2)}$. Then

$$\begin{aligned} g_{\mathsf{K}}(\mu_1)(t) &= \sum_{\tau \in \mathbb{T}} \mathsf{K}(t, \tau)(\mu_1(\tau)) = \sum_{\tau \in \mathbb{T}_{\leq t}} \mathsf{K}(t, \tau)(\mu_1(\tau)) \\ &= \sum_{\tau \in \mathbb{T}_{\leq t}} \mathsf{K}(t, \tau)(\mu_2(\tau)) = \sum_{\tau \in \mathbb{T}} \mathsf{K}(t, \tau)(\mu_2(\tau)) = g_{\mathsf{K}}(\mu_2)(t). \end{aligned}$$

This is the desired causality. ■

One might like to have the causality of the kernel as being necessary for the causality of the associated summation operator. However, to state a general such theorem requires having some relationship between the kernel and the set of inputs that will just be confusing. The basic idea, however, is clear. If the summation kernel is *not* causal, then there will be some $t \in \mathbb{T}$ and an interval $\mathbb{S} \subseteq \mathbb{T}_{>t}$ such that

$$\sum_{\tau \in \mathbb{S}} \|\mathsf{K}(t, \tau)\| \neq 0.$$

Generally speaking, one can expect there to be an input μ for which

$$\sum_{\tau \in \mathbb{S}} \mathsf{K}(t, \tau)(\mu(\tau)) \neq 0.$$

If one can additionally ask that $\mathrm{supp}(\mu) \subseteq \mathbb{S}$, then we would have $g_{\mathsf{K}}(\mu)(t) \neq 0$, even though μ is zero up to time t. This would preclude causality.

With the above considerations at hand, let us consider situations where a causal summation kernel defines a summation kernel system. As we see, the condition of causality of the summation kernel, as well as the causality of the set of input signals as in Definition IV-1.1.16, ensures causality of the system. Note that the result we state here is simpler than the continuous-time version, Theorem 6.7.7, because there is no distinction between the spaces $\ell^p_{\mathrm{loc}}(\mathbb{T}; \mathsf{V})$, $p \in [1, \infty]$, cf. Section IV-1.2.5. Thus we restrict ourselves to consideration of $\ell_{\mathrm{loc}}(\mathbb{T}; \mathsf{V}) = \mathsf{V}^{\mathbb{T}}$.

6.9.7 Theorem (Summation kernel systems with causal kernels and causal inputs) *Let $\mathbb{T} \subseteq \mathbb{Z}(\Delta)$ be a discrete time-domain, let U and Y be finite-dimensional $\mathbb{R}$-vector spaces, and let $p \in [1, \infty]$. Under the following hypotheses, the causal summation kernel $\mathsf{K} \colon \mathbb{T} \times \mathbb{T} \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$, the input space $\mathscr{U}$, and the output space $\mathscr{Y}$ are such that*

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \mathsf{K})$$

is a summation kernel system:

(i) $\mathscr{U} \subseteq \ell_{\mathrm{loc}}(\mathbb{T}; \mathsf{U})$ and there exists $t_0 \in \mathbb{T}$ such that $\operatorname{supp}(\mu) \subseteq \mathbb{T}_{\geq t_0}$ for every $\mu \in \mathscr{U}$;

(ii) $\mathscr{Y} = \ell_{\mathrm{loc}}(\mathbb{T}; \mathsf{Y})$.

Proof By Lemma 6.9.6 and Theorem IV-4.2.47, K is compatible with $\mathscr{U}$. Theorem IV-4.2.47 also gives continuity of the map g_{K}. ■

6.9.3 How general are summation kernel systems?

In Section 6.7.3 we considered the matter of when a linear continuous-time input/output system is an integral kernel system. We saw that we were able to give some quite general conditions involving the Schwartz Kernel Theorem, Theorem IV-4.8.1. Here we shall address the same sort of questions for summation kernel systems, and we shall see that the answers are both more general and simpler.

6.9.8 Proposition (Linear discrete-time input/output systems that are summation kernel systems) *Consider a linear discrete-time input/output system*

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \ell_{\mathrm{loc}}(\mathbb{T}; \mathsf{U}), \ell_{\mathrm{loc}}(\mathbb{T}; \mathsf{Y}), \mathrm{g}).$$

Then there exists a summation kernel K such that $\mathrm{g} = \mathrm{g}_{\mathsf{K}}$.

Proof For $\tau \in \mathbb{T}$ and $u \in \mathsf{U}$, let $\mathsf{K}(t, \tau)(u) = g(\tau_\tau^* \mathsf{P} u)(t)$. If $\mu \in \ell_{\mathrm{loc}}(\mathbb{T}; \mathsf{U})$, then we can write

$$\mu = \sum_{\tau \in \mathbb{T}} \tau_\tau^* \mathsf{P} \mu(\tau).$$

Moreover,

$$\begin{aligned} g(\mu)(t) &= g\left(\sum_{\tau \in \mathbb{T}} \tau_\tau^* \mathsf{P} \mu(\tau) \right)(t) = \left(\sum_{\tau \in \mathbb{T}} g(\tau_\tau^* \mathsf{P} \mu(\tau)) \right)(t) \\ &= \sum_{\tau \in \mathbb{T}} \mathsf{K}(t, \tau) \mu(\tau) = g_{\mathsf{K}}(\mu)(t). \end{aligned}$$

where we move g inside the sum by continuity. ■

6.9.4 Discrete-time convolution systems

In this section we consider a special class of summation kernel systems. These arise from requiring stationarity of the summation kernel system, and the following result captures the manner in which stationarity arises. We focus on systems with time-domain $\mathbb{T} = \mathbb{Z}(\Delta)$, since this can be done without loss of generality in any case.

6.9.9 Proposition (Stationary summation kernel systems) *Let* U *and* Y *be finite-dimensional* $\mathbb{R}$*-vector spaces and let* $\mathsf{K}\colon \mathbb{Z}(\Delta) \times \mathbb{Z}(\Delta) \to \mathrm{L}(\mathsf{U};\mathsf{Y})$ *be an integral kernel compatible with a set* $\mathscr{U}$ *of input signals. Suppose that* $\mathscr{U}$ *is translation invariant, i.e., that* $\tau_{\mathrm{a}}^*\mu \in \mathscr{U}$ *for every* $\mathrm{a} \in \mathbb{Z}(\Delta)$ *and* $\mu \in \mathscr{U}$*. Denote by*

$$\Sigma_{\mathsf{K}} = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathsf{Y}^{\mathbb{Z}(\Delta)}, \mathbb{Z}(\Delta), \mathrm{g}_{\mathsf{K}})$$

the general time system. Then the following statements hold:

(i) if

(a) $\mathscr{U}$ *has the property that, if* $\mathrm{f} \in \ell_{\mathrm{loc}}(\mathbb{Z}(\Delta);\mathbb{R})$ *satisfies*

$$\sum_{\mathrm{t}\in\mathbb{Z}(\Delta)} \mathrm{f}(\mathrm{t})\mu(\mathrm{t}), \qquad \mu \in \mathscr{U},$$

then $\mathrm{f} = 0$*, and*

(b) Σ_{K} *is stationary,*

then there exists $\mathsf{k} \in \ell_{\mathrm{loc}}(\mathbb{Z}(\Delta);\mathrm{L}(\mathsf{U};\mathsf{Y}))$ *such that* $\mathsf{K}(\mathrm{t},\tau) = \mathsf{k}(\mathrm{t}-\tau)$ *for almost every* $(\mathrm{t},\tau) \in \mathbb{R}^2$*;*

(ii) if there exists $\mathsf{k} \in \ell_{\mathrm{loc}}(\mathbb{Z}(\Delta);\mathrm{L}(\mathsf{U};\mathsf{Y}))$ *such that* $\mathsf{K}(\mathrm{t},\tau) = \mathsf{k}(\mathrm{t}-\tau)$ *for every* $(\mathrm{t},\tau) \in \mathbb{Z}(\Delta)^2$*, then* Σ_{K} *is strongly stationary.*

Proof (i) Suppose that Σ_{K} is stationary. Then, for every $a \in \mathbb{Z}(\Delta)$ and for every behaviour (μ,η) for Σ_{K}, $(\tau_a^*\mu, \tau_a^*\eta)$ is also a behaviour. Note that this gives

$$\eta(t) = \sum_{\tau\in\mathbb{Z}(\Delta)} \mathsf{K}(t,\tau)(\mu(\tau))$$

and

$$\eta(t-a) = \sum_{\tau\in\mathbb{Z}(\Delta)} \mathsf{K}(t,\tau)(\mu(\tau-a))$$

for every $t \in \mathbb{Z}(\Delta)$. By a change of summation variable, the second of these equations becomes

$$\eta(t) = \sum_{\tau\in\mathbb{Z}(\Delta)} \mathsf{K}(t+a,\tau+a)(\mu(\tau)).$$

Thus we have

$$\sum_{\tau\in\mathbb{Z}(\Delta)} (\mathsf{K}(t,\tau) - \mathsf{K}(t+a,\tau+a))(\mu(\tau)) = 0$$

for every $t \in \mathbb{Z}(\Delta)$. Thus

$$\mathsf{K}(t,\tau) = \mathsf{K}(t+a,\tau+a), \qquad a \in \mathbb{Z}(\Delta),\ (t,\tau) \in \mathbb{Z}(\Delta)^2.$$

Therefore, for $(t,\tau) \in \mathbb{Z}(\Delta)^2$ we have, by taking $a = -\tau$, $\mathsf{K}(t,\tau) = \mathsf{K}(t-\tau,0)$. Therefore, taking $\mathsf{k}(t) = \mathsf{K}(t,0)$, we get the result.

(ii) We leave this to the reader as Exercise 6.9.2. ■

With this result in mind, we make the following definitions.

6.9.10 Definition (Convolution kernel, convolution operator defined by convolution kernel) Let U and Y be finite-dimensional $\mathbb{R}$-vector spaces.

(i) A ***discrete-time convolution kernel*** from U to Y is a mapping

$$\mathsf{k} \colon \mathbb{Z}(\Delta) \to \mathrm{L}(\mathsf{U}; \mathsf{Y}).$$

Let $\mathscr{U} \subseteq \mathsf{U}^{\mathbb{Z}(\Delta)}$ be a subspace.

(ii) A discrete-time convolution kernel k is ***compatible*** with $\mathscr{U}$ if, for every $\mu \in \mathscr{U}$ and for every $t \in \mathbb{Z}(\Delta)$, $\tau \mapsto \mathsf{k}(t - \tau) \circ \mu(\tau) \in \ell^1(\mathbb{Z}(\Delta); \mathrm{L}(\mathsf{U}; \mathsf{Y}))$.

(iii) If k is compatible with $\mathscr{U}$, the ***discrete-time convolution operator*** defined by k is the mapping

$$g_{\mathsf{k}} \colon \mathscr{U} \to \mathsf{Y}^{\mathbb{Z}(\Delta)}$$

defined by

$$g_{\mathsf{k}}(\mu)(t) = \sum_{\tau \in \mathbb{Z}(\Delta)} \mathsf{k}(t - \tau)(\mu(\tau)), \qquad t \in \mathbb{Z}(\Delta). \qquad \bullet$$

6.9.11 Definition (Discrete-time convolution system) A ***discrete-time convolution system*** is a quintuple $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ where

(i) U (the ***input space***) and Y (the ***output space***) are finite-dimensional $\mathbb{R}$-vector spaces,

(ii) $\mathscr{U} \subseteq \mathsf{U}^{\mathbb{Z}(\Delta)}$ is a subspace of signals with topology (the ***input signals***),

(iii) $\mathscr{Y} \subseteq \mathsf{Y}^{\mathbb{Z}(\Delta)}$ is a subspace of signals with topology (the ***output signals***),

(iv) $\mathsf{k} \in$ is a discrete-time convolution kernel compatible with $\mathscr{U}$, and

(v) the discrete-time convolution operator g_{k} is a continuous linear mapping from $\mathscr{U}$ to $\mathscr{Y}$. $\bullet$

Of course, a discrete-time convolution system

$$\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$$

is a summation kernel system

$$\Sigma' = (\mathsf{U}, \mathsf{Y}, \mathbb{Z}(\Delta), \mathscr{U}, \mathscr{Y}, \mathsf{K})$$

with $\mathsf{K}(t, \tau) = \mathsf{k}(t - \tau)$. Moreover, in Sections IV-4.2.7 and IV-4.2.8, we gave results about convolvable pairs of signals defined on $\mathbb{Z}(\Delta)$ that we can use here to give some specific instances of continuous-time convolution systems. We refer the reader to the above listed sections for precise results as reproducing these would be an unnecessary distraction.

As with Theorem 6.9.4, the results from Sections IV-4.2.7 and IV-4.2.8 are quite restrictive in that they apply only to signals that are summable in some sense, and this is a quite limited class of signals. This can be rectified, both mathematically and practically, by restricting to causal kernels and inputs. Based on Definition 6.9.5, we make the following definition.

6.9.12 Definition (Causal discrete-time convolution kernel) Let U and Y be finite-dimensional $\mathbb{R}$-vector spaces. A discrete-time convolution kernel

$$\mathsf{k} \colon \mathbb{Z}(\Delta) \to \mathrm{L}(\mathsf{U}; \mathsf{Y})$$

is ***causal*** if $\mathsf{k}(t) = 0$ for $t < 0$. •

We can then extend the applicability of the results from Sections IV-4.2.7 and IV-4.2.8 to causal convolution kernels, and with spaces of input and output signals that are only appropriately locally summable. In this respect, we refer the reader to Section IV-4.2.9 that provide some classes of convolution systems with causal kernels. Rather than reproduce all of the results from these sections in our specific setting here, let us simply indicate the steps one must take to adapt the results.

Let U and Y be finite-dimensional $\mathbb{R}$-vector spaces and let k be a causal convolution kernel residing in an appropriate space of locally summable signals. Suppose that $\mathscr{U}$ is a subset of an appropriate space of locally summable signals and that $t_0 \in \mathbb{Z}(\Delta)$ is such that $\mu(t) = 0$ for all $\mu \in \mathscr{U}$ and $t < t_0$. Let $\mathbb{K} \subseteq \mathbb{R}$ be a compact interval satisfying

$$\sup \mathbb{K} \geq \inf \operatorname{supp}(\mu),$$

noting that, when this is not true, then $\mathsf{k} * \mu|\mathbb{K} = 0$. Then, letting $\mathbb{L}$ satisfy

$$\begin{aligned} \inf \mathbb{L} &\leq \min\{0, t_0\}, \\ \sup \mathbb{L} &\geq \max\{\sup \mathbb{K}, \sup \mathbb{K} - t_0\}, \end{aligned}$$

we can use the appropriate variant of, for example, Theorem IV-4.2.47, to give continuity of the input/output g_{k}.

6.9.13 Remark (The "punchline" for discrete-time convolution systems) The technicalities of the results in this section may obscure the simple reasons why discrete-time convolution systems are important. Let us summarise these reasons.

1. Among the summation kernel systems, convolution systems are distinguished by being the stationary systems. This is the content of Proposition 6.9.9.
2. Causality for discrete-time convolution systems is easily characterised by the requirement that the convolution kernel vanish for negative time. Thus causal discrete-time convolution systems give a large and interesting class of causal stationary discrete-time linear systems.

6.9.5 How general are discrete-time convolution systems?

6.9.6 Linear discrete-time state space systems as linear discrete-time input/output systems

A merely mildly astute reader will have noticed that summation kernel system arise in the input/output relations for linear discrete-time state space systems, and

that this summation kernel description simplifies to a convolution system in the case where the state space system has constant coefficients. In this section we make these connections precise, which is comparatively easy given the work that we have done already.

6.9.6.1 The time-varying case Let us get straight to the point and state the main results of this section. We again emphasise the simplifications that arise in the discrete-time case where all of the ℓ^p_{loc}-topologies, $p \in [1, \infty]$, are the same.

6.9.14 Theorem (Summation kernel systems from linear discrete-time state space systems) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system. Let $t_0 \in \mathbb{T}$ *and let*

$$\begin{aligned}\mathscr{U} &\subseteq \{\mu \in \ell_{\text{loc}}(\mathbb{T}; \mathsf{U}) \mid \ \mu(t) = 0, \ t < t_0\},\\ \mathscr{Y} &= \{\eta \in \ell_{\text{loc}}(\mathbb{T}; \mathsf{Y}) \mid \ \eta(t) = 0, \ t < t_0\}.\end{aligned}$$

Then

$$\Sigma_{\text{i/o}}(t_0) = (\mathsf{U}, \mathsf{Y}, \mathbb{T}, \mathscr{U}, \mathscr{Y}, \text{pitm}_\Sigma)$$

is a causal summation kernel system. Moreover,

$$\Phi^\Sigma(t, t_0, x_0, \mu) = \Phi^{\text{d}}_{\mathsf{A}, t_0}(t)(x_0) + g_{\text{pitm}_{\Sigma, t_0}}(\mu)(t) + \mathsf{D}(t) \circ \mu(t).$$

Proof The causality of the integral kernel pitm_Σ ensures that, if $\mu(t) = 0$ for $t < t_0$, then $\text{pitm}_\Sigma * \mu(t) = 0$ for t_0. Continuity of the input/output map for the integral kernel pitm_Σ follows from Theorem 6.9.7. The final formula in the statement of the theorem follows from Proposition 6.8.12. ∎

We see from our decomposition (6.14) of an arbitrary output of a linear discrete-time state space system that the map sending an input μ to the second term in this decomposition is precisely the input/output system $\Sigma_{\text{i/o}}(t_0)$.

6.9.6.2 The constant coefficient case In this case, the main result is the following.

6.9.15 Theorem (Discrete-time convolution systems from linear discrete-time state space systems with constant coefficients) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{Z}(\Delta), \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients, Let

$$\begin{aligned}\mathscr{U} &\subseteq \{\mu \in \ell_{\text{loc}}(\mathbb{Z}(\Delta); \mathsf{U}) \mid \ \mu(t) = 0, \ t < 0\},\\ \mathscr{Y} &= \{\eta \in \ell_{\text{loc}}(\mathbb{Z}(\Delta); \mathsf{Y}) \mid \ \eta(t) = 0, \ t < 0\}.\end{aligned}$$

Then

$$\Sigma_{\mathrm{i/o}} = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathrm{pir}_{\Sigma})$$

is a causal discrete-time convolution system. Moreover,

$$\Phi^{\Sigma}(t, 0, x_0, \mu) = \mathsf{A}^{t/\Delta}(x_0) + g_{\mathrm{pir}_{\Sigma}} * \mu(t) + \mathsf{D}(t) \circ \mu(t).$$

Proof This follows from Theorem 6.9.14. ■

Here we see that the map sending an input to its zero-state/zero-input response, as in Definition 6.8.17, defines the input/output system $\Sigma_{\mathrm{i/o}}$.

6.9.7 Linear discrete-time difference input/output systems

Exercises

6.9.1 For each of the listed attributes, give an example of a linear discrete-time input/output system with that attribute. You are not allowed to choose a system of the sort considered in either of Sections 6.9.2 and 6.9.4.

Here are the attributes:

(a) causal;
(b) not causal;
(c) stationary;
(d) not stationary;
(e) memoryless;
(f) not memoryless.

6.9.2 Complete the proof of Proposition 6.9.9.

6.9.3 For $N \in \mathbb{Z}_{\geq 0}$, consider the function

$$\begin{aligned} d_N \colon \ell_{\mathrm{loc}}(\mathbb{Z}(\Delta); \mathbb{F}) &\to \ell_{\mathrm{loc}}(\mathbb{Z}(\Delta); \mathbb{F}) \\ \mu &\mapsto \tau^*_{N\Delta}\mu. \end{aligned}$$

Answer the following questions.

(a) Show that d_N is a linear discrete-time input/output system.
(b) Determine its system theoretic properties, i.e., is it causal? strongly causal? finitely observable? stationary? strongly stationary? memoryless?
(c) Let $k_0 \in \mathbb{Z}$. Show that, to determine $d_N(\mu)(k\Delta)$ for all $k \geq k_0$, you must know $\mu(k\Delta)$ for $k \geq k_0 - N$.
(d) Argue that the state space for the system starting from $k_0\Delta$ is

$$(\mu((k_0 - N)\Delta), \mu((k_0 - N + 1)\Delta), \ldots, \mu((k_0 - 1)\Delta)).$$

(e) Determine a linear discrete-time state space system whose input/output mapping is the same as d_N.

(f) Compare the situation in this discrete-time case with the continuous-time case presented in Exercise 6.7.3.

6.9.4 Consider the discrete time-domain $\mathbb{T} = \mathbb{Z}(\Delta)$ and the backward difference and forward difference maps as in Definition 3.3.2, but now with different notation to avoid a proliferation of Δ's. These are given by

$$\delta_- : \ell_{\text{loc}}(\mathbb{Z}(\Delta); \mathbb{F}) \to \ell_{\text{loc}}(\mathbb{Z}(\Delta); \mathbb{F}), \quad \delta_+ : \ell_{\text{loc}}(\mathbb{Z}(\Delta); \mathbb{F}) \to \ell_{\text{loc}}(\mathbb{Z}(\Delta); \mathbb{F})$$

with

$$\delta_-(\mu)(t) = (\mu(t) - \mu(t - \Delta)), \quad \delta_+(\mu)(t) = (\mu(t + \Delta) - \mu(t)).$$

Answer the following questions.

(a) Show that δ_- and δ_+ define discrete-time convolution systems and determine their convolution kernels.

(b) Are δ_- and δ_+ causal? strongly causal? memoryless?

6.9.5 A ***discrete-time sliding averager*** takes an input signal $\mu \in \ell_{\text{loc}}(\mathbb{Z}(\Delta); \mathbb{R})$ and returns the signal

$$\eta(t) = \frac{\Delta}{T_+ + T_-} \sum_{j=(t-T_-)/\Delta}^{(t+T_+-\Delta)/\Delta} \mu(j\Delta), \qquad t \in \mathbb{R},$$

for $T_-, T_+ \in \mathbb{Z}_{\geq 0}(\Delta)$ with $T_+ + T_- \in \mathbb{Z}_{>0}(\Delta)$. Answer the following questions.

(a) Show that this is a discrete-time convolution system and determine the convolution kernel.

(b) Is the system causal? strongly causal? stationary? strongly stationary? memoryless?

(c) Compute the output associated with the input $\mu = \mathsf{1}_{\geq 0}$.

(d) Let $\Delta = T_+ = T_- = 1$ and compute the output associated with the input $\mu(t) = \sin(\pi t)$.

6.9.6 Consider a discrete-time convolution system $\Sigma = (\mathbb{R}, \mathbb{R}, \mathbb{Z}(\Delta), \mathscr{U}, \mathscr{Y}, \mathsf{k})$ with $\mathsf{k} \in \ell^1(\mathbb{Z}(\Delta); \mathbb{R})$. Define the ***step response*** of the system to be the output associated with the input $\mu = \mathsf{1}_{\geq 0}$: $\mathsf{1}_\Sigma = \mathsf{k} * \mathsf{1}_{\geq 0}$ (evidently we are assuming that $\mathsf{1}_{\geq 0} \in \mathscr{U}$). Show that

$$\mathsf{1}_\Sigma(m\Delta) = \sum_{j=-\infty}^{m} \mathsf{k}(j\Delta), \quad \mathsf{k}(m\Delta) = (\mathsf{1}_\Sigma(m\Delta) - \mathsf{1}_\Sigma((m-1)\Delta)), \qquad m \in \mathbb{Z}_{>0}.$$

6.9.7 Using the step response from Exercise 6.9.6, suppose that you know that a continuous-time convolution system $\Sigma = (\mathbb{R}, \mathbb{R}, \mathbb{Z}(\Delta), \mathscr{U}, \mathscr{Y}, \mathsf{k})$ has the step response

$$\mathsf{1}_\Sigma(t) = \begin{cases} 1, & t \in [j\Delta, k\Delta), \\ 0, & \text{otherwise} \end{cases}$$

for some $j, k \in \mathbb{Z}$ satisfying $j < k$.

(a) Determine the convolution kernel.
(b) For which values of j and k is the system causal?

The next few exercises have to do with a subject known as "time series analysis." For the exercises below, this concerns the analysis of streams of real, temporal, discrete-time data. Some instances of time series data are those considered in Example IV-1.1.1. In time series analysis, one wishes to develop a model for the process that produces the observed data so that one can predict future values of the data. In deterministic processes, one can effectively do this by some sort of curve-fitting. However, for nondeterministic processes, one must devise a probabilistic model that will predict the future in a statistical sense.

The general setup is this. One has a data stream $\eta \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}(\Delta); \mathbb{R})$ that is to be thought of as a trajectory of a random process. We assume that

$$\eta(k\Delta) = \mu(k\Delta) + \iota(k\Delta),$$

where $\mu \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}(\Delta); \mathbb{R})$ is a known deterministic function and where $\iota \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}(\Delta); \mathbb{R})$ is the ***innovation***. We assume ι is a white noise signal, meaning it has zero mean, constant covariance, and $\iota(j\Delta)$ and $\iota(k\Delta)$ are uncorrelated. We shall work with models where μ is a constant function of time, and is (without loss of generality) zero. The objective is to come up with a model that will predict the data stream at time $(k+1)\Delta$ given its values at times $j\Delta$, $j \in \{0, 1, \ldots, k\}$ and the values of the innovations at times $j\Delta$, $j \in \{0, 1, \ldots, k+1\}$. An important facet of such models are typically determined by how they model the correlation of the data at time $(k+1)\Delta$ and the earlier times and how the effects of the innovations are used to determine the future.

6.9.8 Consider the discrete time-domain $\mathbb{Z}_{\geq 0}(\Delta)$ and the difference equation

$$\eta(k+\Delta) = \alpha\eta(k) + (1-\alpha)\iota(k+\Delta), \qquad k \in \mathbb{Z}_{>0},$$

for signals $\eta, \iota \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}(\Delta), \mathbb{R})$ and for $\alpha \in (0, 1)$ that is determined by matching observed and predicted statistics. This is known as the ***exponential smoother*** for reasons you will explore in this problem.

Answer the following questions.

(a) Show that the solution to the system of difference equations with innovation ι specified and with initial condition $\eta(0) = y_0$ is

$$\eta(k\Delta) = \alpha^k y_0 + (1-\alpha)\sum_{j=0}^{k-1} \alpha^j \iota((k-j)\Delta).$$

(b) Show that, when an initial condition $\eta(0) = 0$ is specified, the previous equation describes a linear discrete-time input/output system with input ι and output η.

(c) What form does this process take as $k \to \infty$?
(d) How far into the future does the innovation at time $k\Delta$ affect the predicted output?

6.9.9 Consider the following difference equation

$$\eta(k\Delta) = b_1\eta((k-1)\Delta)+\cdots+b_{n-1}\eta((k-(n-1))\Delta)+b_n\eta((k-n)\Delta)+\iota(k\Delta), \qquad k \in \mathbb{Z}_{\geq n},$$

for signals $\eta, \iota \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}(\Delta); \mathbb{R})$ and for $b_1, \dots, b_n \in \mathbb{R}$. This is an ***autoregressive model*** of order n, denoted AR(n). Note, for example, that AR(0) is simply a white noise process. The coefficients $b_1, \dots, b_n$ are chosen to fit measured data by matching statistical properties.

You will examine some features of AR(1), which is determined by the equation

$$\eta(k\Delta) = b\eta((k-1)\Delta) + \iota(k\Delta), \qquad k \in \mathbb{Z}_{>0}.$$

For this system, answer the following questions.

(a) Show that the solution to the system of difference equations with innovation ι specified and with initial condition $\eta(0) = y_0$ is

$$\eta(k\Delta) = b^k y_0 + \sum_{j=0}^{k-1} b^j \iota(k-j), \qquad k \in \mathbb{Z}_{>0}.$$

(b) Show that, when an initial condition $\eta(0) = 0$ is specified, the previous equation describes a linear discrete-time input/output system with input ι and output η.
(c) If $|b| < 1$, what form does this process take as $k \to \infty$?
(d) How far into the future does the innovation at time $k\Delta$ affect the predicted output?

6.9.10 Consider the following difference equation

$$\eta(k\Delta) = \iota(k\Delta) + a_1\iota((k-1)\Delta) + \cdots + a_m\iota((k-m)\Delta), \qquad k \in \mathbb{Z}_{\geq m},$$

for signals $\eta, \iota \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}(\Delta); \mathbb{R})$ and for $a_1, \dots, a_m \in \mathbb{R}$. This is known as a ***moving average process*** of order m, denoted MA(m). Note, for example, that MA(0) is simply a white noise process. The coefficients $a_1, \dots, a_m$ are chosen to fit measured data by matching statistical properties.

You will examine some features of MA(1), which is determined by the equation

$$\eta(k\Delta) = \iota(k\Delta) + a\iota((k-1)\Delta), \qquad k \in \mathbb{Z}_{>0}.$$

For this system, answer the following questions.

(a) Show that the solution to the system of difference equations with innovation ι specified and with initial condition $\eta(0) = y_0$ is

$$\eta(k\Delta) = \iota(k\Delta) + a^k\iota(0) + \sum_{j=1}^{k-1} a^j \eta((k-j)\Delta), \qquad k \in \mathbb{Z}_{>0}.$$

(b) Show that, when an initial condition $\eta(0) = 0$ is specified, the previous equation describes a linear discrete-time input/output system with input ι and output η.
(c) If $|a| < 1$, what form does this process take as $k \to \infty$?
(d) How far into the future does the innovation at time $k\Delta$ affect the predicted output?

6.9.11 Consider the following difference equation

$$\begin{aligned}\eta(k\Delta) = b_1\eta((k-1)\Delta) + \cdots + b_{n-1}\eta((k-(n-1))\Delta) + b_n\eta((k-n)\Delta)\\ + \iota(k\Delta) + a_1\iota((k-1)\Delta) + \cdots + a_m\iota((k-m)\Delta), \qquad k \in \mathbb{Z}_{\geq n},\end{aligned}$$

for signals $\eta, \iota \in \ell_{\text{loc}}(\mathbb{Z}_{\geq 0}(\Delta); \mathbb{R})$ and for $b_1, \ldots, b_n, a_1, \ldots, a_m \in \mathbb{R}$. This is known as a ***autoregressive moving average process*** of order (n, m), denoted ARMA(n, m). Note, for example, that ARMA$(0, 0)$ is simply a white noise process. The coefficients $b_1, \ldots, b_n, a_1, \ldots, a_m$ are chosen to fit measured data by matching statistical properties.

You will examine some features of ARMA$(1, 1)$, which is determined by the equation

$$\eta(k\Delta) = b\eta((k-1)\Delta) + \iota(k\Delta) + a\iota((k-1)\Delta), \qquad k \in \mathbb{Z}_{>0}.$$

For this system, answer the following questions.

(a) Show that, when an initial condition $\eta(0) = 0$ is specified, the previous equation describes a linear discrete-time input/output system with input ι and output η.
(b) Show that the solution to the system of difference equations with innovation ι specified and with initial condition $\eta(0) = 0$ is

$$\eta(k\Delta) = b^k y_0 + \sum_{j=0}^{k-1} b^j \iota((k-j+1)\Delta) + a\sum_{j=0}^{k-1} b^j \iota((k-j)\Delta), \qquad k \in \mathbb{Z}_{>0}.$$

(c) If $|b| < 1$, what form does this process take as $k \to \infty$?
(d) How far into the future does the innovation at time $k\Delta$ affect the predicted output?

[Franses and van Dijk 2003]

Chapter 7

Linear systems: Transfer function representations

One of the ways in which linear systems are special is that they admit so-called transfer function representations. In this section we shall examine carefully the rôle of transfer functions for linear systems, both continuous- and discrete-time, and both as state space and input/output representations. We shall focus on stationary systems as these interact most nicely with the notion of a transfer function.

The transfer function descriptions we give rely on the Laplace transforms described in Chapter IV-9. Transfer function representations for stationary linear systems have the property that, under appropriate technical hypotheses, the output is the product of the transfer function with the input. This feature is essentially inherited from the manner in which the Laplace transforms interact with convolution, cf. Sections IV-9.1.3 and IV-9.2.3. It is this simple manner in which the system is manifested by the transfer function that accounts for some of the utility of transfer function representations. Less obvious is that, by using the transfer function representation, one avails oneself of tools of complex analysis since transfer swap the time "t" for the complex variable "z." In any event, the transfer function representation of a system comes from a rather different place than the time-domain representation, and, adopting the view that more knowledge is better, this is a reason for understanding transfer function representations.

Do I need to read this chapter? The material in this chapter is a standard part of the theory of linear time-invariant systems. •

Contents

Section 7.1

Transfer functions for continuous-time linear systems

We start our discussion by a consideration of transfer functions for continuous-time linear systems, such as are introduced in Sections 6.6 and 6.7. As mentioned in the introduction to the chapter, the methods associated with Laplace transforms work best with stationary systems. Thus we will not have anything to say about an huge swath of the systems we presented in Chapter 6. More particularly, we will concentrate on stationary systems that are described in Sections 6.6.2, 6.7.4, and 6.7.7.

Do I need to read this section? If you are reading this chapter, then you will need to read this section. •

7.1.1 Complexification of continuous-time linear systems

The causal CLT converts a $\mathbb{R}$-vector space valued function of time into a $\mathbb{C}$-vector space valued function of a complex variable. To properly describe how the Laplace transform interacts with a continuous-time linear system, we need to indicate how this conversion from "real" to "complex" takes place. If the state spaces, input spaces, and output spaces are not general finite-dimensional $\mathbb{R}$-vector spaces, but actually Euclidean spaces, then this conversion is done in an unthinking way. However, in the more abstract setting we employ, this should be carried out explicitly.

We start with continuous-time convolution systems.

7.1.1 Definition (Complexification of continuous-time convolution systems) The complexification of a continuous-time convolution system $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ is

$$\Sigma_{\mathbb{C}} = (\mathsf{U}_{\mathbb{C}}, \mathsf{Y}_{\mathbb{C}}, \mathscr{U}_{\mathbb{C}}, \mathscr{Y}_{\mathbb{C}}, \mathsf{k}_{\mathbb{C}}),$$

where

(i) $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{Y}_{\mathbb{C}}$ are the complexifications as per Definition I-4.5.60,

(ii) $\mathscr{U}_{\mathbb{C}} = \{\mu \colon \mathbb{R} \to \mathsf{U}_{\mathbb{C}} \mid \operatorname{Re}(\mu), \operatorname{Im}(\mu) \in \mathscr{U}\}$,

(iii) $\mathscr{Y}_{\mathbb{C}} = \{\eta \colon \mathbb{R} \to \mathsf{Y}_{\mathbb{C}} \mid \operatorname{Re}(\eta), \operatorname{Im}(\eta) \in \mathscr{Y}\}$, and

(iv) $\mathsf{k}_{\mathbb{C}} \in \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$ is the complexification of k as per Definition I-5.4.62. •

The resulting system associated with a complexification is then $g_{\mathsf{k}_{\mathbb{C}}} \colon \mathscr{U}_{\mathbb{C}} \to \mathscr{Y}_{\mathbb{C}}$ given by

$$g_{\mathsf{k}_{\mathbb{C}}}(\mu)(t) = \int_{\mathbb{R}} \mathsf{k}_{\mathbb{C}}(t - \tau)(\mu(\tau)) \, \mathrm{d}\tau$$

If one restricts to real inputs, then one ends up with the original system (Exercise 7.1.1).

Now we consider the complexification of differential input/output systems.

7.1.2 Definition (Complexification of linear continuous-time differential input/output systems)

Finally, we indicate how to complexify state space systems.

7.1.3 Definition (Complexification of linear continuous-time state space systems) The complexification of a linear continuous-time state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

with constant coefficients is

$$\Sigma_{\mathbb{C}} = (\mathsf{X}_{\mathbb{C}}, \mathsf{U}_{\mathbb{C}}, \mathsf{Y}_{\mathbb{C}}, \mathscr{U}_{\mathbb{C}}, \mathscr{Y}_{\mathbb{C}}, \mathsf{A}_{\mathbb{C}}, \mathsf{B}_{\mathbb{C}}, \mathsf{C}_{\mathbb{C}}, \mathsf{D}_{\mathbb{C}}),$$

where

(i) $\mathsf{X}_{\mathbb{C}}$, $\mathsf{U}_{\mathbb{C}}$, and $\mathsf{Y}_{\mathbb{C}}$ are the complexifications as per Definition I-4.5.60,

(ii) $\mathscr{U}_{\mathbb{C}} = \{\mu \colon \mathbb{R} \to \mathsf{U}_{\mathbb{C}} \mid \ \mathrm{Re}(\mu), \mathrm{Im}(\mu) \in \mathscr{U}\}$,

(iii) $\mathscr{Y}_{\mathbb{C}} = \{\eta \colon \mathbb{R} \to \mathsf{Y}_{\mathbb{C}} \mid \ \mathrm{Re}(\eta), \mathrm{Im}(\eta) \in \mathscr{Y}\}$, and

(iv) $\mathsf{A}_{\mathbb{C}} \in \mathrm{L}(\mathsf{X}_{\mathbb{C}}; \mathsf{X}_{\mathbb{C}})$, $\mathsf{B}_{\mathbb{C}} \in \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{X}_{\mathbb{C}})$, $\mathsf{C}_{\mathbb{C}} \in \mathrm{L}(\mathsf{X}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$, and $\mathsf{D}_{\mathbb{C}} \in \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$, are the complexifications as per Definition I-5.4.62. •

A controlled trajectory $(\xi, \mu) \in \mathrm{Ctraj}(\Sigma_{\mathbb{C}})$ and a corresponding controlled output $(\eta, \mu) \in \mathrm{Cout}(\Sigma_{\mathbb{C}})$ satisfy the equations

$$\begin{aligned} \dot{\xi}(t) &= \mathsf{A}_{\mathbb{C}} \circ \xi(t) + \mathsf{B}_{\mathbb{C}} \circ \mu(t), \\ \eta(t) &= \mathsf{C}_{\mathbb{C}} \circ \xi(t) + \mathsf{D}_{\mathbb{C}} \circ \mu(t). \end{aligned}$$

We invite the reader to show in Exercise 7.1.3 that the restriction to real inputs gives the same controlled trajectories and controlled outputs as the original system.

7.1.2 Transfer functions for continuous-time convolution systems

Our discussion of transfer function in the continuous-time case starts with the consideration of convolution systems. These systems are necessarily strongly stationary by Proposition 6.7.8. We shall not invoke the assumption of causality at the outset, but we will require that our convolution kernels have support that is bounded on the left so that we can apply to them the causal continuous Laplace transform of Section IV-9.1. Be careful to note, however, that this does *not* mean that the convolution kernels are causal as per Definition 6.7.11.

The following is the essential definition with which we work.

7.1.4 Definition (Transfer function for continuous-time convolution system) Let $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ be a continuous-time convolution system and suppose that $\mathsf{k} \in \mathsf{LT}^{+,p}(\mathbb{R}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$. The ***transfer function*** for Σ is the mapping

$$\begin{aligned} \mathsf{T}_\Sigma \colon \mathbb{C}_{I^1(\mathsf{k})} &\to \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}) \\ z &\mapsto \mathscr{L}^p_{\mathrm{C}}(\mathsf{k})(z). \end{aligned}$$

•

Of course, for a convolution system, the input/output map is the map $g_{\mathsf{k}} \colon \mathscr{U} \to \mathscr{Y}$ defined by

$$g_{\mathsf{k}}(\mu)(t) = \int_{\mathbb{R}} \mathsf{k}(t - \tau)(\mu(\tau)) \, \mathrm{d}\tau.$$

By using the interactions of convolution and the causal CLT as in Propositions IV-9.1.10 and IV-9.1.11, we anticipate that, by taking the causal CLT of the equation, we get

$$\mathscr{L}^p_{\mathrm{C}}(g_{\mathsf{k}}(\mu))(z) = T_\Sigma(z) \mathscr{L}^p_{\mathrm{C}}(\mu)(z).$$

For such a conclusion to hold, there are various impediments: (1) k has to be Laplace transformable; (2) the inputs have to be Laplace transformable; (3) the causal CLT of the output is the "product" of the transfer function and the causal CLT of the input. These impediments arise on top of the matter that, for a convolution system, the convolution kernel has to be compatible with the inputs. Any of these impediments can arise, and provide a limitation to the application of the Laplace transform methods. To bookkeep these issues, we make a definition.

7.1.5 Definition (Laplace transformable continuous-time convolution system) Let $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ be a continuous-time convolution system and let $p, q, r \in [1, \infty]$. Then Σ is an **LT(p, q, r)-*convolution system*** if the following conditions hold:

(i) $\mathsf{k} \in \mathsf{LT}^{r,+}(\mathbb{R}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$;

(ii) $\mathscr{U} \subseteq \mathsf{LT}^{p,+}(\mathbb{R}; \mathsf{U})$;

(iii) $\mathscr{Y} \subseteq \mathsf{LT}^{q,+}(\mathbb{R}; \mathsf{U})$;

(iv) $\mathscr{L}^q_{\mathrm{C}}(\mathsf{k} * \mu)(z) = \mathscr{L}^r_{\mathrm{C}}(\mathsf{k})(z) \mathscr{L}^p_{\mathrm{C}}(\mu)(z)$ for $z \in \mathbb{C}_I$ for some nonempty interval $I \subseteq \mathbb{R}$ for which $\sup I = \infty$.

•

In general, given a continuous-time convolution system $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$, if the formula of Definition 7.1.5(iv) holds, then we say that the input μ satisfies the ***exchange formula***.

The definition brings into focus the circumstances under which a continuous-time convolution system can be profitably handled with the causal CLT. It still remains, however, to determine when a system is an $\mathsf{LT}(p, q, r)$-convolution system. We give two results that characterise some such systems.

A first useful result is the following, which makes use of Young's Inequality for convolution and does not require strict causality of the signals. The result follows from Proposition IV-9.1.11.

7.1.6 Proposition (LT(p, q, r)-convolution systems) *Let* $p, q, r \in [1, \infty]$ *satisfy* $\frac{1}{q} = \frac{1}{r} + \frac{1}{p} - 1$. *Let* $\Sigma = (U, Y, \mathscr{U}, \mathscr{Y}, k)$ *be a continuous-time convolution system for which*

(i) $k \in \mathsf{LT}^{r,+}(\mathbb{R}; L(U; Y))$,

(ii) $\mathscr{U} \subseteq \mathsf{LT}^{p,+}(\mathbb{R}; U)$, *and*

(iii) $\mathscr{Y} \subseteq \mathsf{LT}^{q,+}(\mathbb{R}; Y)$.

Then Σ *is an* $\mathsf{LT}(p, q, r)$*-convolution system and, for* $\mu \in \mathscr{U}$,

$$\operatorname{int}(I^q(k * \mu)) \supseteq \operatorname{int}(I^r(k)) \cap \operatorname{int}(I^p(\mu)).$$

For causal convolution systems, there are additional results one can apply that are useful. The first follows directly from Proposition IV-9.1.10 where we proved the exchange formula for strictly causal signals in $\mathsf{LT}^{\infty,+}(\mathbb{R}; \mathbb{C})$.

7.1.7 Proposition (Strictly causal LT(∞, ∞, ∞)-convolution systems) *If a continuous-time convolution system* $\Sigma = (U, Y, \mathscr{U}, \mathscr{Y}, k)$ *satisfies*

(i) $k \in \mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; L(U; Y))$,

(ii) $\mathscr{U} \subseteq \mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; U)$, *and*

(iii) $\mathscr{Y} \subseteq \mathsf{LT}^{\infty,+}(\mathbb{R}_{\geq 0}; Y)$,

then Σ *is an* $\mathsf{LT}(\infty, \infty, \infty)$*-convolution system and, if* $\mu \in \mathscr{U}$, *then*

$$I^\infty(k * \mu) \supseteq I^\infty(k) \cap I^\infty(\mu).$$

The following result is a useful one for system theory, and in it we make use of the vector space-valued Hardy spaces norms from Section IV-1.4.4, which are derived from the scalar versions described in detail in Chapter III-7. We also assume that the space of linear maps between two vector spaces is equipped with a norm satisfying the submultiplicative property (IV-1.4).

7.1.8 Proposition (Causal LT(2, 2, 1)-convolution systems) *Consider a continuous-time convolution system* $\Sigma = (U, Y, \mathscr{U}, \mathscr{Y}, k)$ *satisfies*

(i) $k \in L^1(\mathbb{R}_{\geq 0}; L(U; Y))$,

(ii) $\mathscr{U} \subseteq L^2(\mathbb{R}_{\geq 0}; U)$, *and*

(iii) $\mathscr{Y} \subseteq L^2(\mathbb{R}_{\geq 0}; U)$.

Then Σ *is an* $\mathsf{LT}(2, 2, 1)$*-convolution system and, for* $\mu \in \mathscr{U}$,

$$\|\mathscr{L}^2_C(k * \mu)\|_{H^2, \mathbb{R}_{\geq 0}} \leq \|\mathscr{L}^1_C(k)\|_{H^\infty, \mathbb{R}_{\geq 0}} \|\mathscr{L}^2_C(\mu)\|_{H^2, \mathbb{R}_{\geq 0}}.$$

Proof That the system is an $\mathsf{LT}(2, 2, 1)$-convolution system follows from Proposition 7.1.7. That $\mathscr{L}^2_C(\mu) \in H^2(\mathbb{C}_{\mathbb{R}_{\geq 0}}; U_{\mathbb{C}})$ and $\mathscr{L}^2_C(k * \mu) \in H^2(\mathbb{C}_{\mathbb{R}_{\geq 0}}; Y_{\mathbb{C}})$ follows from Theorem IV-9.1.17. By Proposition IV-9.1.16, $\mathscr{L}^1_C(k) \in H^\infty(\mathbb{C}_{\mathbb{R}_{\geq 0}}; L(U_{\mathbb{C}}; Y_{\mathbb{C}}))$. The final assertion follows from the assumed submultiplicative property of the norm on $L(U_{\mathbb{C}}; Y_{\mathbb{C}})$. ■

The following result gives a sometimes useful interpretation of the transfer function.

7.1.9 Proposition (The transfer function and exponential inputs) *Let* $\Sigma = (\mathsf{U}; \mathsf{Y}; \mathscr{U}, \mathscr{Y}, \mathsf{k})$ *be a continuous-time convolution system with* $\mathsf{k} \in \mathsf{LT}^{1,+}(\mathbb{R}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$. *Let* $a \in \mathbb{C}_{\mathrm{I}^1(\mathsf{k})}$ *and* $u \in \mathsf{U}_{\mathbb{C}}$, *and suppose that* $\mathsf{E}_a u \in \mathscr{U}_{\mathbb{C}}$. *Then,* $g_{\mathsf{k}_{\mathbb{C}}}(\mathsf{E}_a u)(t) = e^{at} \mathscr{L}^1_{\mathrm{C}}(\mathsf{k})(a)(u)$.

Proof We have

$$g_{\mathsf{k}_{\mathbb{C}}}(\mathsf{E}_a u)(t) = \int_{\mathbb{R}} \mathsf{k}_{\mathbb{C}}(t-\tau)(e^{a\tau}u)\,\mathrm{d}\tau = e^{at} \int_{\mathbb{R}} \mathsf{k}_{\mathbb{C}}(s)(u)e^{-as}\,\mathrm{d}s = e^{at}\mathscr{L}^1_{\mathrm{C}}(\mathsf{k})(a)(u),$$

as claimed. ■

7.1.3 Transfer functions for linear continuous-time differential input/output systems

7.1.4 Transfer functions for linear continuous-time state space systems

Let us now consider the transfer function of a linear continuous-time state space system. As we shall see, these transfer functions have a specific structure that is related directly to the state space structure of these systems.

Let us begin with the definition.

7.1.10 Definition (Transfer function for linear continuous-time state space systems with constant coefficients) For a linear continuous-time state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

with constant coefficients, the ***transfer function*** is the $\mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$-valued function

$$\begin{aligned} \mathsf{T}_\Sigma \colon \mathbb{C}_{(\sigma_{\max}(\mathsf{A}),\infty)} &\to \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}) \\ z &\mapsto \mathsf{C}_{\mathbb{C}} \circ (z\,\mathrm{id}_{\mathsf{X}_{\mathbb{C}}} - \mathsf{A}_{\mathbb{C}})^{-1} \circ \mathsf{B}_{\mathbb{C}} + \mathsf{D}_{\mathbb{C}}, \end{aligned}$$

where

$$\sigma_{\max}(\mathsf{A}) = \max\{\mathrm{Re}(\lambda) \mid \lambda \in \mathrm{spec}(\mathsf{A})\}.$$ •

Let us first establish the connection of the transfer function with the impulse response considered in Section 6.6.3.2.

7.1.11 Theorem (The transfer function and the impulse response) *For a linear continuous-time state space system*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

with constant coefficients, $\mathsf{T}_\Sigma = \mathscr{L}^1_{\mathrm{C}}(\mathrm{ir}_\Sigma)$.

Proof This follows from Proposition 5.4.2 and . ■ CLT of delta-function

Let us enumerate a few properties of the transfer function. To do so, we first do some preparatory work with polynomials and rational functions. We recall from Definition I-4.4.46 the notion of a rational function as a quotient of polynomial functions. The set of rational functions with coefficients in $\mathbb{F}$ and

with indeterminate ξ we denote by $\mathbb{F}(\xi)$. We also recall that, given a polynomial $P \in \mathbb{F}[\xi]$, we can think of "evaluating" P as a function of $x \in \mathbb{F}$ in the obvious way (Proposition I-4.4.9). Thus, if the denominator polynomial in a rational function evaluates to something nonzero at $x \in \mathbb{F}$, then we can similarly evaluate a rational function at x. Given $R \in \mathbb{F}(\xi)$ with coprime fractional representative $R = \frac{N}{D}$ (see Proposition I-4.4.47) we can thus write

$$\mathrm{Ev}_{\mathbb{F}}(R)(x) = \frac{\mathrm{Ev}_{\mathbb{F}}(N)(x)}{\mathrm{Ev}_{\mathbb{F}}(D)(x)}.$$

We shall also make use of the tensor product of vector spaces from Definition I-5.6.11 and of linear maps from Proposition I-5.6.17.

7.1.12 Proposition (Properties of the transfer function for linear continuous-time state space systems with constant coefficients) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{R}, \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear continuous-time state space system with constant coefficients and let T_Σ *be its transfer function. Then the following statements hold:*

(i) there exists $\hat{\mathsf{T}}_\Sigma \in \mathbb{C}(\xi) \otimes \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$ *such that*

$$\mathsf{T}_\Sigma(z) = \mathrm{Ev}_{\mathbb{C}} \otimes \mathrm{id}_{\mathrm{L}(\mathsf{U}_{\mathbb{C}};\mathsf{Y}_{\mathbb{C}})}(\hat{\mathsf{T}}_\Sigma)(z), \qquad z \in \mathbb{C}_{(\sigma_{\max}(\mathsf{A}),\infty)};$$

(ii) if $\mathsf{D} = 0$*, then, for any* $a > \sigma_{\max}(\mathsf{A})$, $\mathsf{T}_\Sigma|\mathbb{C}_{[a,\infty)} \in \mathsf{H}^2(\mathbb{C}_{[a,\infty)}; \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}))$;

(iii) for any $a > \sigma_{\max}(\mathsf{A})$, $\mathsf{T}_\Sigma|\mathbb{C}_{[a,\infty)} \in \mathsf{H}^\infty(\mathbb{C}_{[a,\infty)}; \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}))$.

Proof Throughout the proof, we choose bases for X, U, and Y, and work with matrix representatives $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$. We suppose that

$$n = \dim_{\mathbb{R}}(\mathsf{X}), \quad m = \dim_{\mathbb{R}}(\mathsf{U}), \quad \dim_{\mathbb{R}}(\mathsf{Y}).$$

(i) Note that $T_\Sigma(z)$ can be computed by matrix multiplication and matrix inversion, with the dependence on z coming from the term involving the matrix inverse $(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}$. If we make reference to Theorem I-5.3.10, we see that each component of this inverse will be a quotient of polynomials in z, the numerator arising from a determinant of an $(n-1) \times (n-1)$-matrix with terms that are either linear in z or constant with respect to z. Thus each such determinant will be a polynomial of degree at most $n-1$. The denominator polynomial will be the characteristic polynomial of $\boldsymbol{A}$, which, therefore, has degree n by Proposition I-5.8.17.

(ii) Note that $I^\infty(\mathrm{pir}_\Sigma) \supseteq (\sigma_{\max}(\mathsf{A}), \infty)$. Therefore, if $a > \sigma_{\max}(\mathsf{A})$ then

$$\|\mathrm{e}^{-at}\mathrm{pir}_\Sigma(t)\| \leq M\mathrm{e}^{-(a-\sigma_{\max})t}, \qquad t \in \mathbb{R}_{\geq 0}.$$

Thus $t \mapsto \mathrm{e}^{-at}\mathrm{pir}_\Sigma(t)$ is in $\mathsf{L}^2(\mathbb{R}_{\geq 0}; \mathrm{L}(\mathsf{U}; \mathsf{Y}))$. By Theorem IV-9.1.17 we conclude that $\mathscr{L}^2_{\mathrm{C}}(\mathrm{pir}_\Sigma) \in \mathsf{H}^2(\mathbb{C}_{[a,\infty)}; \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}))$, and so the result follows from the Theorem 7.1.11.

(iii) This follows by definition and the preceding part of the proof. ■

induced norm on H^2

Exercises

7.1.1 Show that the restriction of the complexification of a continuous-time convolution system to real inputs agrees with the original system.

7.1.2 Show that the restriction of the complexification of a linear continuous-time differential input/output system to real inputs agrees with the original system.

7.1.3 Show that the restriction of the complexification of a linear continuous-time state space system to real inputs agrees with the original system.

7.1.4 For the Butterworth filter of Exercise 6.7.5, determine its transfer function.

7.1.5 For the continuous-time sliding averager of Exercise 6.7.6, answer the following questions:

(a) determine its transfer function;

(b) comment on the causality of the system, given its transfer function.

7.1.6 For the first three linear continuous-time state space systems of Exercise 6.6.2, compute their transfer functions.

7.1.7 For the linear continuous-time state space system you derived in Exercise 6.6.3, compute its transfer function.

7.1.8 For the linear continuous-time state space system of Exercise 6.6.4, compute its transfer function.

7.1.9 For the linear continuous-time state space system of Exercise 6.6.5, compute its transfer function.

Section 7.2

Transfer functions for discrete-time linear systems

We continue our discussion by a consideration of transfer functions for discrete-time linear systems, such as are introduced in Sections 6.8 and 6.9. As in Section 7.1, we will concentrate on stationary systems that are described in Sections 6.8.2, 6.9.4, and 6.9.7.

Do I need to read this section? If you are reading this chapter, then you will need to read this section. •

7.2.1 Complexification of discrete-time linear systems

The causal DLT converts a $\mathbb{R}$-vector space valued function of time into a $\mathbb{C}$-vector space valued function of a complex variable. To properly describe how the Laplace transform interacts with a continuous-time linear system, we need to indicate how this conversion from "real" to "complex" takes place. If the state spaces, input spaces, and output spaces are not general finite-dimensional $\mathbb{R}$-vector spaces, but actually Euclidean spaces, then this conversion is done in an unthinking way. However, in the more abstract setting we employ, this should be carried out explicitly.

We start with discrete-time convolution systems.

7.2.1 Definition (Complexification of discrete-time convolution systems) The complexification of a discrete-time convolution system $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ is

$$\Sigma_{\mathbb{C}} = (\mathsf{U}_{\mathbb{C}}, \mathsf{Y}_{\mathbb{C}}, \mathscr{U}_{\mathbb{C}}, \mathscr{Y}_{\mathbb{C}}, \mathsf{k}_{\mathbb{C}}),$$

where

(i) $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{Y}_{\mathbb{C}}$ are the complexifications as per Definition I-4.5.60,

(ii) $\mathscr{U}_{\mathbb{C}} = \{\mu \colon \mathbb{Z}(\Delta) \to \mathsf{U}_{\mathbb{C}} \mid \operatorname{Re}(\mu), \operatorname{Im}(\mu) \in \mathscr{U}\}$,

(iii) $\mathscr{Y}_{\mathbb{C}} = \{\eta \colon \mathbb{Z}(\Delta) \to \mathsf{Y}_{\mathbb{C}} \mid \operatorname{Re}(\eta), \operatorname{Im}(\eta) \in \mathscr{Y}\}$, and

(iv) $\mathsf{k}_{\mathbb{C}} \in \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$ is the complexification of k as per Definition I-5.4.62. •

The resulting system associated with a complexification is then $g_{\mathsf{k}_{\mathbb{C}}} \colon \mathscr{U}_{\mathbb{C}} \to \mathscr{Y}_{\mathbb{C}}$ given by

$$g_{\mathsf{k}_{\mathbb{C}}}(k\Delta)(t) = \sum_{j \in \mathbb{Z}} \mathsf{k}_{\mathbb{C}}((k-j)\Delta)(\mu(j\Delta))$$

If one restricts to real inputs, then one ends up with the original system (Exercise 7.2.1).

Now we consider the complexification of differential input/output systems.

7.2.2 Definition (Complexification of linear discrete-time differential input/output systems)

Finally, we indicate how to complexify state space systems.

7.2.3 Definition (Complexification of linear discrete-time state space systems) The complexification of a linear discrete-time state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

with constant coefficients is

$$\Sigma_{\mathbb{C}} = (\mathsf{X}_{\mathbb{C}}, \mathsf{U}_{\mathbb{C}}, \mathsf{Y}_{\mathbb{C}}, \mathscr{U}_{\mathbb{C}}, \mathscr{Y}_{\mathbb{C}}, \mathsf{A}_{\mathbb{C}}, \mathsf{B}_{\mathbb{C}}, \mathsf{C}_{\mathbb{C}}, \mathsf{D}_{\mathbb{C}}),$$

where

(i) $\mathsf{X}_{\mathbb{C}}$, $\mathsf{U}_{\mathbb{C}}$, and $\mathsf{Y}_{\mathbb{C}}$ are the complexifications as per Definition I-4.5.60,

(ii) $\mathscr{U}_{\mathbb{C}} = \{\mu \colon \mathbb{Z}(\Delta) \to \mathsf{U}_{\mathbb{C}} \mid \operatorname{Re}(\mu), \operatorname{Im}(\mu) \in \mathscr{U}\}$,

(iii) $\mathscr{Y}_{\mathbb{C}} = \{\eta \colon \mathbb{Z}(\Delta) \to \mathsf{Y}_{\mathbb{C}} \mid \operatorname{Re}(\eta), \operatorname{Im}(\eta) \in \mathscr{Y}\}$, and

(iv) $\mathsf{A}_{\mathbb{C}} \in \mathrm{L}(\mathsf{X}_{\mathbb{C}}; \mathsf{X}_{\mathbb{C}})$, $\mathsf{B}_{\mathbb{C}} \in \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{X}_{\mathbb{C}})$, $\mathsf{C}_{\mathbb{C}} \in \mathrm{L}(\mathsf{X}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$, and $\mathsf{D}_{\mathbb{C}} \in \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$, are the complexifications as per Definition I-5.4.62. •

A controlled trajectory $(\xi, \mu) \in \operatorname{Ctraj}(\Sigma_{\mathbb{C}})$ and a corresponding controlled output $(\eta, \mu) \in \operatorname{Cout}(\Sigma_{\mathbb{C}})$ satisfy the equations

$$\begin{aligned} \xi(t + h) &= \mathsf{A}_{\mathbb{C}} \circ \xi(t) + \mathsf{B}_{\mathbb{C}} \circ \mu(t), \\ \eta(t) &= \mathsf{C}_{\mathbb{C}} \circ \xi(t) + \mathsf{D}_{\mathbb{C}} \circ \mu(t). \end{aligned}$$

We invite the reader to show in Exercise 7.2.3 that the restriction to real inputs gives the same controlled trajectories and controlled outputs as the original system.

7.2.2 Transfer functions for discrete-time convolution systems

Our discussion of transfer function in the discrete-time case starts with the consideration of convolution systems. These systems are necessarily strongly stationary by Proposition 6.9.9. We shall not invoke the assumption of causality at the outset, but we will require that our convolution kernels have support that is bounded on the left so that we can apply to them the causal discrete Laplace transform of Section IV-9.2. Be careful to note, however, that this does *not* mean that the convolution kernels are causal as per Definition 6.9.12.

The following is the essential definition with which we work.

7.2.4 Definition (Transfer function for discrete-time convolution system) Let $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ be a discrete-time convolution system and suppose that $\mathsf{k} \in \mathrm{LT}^{+,p}(\mathbb{Z}(\Delta); \mathrm{L}(\mathsf{U}; \mathsf{Y}))$. The ***transfer function*** for Σ is the mapping

$$\begin{aligned} \mathsf{T}_{\Sigma} \colon \mathbb{A}_{I^1(\mathsf{k})} &\to \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}) \\ z &\mapsto \mathscr{L}^p_{\mathrm{D}}(\mathsf{k})(z). \end{aligned}$$

•

Of course, for a convolution system, the input/output map is the map $g_{\mathsf{k}} : \mathscr{U} \to \mathscr{Y}$ defined by

$$g_{\mathsf{k}}(\mu)(k\Delta) = \sum_{j \in \mathbb{Z}} \mathsf{k}((k-j)\Delta)(\mu(j\Delta\tau)).$$

By using the interactions of convolution and the causal DLT as in Propositions IV-9.2.9 and IV-9.2.10, we anticipate that, by taking the causal DLT of the equation, we get

$$\mathscr{L}^p_{\mathrm{D}}(g_{\mathsf{k}}(\mu))(z) = T_\Sigma(z)\mathscr{L}^p_{\mathrm{D}}(\mu)(z).$$

For such a conclusion to hold, there are various impediments: (1) k has to be Laplace transformable; (2) the inputs have to be Laplace transformable; (3) the causal DLT of the output is the "product" of the transfer function and the causal DLT of the input. These impediments arise on top of the matter that, for a convolution system, the convolution kernel has to be compatible with the inputs. Any of these impediments can arise, and provide a limitation to the application of the Laplace transform methods. To bookkeep these issues, we make a definition.

7.2.5 Definition (Laplace transformable discrete-time convolution system) Let $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ be a discrete-time convolution system and let $p, q, r \in [1, \infty]$. Then Σ is an **LT(p, q, r)-*convolution system*** if the following conditions hold:

(i) $\mathsf{k} \in \mathsf{LT}^{r,+}(\mathbb{Z}(\Delta); \mathrm{L}(\mathsf{U}; \mathsf{Y}))$;

(ii) $\mathscr{U} \subseteq \mathsf{LT}^{p,+}(\mathbb{Z}(\Delta); \mathsf{U})$;

(iii) $\mathscr{Y} \subseteq \mathsf{LT}^{q,+}(\mathbb{Z}(\Delta); \mathsf{U})$;

(iv) $\mathscr{L}^q_{\mathrm{D}}(\mathsf{k} * \mu)(z) = \mathscr{L}^r_{\mathrm{D}}(\mathsf{k})(z)\mathscr{L}^p_{\mathrm{D}}(\mu)(z)$ for $z \in \mathbb{A}_I$ for some nonempty interval $I \subseteq \mathbb{R}$ for which $\sup I = \infty$. •

As with continuous-time systems, if the formula of Definition 7.2.5(iv) holds, then we say that the input μ satisfies the ***exchange formula***.

The definition brings into focus the circumstances under which a discrete-time convolution system can be profitably handled with the causal DLT. It still remains, however, to determine when a system is an $\mathsf{LT}(p, q, r)$-convolution system. We give two results that characterise some such systems.

A first useful result is the following, which makes use of Young's Inequality for convolution and does not require strict causality of the signals. The result follows from Proposition IV-9.2.10.

7.2.6 Proposition (LT(p, q, r)-convolution systems) *Let* $\mathrm{p}, \mathrm{q}, \mathrm{r} \in [1, \infty]$ *satisfy* $\frac{1}{\mathrm{q}} = \frac{1}{\mathrm{r}} + \frac{1}{\mathrm{p}} - 1$. *Let* $\Sigma = (\mathsf{U}, \mathsf{Y}, \mathscr{U}, \mathscr{Y}, \mathsf{k})$ *be a discrete-time convolution system for which*

(i) $\mathsf{k} \in \mathsf{LT}^{\mathrm{r},+}(\mathbb{Z}(\Delta); \mathrm{L}(\mathsf{U}; \mathsf{Y}))$,

(ii) $\mathscr{U} \subseteq \mathsf{LT}^{\mathrm{p},+}(\mathbb{Z}(\Delta); \mathsf{U})$, *and*

(iii) $\mathscr{Y} \subseteq \mathsf{LT}^{\mathrm{q},+}(\mathbb{Z}(\Delta); \mathsf{Y})$.

Then Σ *is an* $\mathsf{LT}(p, q, r)$*-convolution system and, for* $\mu \in \mathscr{U}$*,*

$$\operatorname{int}(I^q(k * \mu)) \supseteq \operatorname{int}(I^r(k)) \cap \operatorname{int}(I^p(\mu)).$$

For causal convolution systems, there are additional results one can apply that are useful. The first follows directly from Proposition IV-9.2.9 where we proved the exchange formula for strictly causal signals in $\mathsf{LT}^{\infty,+}(\mathbb{Z}(\Delta); \mathbb{C})$.

7.2.7 Proposition (Strictly causal $\mathsf{LT}(\infty, \infty, \infty)$-convolution systems) *If a continuous-time convolution system* $\Sigma = (U, Y, \mathscr{U}, \mathscr{Y}, k)$ *satisfies*

(i) $k \in \mathsf{LT}^{\infty,+}(\mathbb{Z}_{\geq 0}(\Delta); L(U; Y))$*,*

(ii) $\mathscr{U} \subseteq \mathsf{LT}^{\infty,+}(\mathbb{Z}_{\geq 0}(\Delta); U)$*, and*

(iii) $\mathscr{Y} \subseteq \mathsf{LT}^{\infty,+}(\mathbb{Z}_{\geq 0}(\Delta); Y)$*,*

then Σ *is an* $\mathsf{LT}(\infty, \infty, \infty)$*-convolution system and, if* $\mu \in \mathscr{U}$*, then*

$$I^\infty(k * \mu) \supseteq I^\infty(k) \cap I^\infty(\mu).$$

The following result is a useful one for system theory, and in it we make use of the vector space-valued Hardy spaces norms from Section IV-1.4.4, which are derived from the scalar versions described in detail in Chapter III-7. We also assume that the space of linear maps between two vector spaces is equipped with a norm satisfying the submultiplicative property (IV-1.4).

7.2.8 Proposition (Causal $\mathsf{LT}(2, 2, 1)$-convolution systems) *Consider a discrete-time convolution system* $\Sigma = (U, Y, \mathscr{U}, \mathscr{Y}, k)$ *satisfies*

(i) $k \in \ell^1(\mathbb{Z}_{\geq 0}(\Delta); L(U; Y))$*,*

(ii) $\mathscr{U} \subseteq \ell^2(\mathbb{Z}_{\geq 0}(\Delta); U)$*, and*

(iii) $\mathscr{Y} \subseteq \ell^2(\mathbb{Z}_{\geq 0}(\Delta); U)$*.*

Then Σ *is an* $\mathsf{LT}(2, 2, 1)$*-convolution system and, for* $\mu \in \mathscr{U}$*,*

$$\|\mathscr{L}^2_D(k * \mu)\|_{H^2,[1,\infty)} \leq \|\mathscr{L}^1_D(k)\|_{H^\infty,[1,\infty)} \|\mathscr{L}^2_D(\mu)\|_{H^2,[1,\infty)}.$$

Proof That the system is an $\mathsf{LT}(2, 2, 1)$-convolution system follows from Proposition 7.2.7. That $\mathscr{L}^2_D(\mu) \in H^2(\mathbb{A}_{[1,\infty)}; U_{\mathbb{C}})$ and $\mathscr{L}^2_D(k * \mu) \in H^2(\mathbb{A}_{[1,\infty)}; Y_{\mathbb{C}})$ follows from Theorem IV-9.2.15. By Proposition IV-9.2.14, $\mathscr{L}^1_D(k) \in H^\infty(\mathbb{A}_{[1,\infty)}; L(U_{\mathbb{C}}; Y_{\mathbb{C}}))$. The final assertion follows from the assumed submultiplicative property of the norm on $L(U_{\mathbb{C}}; Y_{\mathbb{C}})$. ■

The following result gives a sometimes useful interpretation of the transfer function.

7.2.9 Proposition (The transfer function and exponential inputs) *Let* $\Sigma = (\mathsf{U}; \mathsf{Y}; \mathscr{U}, \mathscr{Y}, \mathsf{k})$ *be a discrete-time convolution system with* $\mathsf{k} \in \mathsf{LT}^{1,+}(\mathbb{Z}(\Delta); \mathrm{L}(\mathsf{U}; \mathsf{Y}))$. *Let* $a \in \mathbb{A}_{\mathrm{I}^1(\mathsf{k})}$ *and* $u \in \mathsf{U}_{\mathbb{C}}$, *and suppose that* $\mathsf{P}_a u \in \mathscr{U}_{\mathbb{C}}$. *Then,* $g_{\mathsf{k}_{\mathbb{C}}}(\mathsf{P}_a u)(k\Delta) = a^k \mathscr{L}^1_{\mathrm{D}}(\mathsf{k})(a)(u)$.

Proof We have

$$g_{\mathsf{k}_{\mathbb{C}}}(\mathsf{P}_a u)(k\Delta) = \Delta \sum_{j \in \mathbb{Z}} \mathsf{k}_{\mathbb{C}}((k-j)\Delta)(a^j u) = a^k \Delta \sum_{l \in \mathbb{Z}} \mathsf{k}_{\mathbb{C}}(l\Delta)(u) a^{-l} = a^k \mathscr{L}^1_{\mathrm{D}}(\mathsf{k})(a)(u),$$

as claimed. ∎

7.2.3 Transfer functions for linear discrete-time differential input/output systems

7.2.4 Transfer functions for linear discrete-time state space systems

Let us now consider the transfer function of a linear continuous-time state space system. As we shall see, these transfer functions have a specific structure that is related directly to the state space structure of these systems.

Let us begin with the definition.

7.2.10 Definition (Transfer function for linear discrete-time state space systems with constant coefficients) For a linear discrete-time state space system

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{Z}(\Delta), \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

with constant coefficients, the ***transfer function*** is the $\mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$-valued function

$$\begin{aligned} \mathsf{T}_\Sigma \colon \mathbb{A}_{(\rho_{\max}(\mathsf{A}),\infty)} &\to \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}) \\ z &\mapsto \mathsf{C}_{\mathbb{C}} \circ (z \, \mathrm{id}_{\mathsf{X}_{\mathbb{C}}} - \mathsf{A}_{\mathbb{C}})^{-1} \circ \mathsf{B}_{\mathbb{C}} + \mathsf{D}_{\mathbb{C}}, \end{aligned}$$

where

$$\rho_{\max}(\mathsf{A}) = \max\{|\lambda| \mid \lambda \in \mathrm{spec}(\mathsf{A})\}.$$ •

Let us first establish the connection of the transfer function with the impulse response considered in Section 6.9.6.2.

7.2.11 Theorem (The transfer function and the impulse response) *For a linear discrete-time state space system*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{Z}(\Delta), \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

with constant coefficients, $\mathsf{T}_\Sigma = \mathscr{L}^1_{\mathrm{D}}(\mathrm{ir}_\Sigma)$.

Proof We have

$$\mathrm{pir}_\Sigma(k\Delta) = 1_{\geq 0}((k-1)\Delta)\mathsf{C} \circ \mathsf{P}_{\mathsf{A}}(k-1) \circ \mathsf{B}.$$

DLT of pulse

By Proposition 5.8.2, Exercise IV-9.2.4, and , the result follows. ∎

Let us next enumerate a few properties of the transfer function, analogous to those we have seen for continuous-time systems.

7.2.12 Proposition (Properties of the transfer function for linear discrete-time state space systems with constant coefficients) *Let*

$$\Sigma = (\mathsf{X}, \mathsf{U}, \mathsf{Y}, \mathbb{Z}(\Delta), \mathscr{U}, \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$$

be a linear discrete-time state space system with constant coefficients and let T_Σ *be its transfer function. Then the following statements hold:*

(i) there exists $\hat{\mathsf{T}}_\Sigma \in \mathbb{C}(\xi) \otimes \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}})$ *such that*

$$\mathsf{T}_\Sigma(z) = \mathrm{Ev}_{\mathbb{C}} \otimes \mathrm{id}_{\mathrm{L}(\mathsf{U}_{\mathbb{C}};\mathsf{Y}_{\mathbb{C}})}(\hat{\mathsf{T}}_\Sigma)(z), \qquad z \in \mathbb{A}_{(\rho_{\max}(\mathsf{A}),\infty)};$$

(ii) if $\mathsf{D} = 0$*, then, for any* $a > \rho_{\max}(\mathsf{A})$, $\mathsf{T}_\Sigma|\mathbb{A}_{[a,\infty)} \in \mathsf{H}^2(\mathbb{A}_{[a,\infty)}; \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}))$*;*

(iii) for any $a > \rho_{\max}(\mathsf{A})$, $\mathsf{T}_\Sigma|\mathbb{A}_{[a,\infty)} \in \mathsf{H}^\infty(\mathbb{A}_{[a,\infty)}; \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}))$.

Proof (i) The proof of Proposition 7.1.12(i) applies here as well.

(ii) Note that $I^\infty(\mathrm{pir}_\Sigma) \supseteq (\rho_{\max}(\mathsf{A}), \infty)$. Therefore, if $a > \rho_{\max}(\mathsf{A})$ then

$$\|a^{-k}\mathrm{pir}_\Sigma(k\Delta)\| \leq Ma^{-k}\rho_{\max}^k, \qquad k \in \mathbb{Z}_{\geq 0}.$$

Since $\frac{\rho_{\max}}{a} < 1$, $k\Delta \mapsto a^{-k}\mathrm{pir}_\Sigma(k\Delta)$ is in $\ell^2(\mathbb{Z}_{\geq 0}(\Delta); \mathrm{L}(\mathsf{U}; \mathsf{Y}))$. By Theorem IV-9.2.15 we conclude that $\mathscr{L}^2_C(\mathrm{pir}_\Sigma) \in \mathsf{H}^2(\mathbb{A}_{[a,\infty)}; \mathrm{L}(\mathsf{U}_{\mathbb{C}}; \mathsf{Y}_{\mathbb{C}}))$, and so the result follows from the Theorem 7.2.11.

(iii) This follows by definition and the preceding part of the proof. ■

induced norm on H^2

Exercises

7.2.1 Show that the restriction of the complexification of a discrete-time convolution system to real inputs agrees with the original system.

7.2.2 Show that the restriction of the complexification of a linear discrete-time difference input/output system to real inputs agrees with the original system.

7.2.3 Show that the restriction of the complexification of a linear discrete-time state space system to real inputs agrees with the original system.

7.2.4 For the discrete-time delay of Exercise 6.9.3, answer the following questions:

(a) compute its transfer function;

(b) comment on the causality of the system, given its transfer function.

7.2.5 For the discrete-time sliding averager of Exercise 6.9.5, determine its transfer function.

7.2.6 For the exponential smoother of Exercise 6.9.8, compute its transfer function.

7.2.7 For the autoregressive model of Exercise 6.9.9, compute its transfer function.

7.2.8 For the moving average process of Exercise 6.9.10, compute its transfer function.

7.2.9 For the autoregressive moving average process of Exercise 6.9.11, compute its transfer function.

7.2.10 For the backward and forward difference systems of Exercise 6.9.4, compute their transfer functions.

7.2.11 For the first three linear discrete-time state space systems of Exercise 6.8.2, compute their transfer functions.

7.2.12 For the linear discrete-time state space system you derived in Exercise 6.8.4, compute its transfer function.

Section 7.3

Polynomial matrix systems

This version: 2022/03/07

Chapter 8

Linear systems: Frequency-domain representations

Closely related to transfer function representations for linear systems are frequency response representations.

Section 8.1

The continuous-continuous Fourier transform and continuous-time linear systems

Section 8.2

The continuous-discrete Fourier transform and discrete-time linear systems

This version: 2022/03/07

Chapter 9

Controllability and observability

The topics in this chapter, controllability and observability are important and venerable parts of control theory. Here we consider these first in the context of general system theory. Then we consider special cases, and finally focus in detail on linear systems.

Section 9.1

Controllability and observability for general systems

Section 9.2

Controllability and observability for systems described by ordinary differential and ordinary difference equations

Section 9.3

Controllability for finite-dimensional linear systems

Section 9.4

Observability for continuous-time state space systems

This version: 2022/03/07

Chapter 10

State space stability

In the preceding two chapters we considered some methods for solving ordinary differential equations, dealing almost exclusively with linear equations. In Section 5.1 we motivated our rationale for this by illustrating that systems of ordinary differential equations can be linearised, although we did not at that time indicate how this process of linearisation might be useful. In this chapter we shall see, among other things, a concrete illustration of why one is interested in linear ordinary differential equations, namely that understanding them can help one understand the stability of systems that are not necessarily linear. Indeed, in this chapter we shall engage in a general discussion of stability, and this connection to linear ordinary differential equations will be just one of the topics considered.

We shall begin our general presentation in Section 10.2 with definitions of various types of stability and examples that illustrate these. We shall give many definitions here, and shall only consider a few of them in any detail subsequently. However, the full slate of definitions is useful for establishing context. In Section 10.3 we consider the stability of systems of linear ordinary differential equations, where the extra structure, especially in the case of systems with constant coefficients, allows a complete description of stability. Two methods, called "Lyapunov's First and Second Method," for stability analysis for systems of (not necessarily linear) ordinary differential equations are considered in Sections 10.7 and 10.5. Lyapunov's First Method allows the determination of the stability of a system of differential equations from its linearisation in some cases.

Contents

Section 10.1

Stability for general systems

Section 10.2

Stability definitions

In this section we state the standard stability definitions for a system of ordinary differential equations. Thus we are working with an ordinary differential equation $\boldsymbol{F}$ with right-hand side

$$\widehat{\boldsymbol{F}}\colon \mathbb{T} \times U \to \mathbb{R}^n,$$

where $U \subseteq \mathbb{R}^n$ is an open subset of $\mathbb{R}^n$. In order to ensure local existence and uniqueness of solutions, we shall make the following assumptions on $\boldsymbol{F}$.

10.2.1 Assumption (Right-hand side assumptions for stability definitions) We suppose that

(i) the map $t \mapsto \widehat{\boldsymbol{F}}(t, x)$ is continuous for each $x \in U$,

(ii) the map $x \mapsto \widehat{\boldsymbol{F}}(t, x)$ is Lipschitz for each $t \in \mathbb{T}$, and

(iii) for each $x \in U$ and for each $r \in \mathbb{R}_{>0}$, there exist continuous functions $g, L\colon \mathbb{T} \to \mathbb{R}_{\geq 0}$ such that

$$\|\widehat{\boldsymbol{F}}(t, y)\| \leq g(t), \qquad (t, y) \in \mathbb{T} \times \mathsf{B}(r, x),$$

and

$$\|\widehat{\boldsymbol{F}}(t, y_1) - \widehat{\boldsymbol{F}}(t, y_2)\| \leq L(t)\|y_1 - y_2\|, \qquad t \in \mathbb{T},\ y_1, y_2 \in \mathsf{B}(r, x). \qquad \bullet$$

10.2.1 Definitions for general systems

The first thing one should address when talking about stability is "stability of what?" Almost always—and always for us—we will be thinking about stability of a solution $t \mapsto \xi_0(t)$ of a system of ordinary differential equations $\boldsymbol{F}$. In all cases, stability of a solution intuitively means that other solutions starting nearby remain nearby at $t \to \infty$. However, this intuitive idea needs to be made precise. As part of this, we make the following definitions.

10.2.2 Definition (ϵ-neighbourhood of a curve) Let $U \subseteq \mathbb{R}^n$ be open, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $\gamma\colon \mathbb{T} \to U$ be a curve. The set

$$\mathcal{N}(\gamma, \epsilon) = \{x \in U \mid \|x - \gamma(t)\| < \epsilon \text{ for some } t \in \mathbb{T}\}$$

is the ***ϵ-neighbourhood*** of γ. $\bullet$

10.2.3 Definition (Distance to a set) Let $U \subseteq \mathbb{R}^n$ be open and let $S \subseteq U$. The function

$$\begin{aligned} \mathrm{d}_S\colon U &\to \mathbb{R}_{\geq 0} \\ x &\mapsto \inf\{\|x - y\| \mid y \in S\} \end{aligned}$$

is the ***distance function to*** $\mathbf{S}$. $\bullet$

We can now state our stability definitions.

10.2.4 Definition (Stability of solutions) Let $\boldsymbol{F}$ be a system of ordinary differential equations satisfying Assumption 10.2.1 and suppose that $\sup \mathbb{T} = \infty$.[1] Let $\xi_0 \colon \mathbb{T}' \to U$ be a solution for $\boldsymbol{F}$, supposing that $\sup \mathbb{T}' = \infty$. The solution ξ_0 is:

(i) ***Lyapunov stable***, or merely ***stable***, if, for any $\epsilon \in \mathbb{R}_{>0}$ and $t_0 \in \mathbb{T}'$, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{x} \in U$ satisfies $\|\xi_0(t_0) - \boldsymbol{x}\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x},$$

is defined on $[t_0, \infty)$ and satisfies $\|\xi(t) - \xi_0(t)\| < \epsilon$ for $t \geq t_0$;

(ii) ***asymptotically stable*** if it is stable and if, for every $t_0 \in \mathbb{T}'$, there exists $\delta \in \mathbb{R}_{>0}$ such that, for $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{x} \in U$ satisfies $\|\xi_0(t_0) - \boldsymbol{x}\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x},$$

is defined on $[t_0, \infty)$ and satisfies $\|\xi(t) - \xi_0(t)\| < \epsilon$ for $t \geq t_0 + T$;

(iii) ***exponentially stable*** if it is stable and if, for every $t_0 \in \mathbb{T}'$, there exists $M, \delta, \sigma \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{x} \in U$ satisfies $\|\xi_0(t_0) - \boldsymbol{x}\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x},$$

is defined on $[t_0, \infty)$ and satisfies $\|\xi(t) - \xi_0(t)\| \leq M\mathrm{e}^{-\sigma(t-t_0)}$;

(iv) ***orbitally stable*** if, for any $\epsilon \in \mathbb{R}_{>0}$ and $t_0 \in \mathbb{T}'$, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{x} \in U$ satisfies $\|\xi_0(t_0) - \boldsymbol{x}\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x},$$

is defined on $[t_0, \infty)$ and satisfies $\xi(t) \in \mathcal{N}(\xi_0, \epsilon)$ for $t \geq t_0$;

(v) ***asymptotically orbitally stable*** if it is orbitally stable and if, for every $t_0 \in \mathbb{T}'$, there exists $\delta \in \mathbb{R}_{>0}$ such that, for $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{x} \in U$ satisfies $\|\xi_0(t_0) - \boldsymbol{x}\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x},$$

is defined on $[t_0, \infty)$ and satisfies $\mathrm{d}_{\mathrm{image}(\xi_0)}(\xi(t)) < \epsilon$ for $t \geq t_0 + T$;

(vi) ***exponentially orbitally stable*** if it is orbitally stable and if, for every $t_0 \in \mathbb{T}'$, there exists $M, \sigma, \delta \in \mathbb{R}_{>0}$ such that, if $\boldsymbol{x} \in U$ satisfies $\|\xi_0(t_0) - \boldsymbol{x}\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x},$$

is defined on $[t_0, \infty)$ and satisfies $\mathrm{d}_{\mathrm{image}(\xi_0)}(\xi(t)) \leq M\mathrm{e}^{-\sigma(t-t_0)}$;

[1] Thus $\mathbb{T}$ is a time-interval that is unbounded on the right, i.e., either $\mathbb{T} = [a, \infty)$ or $\mathbb{T} = (a, \infty)$ for some $a \in \mathbb{R}$.

(vii) ***uniformly Lyapunov stable***, or merely ***uniformly stable***, if, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $(t_0, x) \in \mathbb{T}' \times U$ satisfies $\|\xi_0(t_0) - x\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

is defined on $[t_0, \infty)$ and satisfies $\|\xi(t) - \xi_0(t)\| < \epsilon$ for $t \geq t_0$;

(viii) ***uniformly asymptotically stable*** if it is uniformly stable and if there exists $\delta \in \mathbb{R}_{>0}$ such that, for $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that, if $(t_0, x) \in \mathbb{T}' \times U$ satisfies $\|\xi_0(t_0) - x\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

is defined on $[t_0, \infty)$ and satisfies $\|\xi(t) - \xi_0(t)\| < \epsilon$ for $t \geq t_0 + T$;

(ix) ***uniformly exponentially stable*** if it is uniformly stable and if there exists $M, \sigma, \delta \in \mathbb{R}_{>0}$ such that, if $(t_0, x) \in \mathbb{T}' \times U$ satisfies $\|\xi_0(t_0) - x\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

is defined on $[t_0, \infty)$ and satisfies $\|\xi(t) - \xi_0(t)\| \leq M e^{-\sigma(t-t_0)}$;

(x) ***uniformly orbitally stable*** if, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $(t_0, x) \in \mathbb{T}' \times U$ satisfies $\|\xi_0(t_0) - x\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

is defined on $[t_0, \infty)$ and satisfies $\xi(t) \in \mathcal{N}(\xi_0, \epsilon)$ for $t \geq t_0$;

(xi) ***uniformly asymptotically orbitally stable*** if it is uniformly orbitally stable and if there exists $\delta \in \mathbb{R}_{>0}$ such that, for $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that, if $(t_0, x) \in \mathbb{T}' \times U$ satisfies $\|\xi_0(t_0) - x\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

is defined on $[t_0, \infty)$ and satisfies $\mathrm{d}_{\mathrm{image}(\xi_0)}(\xi(t)) < \epsilon$ for $t \geq t_0 + T$;

(xii) ***uniformly exponentially orbitally stable*** if it is uniformly orbitally stable and if there exists $M, \sigma, \delta \in \mathbb{R}_{>0}$ such that, if $(t_0, x) \in \mathbb{T}' \times U$ satisfies $\|\xi_0(t_0) - x\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

is defined on $[t_0, \infty)$ and satisfies $\mathrm{d}_{\mathrm{image}(\xi_0)}(\xi(t)) \leq M e^{-\sigma(t-t_0)}$;

(xiii) ***unstable*** if it is not stable. •

While this seems like an absurdly large number of definitions, it is made to appear larger by there being a few concepts, represented in all possible combinations. Let us describe the essential dichotomies and trichotomies.

1. *Stable/(asymptotically stable)/(exponentially stable).* The idea of the dichotomy of stable/(asymptotically stable) is that stability has to do with solutions remaining close if their initial conditions are close, while asymptotic stability has to do with solutions with close initial conditions getting closer and closer as time goes by. The notion of exponential stability is similar to that of asymptotic stability, but places some constraints on the rate at which solutions with nearby initial conditions approach one another.
2. *Stable/(orbitally stable).* The stable/(orbitally stable) dichotomy has to do with how one measures the "closeness" of solutions with nearby initial conditions. When dealing with stability, as opposed to orbital stability, one asks that, at all times, solutions remain close. Orbital stability is weaker in that we do not ask that solutions at the same time are close, but rather that one solution at one time is close to another solution, but possibly at a different time.
3. *Stable/(uniformly stable).* The dichotomy here here has to do with the rôle of the initial time t_0 in the definition. In uniform stability, the parameters δ, M, and σ are independent of the initial time t_0, whereas with (nonuniform) stability, these parameters depend on t_0. This is a more or less standard occurrence of the notion of "uniform," and if a reader is encountering this notion for the first time, it is best to acquire a feeling for what it represents.

Now that we have presented our definitions and tried to understand what they mean, let us explore them a little. First let us consider the relationships between the various notions of stability. To do this it is most convenient to arrange the various definitions in a diagram. To control the clutter in the diagram and other places, we use some obvious abbreviations:

(U)S	(uniformly) stable
(U)AS	(uniformly) asymptotically stable
(U)ES	(uniformly) exponentially stable
(U)OS	(uniformly) orbitally stable
(U)AOS	(uniformly) asymptotically orbitally stable
(U)EOS	(uniformly) exponentially orbitally stable

With these abbreviations, we have the diagram in Figure 10.1 illustrating the relationships between the various forms of stability. All of the implications in the diagram follow more or less immediately from the definitions.

Next let us see that, in the case of most interest to us where the solution ξ_0 is an equilibrium solution, the preceding definitions simplify by a factor of $\frac{1}{2}$. Thus, in this discussion, we have an equilibrium state x_0 for F, i.e., $\widehat{F}(t, x_0) = 0$ for all $t \in \mathbb{T}$. In this case, as per Proposition 5.1.5, we have the equilibrium solution ξ_0 defined by $\xi_0(t) = 0$, $t \in \mathbb{T}$. The usual linguistic simplification is to speak, not of the stability of this equilibrium solution, but of the stability of the equilibrium state x_0 since the latter prescribes the former.

The next result records the simplifications that occur in the stability definitions in this case.

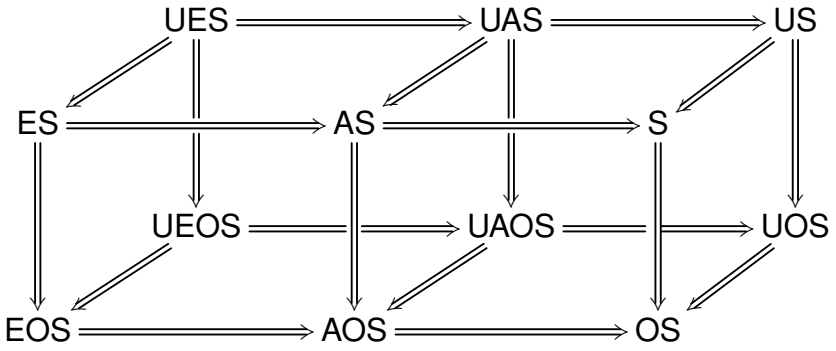


Figure 10.1 Relationships between the various forms of stability

10.2.5 Proposition (Collapsing of stability definitions for equilibria) *Let* $\mathbf{F}$ *be a system of ordinary differential equations satisfying Assumption 10.2.1 and suppose that* $\sup \mathbb{T} = \infty$. *For an equilibrium state* $\mathbf{x}_0$ *for* $\mathbf{F}$, *we have the following implications:*

(i) $\mathsf{OS} \implies \mathsf{S}$;
(ii) $\mathsf{AOS} \implies \mathsf{AS}$;
(iii) $\mathsf{EOS} \implies \mathsf{ES}$;
(iv) $\mathsf{UOS} \implies \mathsf{US}$;
(v) $\mathsf{UAOS} \implies \mathsf{AOS}$;
(vi) $\mathsf{UEOS} \implies \mathsf{UES}$.

In short, all forms of orbital stability are implied by their nonorbital counterparts in the case of equilibrium solutions.

Moreover, if $\mathbf{F}$ *is autonomous, then we additionally have the following implications:*

(vii) $\mathsf{S} \implies \mathsf{US}$;
(viii) $\mathsf{AS} \implies \mathsf{UAS}$;
(ix) $\mathsf{ES} \implies \mathsf{UES}$.

Proof In all cases, this amounts to the observation that, if ξ_0 is the equilibrium solution $\xi_0(t) = x_0$, then $\mathcal{N}(\xi_0, \epsilon) = \mathsf{B}(\epsilon, x_0)$, and so

1. $x \in \mathcal{N}(\xi_0, \epsilon)$ if and only if $\|x - x_0\| < \epsilon$ and
2. $\mathrm{d}_{\mathrm{image}(\xi_0)}(x) = \|x - x_0\|$. ■

For the final assertion of the proposition, we shall explicitly give the proof that $\mathsf{S} \implies \mathsf{US}$, the other implications following using the same idea. Let $\epsilon \in \mathbb{R}_{>0}$. Since x_0 is stable, for $t_0 \in \mathbb{T}$, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $x \in U$ satisfies $\|x - x_0\| < \delta$, the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

exists for $t \geq t_0$ and satisfies $\|\xi(t) - x_0\| < \epsilon$ for $t \geq t_0$. Now let $\hat{t}_0 \in \mathbb{T}$. Then, let $x \in U$ be such that $\|x - x_0\| < \delta$ and let $\xi \colon \mathbb{T} \to U$ and $\hat{\xi} \colon \mathbb{T} \to U$ be the solutions to the initial value problems

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

and

$$\dot{\hat{\xi}}(t) = \widehat{F}(t, \hat{\xi}(t)), \quad \hat{\xi}(t_0') = x,$$

respectively. By Exercise 3.1.19 we have $\hat{\xi}(t) = \xi(t - (\hat{t}_0 - t_0))$. Therefore, $\hat{\xi}$ is defined for $t \geq \hat{t}_0$ and

$$\|\hat{x}(t) - x_0\| = \|x(t - (\hat{t}_0 - t_0)) - x_0\| < \epsilon$$

for $t \geq \hat{t}_0$. This shows that the choice of δ can be made independently of the initial time t_0, and so x_0 is uniformly stable.

We conclude our discussion of stability definitions with a warning of some lurking dangers in these definitions.

10.2.6 Remarks (Caveats concerning stability definitions)

1. First let us provide some good news. For stability of equilibria—by far the most widely used and interesting case—the definitions we give are completely standard and coherent and offer no difficulties in their use.
2. It is often possible to reduce the study of stability of nonequilibrium solutions to the study of equilibria. Let us illustrate how this is done. We suppose that we have an ordinary differential equation F with right-hand side

$$\widehat{F} \colon \mathbb{T} \times U \to \mathbb{R}^n$$

with $\sup \mathbb{T} = \infty$. Let us suppose that we have a solution $\xi_0 \colon \mathbb{T} \to U$ for F, whose stability we wish to examine. In order to do this, we suppose that there exists $r \in \mathbb{R}_{>0}$ such that the "tube"

$$T(r, \xi_0) = \{\xi_0(t) + x' \mid t \in \mathbb{T},\ x' \in \mathsf{B}(r, 0)\}$$

of radius r about ξ_0 is a subset of U. We then define a "time-varying change of coordinates"

$$\begin{aligned} \Phi \colon \mathbb{T} \times T(r, \xi_0) &\to \mathbb{T} \times \mathsf{B}(r, 0) \\ (t, x) &\mapsto (t, x - \xi_0(t)). \end{aligned}$$

We then define a differential equation G with right-hand side

$$\begin{aligned} \widehat{G} \colon \mathbb{T} \times \mathsf{B}(r, 0) &\to \mathbb{R}^n \\ (t, y) &\mapsto \widehat{F} \circ \Phi^{-1}(t, y), \end{aligned}$$

whose state space is $\mathsf{B}(r, 0)$. Note that, if $\xi \colon \mathbb{T}' \to U$ is a solution for F for which $\xi(t) - \xi_0(t) \in \mathsf{B}(r, 0)$, then the function $\eta(t) = \xi(t) - \xi_0(t)$ is a solution for G. Indeed,

$$\dot{\eta}(t) = \dot{\xi}(t) - \dot{\xi}_0(t) = \widehat{F}(t, \xi(t)) = \widehat{F}(t, \eta(t) + \xi_0(t)) = \widehat{G}(t, \eta(t)).$$

Moreover, since $\Phi \circ \xi_0(t) = (t, 0)$ for every $t \in \mathbb{T}$, the solution ξ_0 is mapped to the equilibrium solution $\eta_0 \colon t \mapsto 0$. Therefore, the study of the stability of solution

ξ_0 is reduced to the study of the equilibrium solution at $\mathbf{0}$. In this way, the study of nonequilibrium solutions can sometimes be reduced to the study of equilibrium solutions. Note, also, that, even if $\boldsymbol{F}$ is autonomous, the resulting differential equation $\boldsymbol{G}$ will be nonautonomous.

3. Now for the bad news. For stability of nonequilibrium solutions, there are some possible problems with the definitions that need to be understood. The problems manifest themselves in at least two different ways, and these two ways are not unrelated.
   (a) The ϵ-neighbourhood of a solution is measured using a specific notion of distance coming from the Euclidean norm. It is possible that this is not the most meaningful way of measuring distance, and that, upon choosing another way of measuring distance, one can get inconsistent conclusions when applying stability tests. For example, one might use one method of measuring distance and conclude stability, while another method of measuring distance yields instability. To see examples of where this can happen requires understanding "other ways of measuring distance," and this is not something we shall do here.
   (b) The definitions we give can vary with coordinate systems. That is, one can render a stable (or unstable) system unstable (or stable) by using different coordinates. The reader is asked to explore this in Exercise 10.2.1.

   These caveats need to be kept in mind when working with the stability of nonequilibrium solutions. •

10.2.2 Special definitions for linear systems

We begin with some definitions for stability that are suitable for linear equations.

10.2.7 Definition (Stability for linear systems) Let F be a system of linear homogeneous ordinary differential equations in an n-dimensional $\mathbb{R}$-vector space V and with right-hand side $\widehat{F}(t, x) = A(t)(x)$ for $A\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$. Suppose that $\sup \mathbb{T} = \infty$. Let $\zeta\colon \mathbb{T} \to \mathsf{V}$ be the zero solution $\zeta(t) = 0$, $t \in \mathbb{T}$.

(i) The equation F is S (resp. AS, ES, US, UAS, UES) if the zero solution ζ is S (resp. AS, ES, US, UAS, UES).

The equation F is:

(ii) ***globally stable*** if, for each $t_0 \in \mathbb{T}$, there exists $C \in \mathbb{R}_{>0}$ such that, for $x \in \mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq C\|x\|$ for $t \geq t_0$;

(iii) ***globally asymptotically stable*** if, for each $t_0 \in \mathbb{T}$ and each $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that, for $x \in \mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq \epsilon\|x\|$ for $t \geq t_0 + T$;

(iv) ***globally exponentially stable*** if, for each $t_0 \in \mathbb{T}$, there exists $M, c \in \mathbb{R}_{>0}$ such that, for $x \in \mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq M\|x\|\mathrm{e}^{-c(t-t_0)}$ for $t \geq t_0$;

(v) ***globally uniformly stable*** if there exists $C \in \mathbb{R}_{>0}$ such that, for $(t_0, x) \in \mathbb{T} \times \mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq C\|x\|$ for $t \geq t_0$;

(vi) ***globally uniformly asymptotically stable*** if it is globally uniformly stable and if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that, for $(t_0, x) \in \mathbb{T} \times \mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq \epsilon\|x\|$ for $t \geq t_0 + T$;

(vii) ***globally uniformly exponentially stable*** if there exists $M, c \in \mathbb{R}_{>0}$ such that, for $(t_0, x) \in \mathbb{T} \times \mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq M\|x\|\mathrm{e}^{-c(t-t_0)}$ for $t \geq t_0$. •

Part (i) of the definition is merely the statement of the convention that, when talking about stability for linear ordinary differential equations, one is interested in the stability of the equilibrium state at 0. For this reason, given Proposition 10.2.5, we do not discuss orbital stability for linear equations. The remaining six definitions above are quite particular to linear equations.

We can add obviously to our list of abbreviations.

(U)GS	(uniformly) globally stable
(U)GAS	(uniformly) globally asymptotically stable
(U)GES	(uniformly) globally exponentially stable

There is a little subtlety to the preceding definitions that merits exploration, and this is that (1) the definition of GAS does not include GS as part of the definition, (2) the definition of UGES does not include UGS, whereas (3) for UGAS and UGES, we *do* include the requirement that the equation also be UGS. As we shall see in the proof of Theorem 10.2.9 below, it is the case that GAS $\implies$ GS. It is obvious from the definition that UGES $\implies$ UGS. However, it is *not* true that UGS can be omitted in the definitions of UGAS and UGES, as the following example shows.

10.2.8 Example (UGS must be a part of the definition of UGAS and UGES) We shall construct a system of linear homogeneous ordinary differential equations F in $\mathsf{V} = \mathbb{R}$ with right hand-side $\widehat{F}(t, x) = a(t)x$ and with the following properties:

1. F is not UGS;
2. for $\epsilon \in \mathbb{R}_{>0}$ there exists $T \in \mathbb{R}_{>0}$ with the property that, for $(t_0, x) \in \mathbb{T} \times \mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = a(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $|\xi(t)| < \epsilon|x|$ for $t \geq t_0 + T$.

The example is a little convoluted.

We take $\mathbb{T} = \mathbb{R}_{\geq 0}$ and define $a \colon \mathbb{T} \to \mathbb{R}$ in the following way.

1. Define sequences $(a_k)_{k \in \mathbb{Z}_{\geq 0}}$, $(b_k)_{k \in \mathbb{Z}_{\geq 0}}$, and $(\Delta_k)_{k \in \mathbb{Z}_{\geq 0}}$ as follows:
   (a) $\Delta_k = 2^{-k-1}$, $k \in \mathbb{Z}_{\geq 0}$;
   (b) $b_k = k2^{k+1}$, $k \in \mathbb{Z}_{\geq 0}$;
   (c) define $a_1 = 1$ and then define a_k, $k \geq 2$, by

$$b_{k-1}\Delta_{k-1} - a_k(1 - \Delta_k) + b_k\Delta_k + b_{k+1}\Delta_{k+1} = -1.$$

2. If $t \in \mathbb{T}$, let $k \in \mathbb{Z}_{\geq 0}$ be such that $t \in [k, k + 1)$, and then define

$$a(t) = \begin{cases} -a_k, & t \in [k, k + \Delta_{k+1}), \\ b_k, & t \in [k + \Delta_{k+1}, k + 1). \end{cases}$$

Note that a is not continuous, however, it can be modified to be continuous and still have the desired properties.

To show that F, defined by a, has the desired properties, we first show that F has the property 1 above. For $k \in \mathbb{Z}_{\geq 0}$ define $t_k = k + 1$ and $t_{0,k} = k + \Delta_k$. Let $x = 1 \in \mathsf{V}$ and let $\xi_k \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}_k(t) = a(t)\xi_k(t), \quad \xi_k(t_{0,k}) = x,$$

for $k \in \mathbb{Z}_{\geq 0}$. Note that

$$|\xi_k(t_k)| = \left| x\mathrm{e}^{-\int_{t_{0,k}}^{t_k} a(\tau)\,\mathrm{d}\tau} \right| = |x|\mathrm{e}^{k}.$$

This prohibits uniform global stability for F.

Next we show that F has the property 2 above. Thus let $\epsilon \in \mathbb{R}_{>0}$ and define $T \in \mathbb{Z}_{>0}$ such that $\mathrm{e}^{-(T-3)} < \epsilon$. Let $t_0 \in \mathbb{T}$ and let $t \geq t_0 + T$. Let $k_1 \in \mathbb{Z}_{\geq 0}$ be such that $t_0 \in [k_1, k_1 + 1)$, let $k_2 \in \mathbb{Z}_{>0}$ be such that $t \in [k_2, k_2 + 1)$. Note that

$$t - t_0 \geq T \implies k_2 - k_1 + 1 > T \implies k_2 - k_1 - 2 > T - 3.$$

Now we estimate

$$\begin{aligned}\int_{t_0}^{t_0+t} a(\tau)\,\mathrm{d}\tau &= \int_{t_0}^{k_1+1} a(\tau)\,\mathrm{d}\tau + \sum_{k=k_1+1}^{k_2-1} \int_k^{k+1} a(\tau)\,\mathrm{d}\tau + \int_{k_2}^{t} a(\tau)\,\mathrm{d}\tau\\ &\le b_{k_1}\Delta_{k_1} + \sum_{k=k_1+1}^{k_2-1} (-a_k(1-\Delta_k) + b_k\Delta_k) + b_{k_2}\Delta_{k_2}\\ &\le \sum_{k_1+1}^{k_2-1} (b_{k-1}\Delta_{k-1} - a_k(1-\Delta_k) + b_k\Delta_k + b_{k+1}\Delta_{k+1})\\ &= -\sum_{k=k_1+1}^{k_2-1} 1 = -(k_2-k_1-2) < -(T-3).\end{aligned}$$

Now let $x \in \mathsf{V}$ and let $\xi\colon \mathbb{T} \to \mathsf{V}$ satisfy the initial value problem

$$\dot{\xi}(t) = a(t)\xi(t), \quad \xi(t_0) = x.$$

Then

$$|\xi(t)| = \left| x\mathrm{e}^{-\int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau} \right| \le |x|\mathrm{e}^{(T-3)} < \epsilon|x|,$$

for $t \ge t_0 + T$, giving the desired conclusion. •

Let us further explore these definitions by (1) exploring their relationships with the notions of stability from Definition 10.2.4 and (2) exploring the relationships between these new notions.

First the first. . .

10.2.9 Theorem (Equivalence of stability and global stability for linear ordinary differential equations) *Consider the system of linear homogeneous ordinary differential equations* F *with right-hand side (10.5) and suppose that* $\mathrm{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V};\mathsf{V})$ *is continuous. Suppose that* $\sup \mathbb{T} = \infty$. *Then* F *is* S *(resp.* AS, ES, US, UAS, UES*) if and only if it is* GS *(resp.* GAS, GES, GUS, GUAS, GUES*).*

Proof (GS $\implies$ S) Let $t_0 \in \mathbb{T}$ and let $C \in \mathbb{R}_{>0}$ be such that the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \le C\|x\|$ for $t \ge t_0$. Let $\epsilon \in \mathbb{R}_{>0}$ and take $\delta = \frac{\epsilon}{C}$. Now let $x \in \mathsf{V}$ satisfy $\|x\| < \delta$ and let $\xi\colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x.$$

We then have

$$\|\xi(t)\| \le C\|x\| = \frac{\epsilon}{\delta}\|x\| \le \epsilon,$$

for $t \ge t_0$, giving stability of F.

(S $\implies$ GS) Let $t_0 \in \mathbb{T}$ and let $\delta \in \mathbb{R}_{>0}$ have the property that, if $\|x\| \leq \delta$, then the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq 1$ for $t \geq t_0$. Define $C = \delta^{-1}$. Now let $x \in \mathsf{V}$ and let $\xi \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x. \tag{10.1}$$

First suppose that $x \neq 0$ and define $\hat{x} = \delta \frac{x}{\|x\|}$ so that $\|\hat{x}\| = \delta$. Thus the solution $\hat{\xi} \colon \mathbb{T} \to \mathsf{V}$ to the initial value problem

$$\dot{\hat{\xi}}(t) = A(t)(\hat{\xi}(t)), \quad \hat{\xi}(t_0) = \hat{x},$$

satisfies $\|\hat{\xi}(t)\| \leq 1$ for $t \geq t_0$. However,

$$\xi(t) = \Phi_A(t, t_0)(x) = \Phi_A(t, t_0)\left(\frac{\|x\|}{\delta}\hat{x}\right) = \frac{\|x\|}{\delta}\Phi_A(t, t_0)(\hat{x}) = C\|x\|\hat{\xi}(t).$$

Therefore,

$$\|\xi(t)\| = C\|x\|\,\|\hat{\xi}(t)\| \leq C\|x\|.$$

If $x = 0$ this relation clearly holds since the solution to the initial value problem (10.1) is simply $\xi(t) = 0$, $t \in \mathbb{T}$. Thus F is globally stable.

(GAS $\implies$ AS) First we show that GAS $\implies$ GS (which implies S as we have already proved). Let $t_0 \in \mathbb{T}$, let $x \in \mathsf{V}$, and let $\xi \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x.$$

First suppose that $x \neq 0$. Since

$$\lim_{t \to \infty} \frac{\|\xi(t)\|}{\|x\|} = 0$$

and since ξ is continuous (indeed, of class C^1), it follows that $t \mapsto \frac{\|\xi(t)\|}{\|x\|}$ is bounded, i.e., there exists $C \in \mathbb{R}_{>0}$ such that

$$\frac{\|\xi(t)\|}{\|x\|} \leq C \quad \implies \quad \|\xi(t)\| \leq C\|x\|.$$

This relationship also holds when $x = 0$, we conclude global stability of F.

Now let $t_0 \in \mathbb{T}$ and take $\delta = \frac{1}{2}$. Let $\epsilon \in \mathbb{R}_{>0}$, and take $T \in \mathbb{R}_{>0}$ such that the solution $\xi \colon \mathbb{T} \to \mathsf{V}$ to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq \epsilon\|x\|$ for $t \geq t_0 + T$. Now suppose that $\|x\| < \delta = \frac{1}{2}$, and let $\xi \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x.$$

Then, for $t \geq t_0 + T$,

$$\|\xi(t)\| \leq \epsilon\|x\| < \epsilon.$$

This shows that F is asymptotically stable.

(AS $\implies$ GAS) Let $t_0 \in \mathbb{T}$ and let $\delta \in \mathbb{R}_{>0}$ have the property that, given $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that, if $\|x\| < \delta$, then the solution $\xi \colon \mathbb{T} \to \mathsf{V}$ to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| < \epsilon$ for $t \geq t_0 + T$.

Let $\epsilon \in \mathbb{R}_{>0}$ and let $T \in \mathbb{R}_{>0}$ be such that, if $\|x\| < \delta$, then the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| < \frac{\epsilon\delta}{2}$ for $t \geq t_0 + T$. Let $x \in \mathsf{V}$ and let $\xi \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x.$$

Let $\hat{x} = \delta\frac{x}{2\|x\|}$ and let $\hat{\xi} \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\hat{\xi}}(t) = A(t)(\hat{\xi}(t)), \quad \hat{\xi}(t_0) = \hat{x}.$$

Since $\|\hat{x}\| = \frac{\delta}{2} < \delta$, $\|\hat{\xi}(t)\| < \frac{\epsilon\delta}{2}$ for $t \geq t_0 + T$. We also have

$$\xi(t) = \Phi_A(t, t_0)(x) = \Phi_A(t, t_0)\left(\frac{2\|x\|}{\delta}\hat{x}\right) = \frac{2\|x\|}{\delta}\Phi_A(t, t_0)(\hat{x}) = \frac{2\|x\|}{\delta}\|x\|\hat{\xi}(t).$$

Thus

$$\|\xi(t)\| \leq \frac{2}{\delta}\|x\|\,\|\hat{\xi}(t)\| < \epsilon\|x\|,$$

for $t \geq t_0 + T$, and so F is globally asymptotically stable.

(GES $\implies$ ES) First we note that GES $\implies$ GS (which implies S, as we have already seen). Indeed, the proof that GAS $\implies$ GS we gave above also applies if we replace "GAS" with "GES."

Now let $t_0 \in \mathbb{T}$ and let $\tilde{M}, \tilde{\sigma} \in \mathbb{R}_{>0}$ be such that, for $v \in \mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq \tilde{M}\|x\|\mathrm{e}^{-\tilde{\sigma}(t-t_0)}$ for $t \geq t_0$. Now let $\delta = \frac{1}{2}$ and take $M = \tilde{M}$ and $\sigma = \tilde{\sigma}$. Then, for $\|x\| < \delta = \frac{1}{2}$, let $\xi \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x.$$

We then have

$$\|\xi(t)\| \leq \tilde{M}\|x\|\mathrm{e}^{\tilde{\sigma}(t-t_0)} \leq M\mathrm{e}^{-\sigma(t-t_0)},$$

showing that F is exponentially stable.

(ES $\implies$ GES) Let $t_0 \in \mathbb{T}$ and let $\tilde{M}, \delta, \tilde{\sigma} \in \mathbb{R}_{>0}$ be such that, if $\|x\| < \delta$, then the solution $\xi \colon \mathbb{T} \to \mathsf{V}$ to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq \tilde{M}e^{-\tilde{\sigma}(t-t_0)}$ for $t \geq t_0$.

Take $M = \frac{2\tilde{M}}{\delta}$ and $\sigma = \tilde{\sigma}$. Now let $x \in \mathsf{V}$ and let $\xi \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x.$$

Let $\hat{x} = \delta\frac{x}{2\|x\|}$ and let $\hat{\xi} \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\hat{\xi}}(t) = A(t)(\hat{\xi}(t)), \quad \hat{\xi}(t_0) = \hat{x}.$$

Since $\|\hat{x}\| = \frac{\delta}{2} < \delta$, $\|\hat{\xi}(t)\| \leq \tilde{M}e^{-\tilde{\sigma}(t-t_0)}$ for $t \geq t_0$. Then, as in the proof that $\mathsf{AS} \implies \mathsf{GAS}$,

$$\xi(t) = \frac{2\|x\|}{\delta}\hat{\xi}(t),$$

and so

$$\|\xi(t)\| = \frac{2}{\delta}\|x\| \, \|\hat{\xi}(t)\| \leq \frac{2\tilde{M}}{\delta}\|x\|e^{-\tilde{\sigma}(t-t_0)} = M\|x\|e^{-\sigma(t-t_0)},$$

for $t \geq t_0$, showing that F is globally exponentially stable.

The remainder of the proof concerns the results we have already proved, but with the property "uniform" being applied to all hypotheses and conclusions. The proofs are entirely similar to those above. We shall, therefore, only work this out in one of the three cases, the other two following in an entirely similar manner.

($\mathsf{GUS} \implies \mathsf{US}$) Let $C \in \mathbb{R}_{>0}$ be such that the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq C\|x\|$ for $t \geq t_0$. Let $\epsilon \in \mathbb{R}_{>0}$ and take $\delta = \frac{\epsilon}{C}$. Now let $x \in \mathsf{V}$ satisfy $\|x\| < \delta$ and let $\xi \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x.$$

We then have

$$\|\xi(t)\| \leq C\|x\| = \frac{\epsilon}{\delta}\|x\| \leq \epsilon,$$

for $t \geq t_0$, giving stability of F.

($\mathsf{US} \implies \mathsf{GUS}$) Let $\delta \in \mathbb{R}_{>0}$ have the property that, if $\|x\| \leq \delta$, then the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| \leq 1$ for $t \geq t_0$. Define $C = \delta^{-1}$. Now let $x \in \mathsf{V}$ and let $\xi \colon \mathbb{T} \to \mathsf{V}$ be the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x. \tag{10.2}$$

First suppose that $x \neq 0$ and define $\hat{x} = \delta\frac{x}{\|x\|}$ so that $\|\hat{x}\| = \delta$. Thus the solution $\hat{\xi} \colon \mathbb{T} \to \mathsf{V}$ to the initial value problem

$$\dot{\hat{\xi}}(t) = A(t)(\hat{\xi}(t)), \quad \hat{\xi}(t_0) = \hat{x},$$

satisfies $\|\hat{\xi}(t)\| \leq 1$ for $t \geq t_0$. However,

$$\xi(t) = \Phi_A(t, t_0)(x) = \Phi_A(t, t_0)\left(\frac{\|x\|}{\delta}\hat{x}\right) = \frac{\|x\|}{\delta}\Phi_A(t, t_0)(\hat{x}) = C\|x\|\hat{\xi}(t).$$

Therefore,

$$\|\xi(t)\| = C\|x\|\,\|\hat{\xi}(t)\| \leq C\|x\|.$$

If $x = 0$ this relation clearly holds since the solution to the initial value problem (10.2) is simply $\xi(t) = 0$, $t \in \mathbb{T}$. Thus F is globally stable. ■

Now let us examine some relationships between these special notions of stability for linear equations.

10.2.10 Theorem (Equivalence of uniform asymptotic and uniform exponential stability for linear ordinary differential equations) *Consider the system of linear homogeneous ordinary differential equations* F *with right-hand side (10.5) and suppose that* $\mathrm{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$ *is continuous. Suppose that* $\sup \mathbb{T} = \infty$. *Then* F *is* UGAS *if and only if it is* UGES.

Proof It is clear that UGES implies UGAS, so we will only prove the converses.

(UGAS $\implies$ UGES) By definition of uniform asymptotic stability, there exists $C, T \in \mathbb{R}_{>0}$ such that

$$\|\Phi_A(t, t_0)(x)\| \leq C\|x\|$$

and

$$\|\Phi_A(t, t_0)(x)\| \leq \frac{1}{2}\|x\|, \qquad t \geq t_0 + T,$$

for all $(t_0, x) \in \mathbb{T} \times \mathsf{V}$. Then, for $k \in \mathbb{Z}_{>0}$, $(t_0, x) \in \mathbb{T} \times \mathsf{V}$, and $t \geq t_0 + kT$,

$$\begin{aligned}&\|\Phi_A(t, t_0)(x)\|\\ &\qquad = \|\Phi_A(t, t_0 + kT) \circ \Phi_A(t_0 + kT, t_0 + (k-1)T) \circ \cdots \circ \Phi_A(t_0 + T, t_0)(x)\| \leq \frac{C}{2^k}\|x\|.\end{aligned}$$

Now define $M = C$ and $\sigma = \frac{\ln 2}{T}$ and let $(t_0, x) \in \mathbb{T} \times \mathsf{V}$ and $t \geq t_0$. Then $t \in [t_0, t_0 + kT)$ for some uniquely defined $k \in \mathbb{Z}_{>0}$, and then

$$\|\Phi_A(t, t_0)(x)\| \leq \frac{C}{2^k}\|x\| = M\mathrm{e}^{-\sigma kT} \leq M\mathrm{e}^{\sigma(t-t_0)},$$

as desired. ■

Note that the conclusions of the theorem are not true if we eliminate "uniform" in the hypotheses.

10.2.11 Example (Global asymptotic stability does not imply global exponential stability) We consider the system of linear homogeneous ordinary differential equations F in $\mathsf{V} = \mathbb{R}$ and with

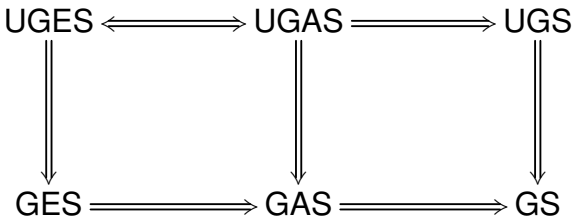
$$\widehat{F}(t, x) = -\frac{x}{t},$$

and we take $\mathbb{T} = [1, \infty)$. This equation can be solved using the methods of Section 4.1.1 to give

$$\xi(t) = \frac{t_0 \xi(t_0)}{t},$$

and from this we conclude that, for any initial condition $\xi(t_0)$, $\lim_{t\to\infty} \xi(t) = 0$ (i.e., we have GAS) but that we do not have exponential stability. •

Let us summarise the relationships between the various notions of stability for systems of linear homogeneous ordinary differential equations in a diagram:



The arrows not present in the diagram represent implications that do not, in fact, hold.

10.2.3 Examples

In this section, we give some examples to illustrate some of the ways in which the different notions of stability are separated in practice.

10.2.12 Example (Stable versus unstable versus asymptotically stable I) We consider the ordinary differential equation F with state space $U = \mathbb{R}$ and with right-hand side $\widehat{F}(t, x) = ax$ with $a \in \mathbb{R}$. This is a simple linear ordinary differential equation and has solution $\xi(t) = \xi(t_0)e^{a(t-t_0)}$. We shall consider the stability of the equilibrium point $x_0 = 0$. We have three cases.

1. $a < 0$: In this case we note two things. First of all, $|\xi(t)| \leq |\xi(t_0)|$ for $t \geq t_0$, from which we conclude that the equilibrium at $x_0 = 0$ is stable. (Formally, let $\epsilon \in \mathbb{R}_{>0}$. Then, if we take $\delta = \epsilon$, we have

$$|\xi(t_0) - 0| \leq \delta \implies |\xi(t) - 0| < \epsilon, \qquad t \geq t_0.$$

which is what is required to prove stability of the equilibrium $x_0 = 0$.) Also, $\lim_{t\to\infty} |\xi(t) - 0| = 0$, which gives asymptotic stability of $x_0 = 0$. Moreover, in this case we also have $|\xi(t)| = |\xi(t_0)|e^{a(t-t_0)}$, and so we further have exponential stability.

2. $a = 0$: Here we have $\xi(t) = \xi(t_0)$ for all t. Therefore, we have stability, but not asymptotic stability of the equilibrium $x_0 = 0$. (Formally, let $\epsilon \in \mathbb{R}_{>0}$. Then, taking $\delta = \epsilon$, we have

$$|\xi(t_0) - 0| < \delta \implies |\xi(t) - 0| < \epsilon, \qquad t \geq t_0.)$$

3. $a > 0$: Here, as long as $\xi(t_0) \neq 0$, we have $\lim_{t\to\infty}|\xi(t)| = \infty$, and this suffices to show that the equilibrium $x_0 = 0$ is unstable. (Formally, we must show that there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for any $\delta \in \mathbb{R}_{>0}$ there exists $\xi(t_0) \in \mathbb{R}$ and $T \in \mathbb{R}_{>0}$ such that, $|\xi(t_0)| < \delta$ and $|\xi(t_0 + T)| \geq \epsilon$. We can take $\epsilon = 1$ and, given $\delta \in \mathbb{R}_{>0}$, we can take $\xi(t_0) = \frac{\delta}{2}$ and $T \in \mathbb{R}_{>0}$ such that $\mathrm{e}^{aT} \geq \frac{2}{\delta}$.) •

10.2.13 Example (Stable versus unstable versus asymptotically stable II) We consider another example illustrating the same trichotomy as the preceding example, but one that generates some pictures that one can keep in mind when thinking about concepts of stability. We consider the ordinary differential equation $\boldsymbol{F}$ with state space $U = \mathbb{R}^2$ and with right-hand side $\widehat{F}(t, (x_1, x_2)) = (x_2, -x_1 - 2\delta x_2)$ for $|\delta| < 1$. We shall be concerned with the stability of the equilibrium point $\boldsymbol{x}_0 = (0, 0)$. Solutions $\boldsymbol{\xi}\colon \mathbb{T} \to \mathbb{R}^2$ satisfy

$$\begin{aligned}\dot{\xi}_1(t) &= \xi_2(t),\\ \dot{\xi}_2(t) &= -\xi_1(t) - 2\delta\xi_2(t).\end{aligned}$$

This is a linear homogeneous ordinary differential equation with constant coefficients determined by the matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & -\delta \end{bmatrix}.$$

We compute the eigenvalues of $\boldsymbol{A}$ to be

$$\lambda_1 = -\delta + \mathrm{i}\sqrt{1 - \delta^2}, \quad \lambda_2 = -\delta - \mathrm{i}\sqrt{1 - \delta^2}.$$

Thus we have two distinct complex eigenvalues. We can then apply Procedures 5.2.23 and 5.2.26 to compute

$$\mathrm{e}^{\boldsymbol{A}t} = \mathrm{e}^{-\delta t}\begin{bmatrix} \cos(\sqrt{1-\delta^2}t) + \frac{\delta}{\sqrt{1-\delta^2}}\sin(\sqrt{1-\delta^2}t) & \frac{1}{\sqrt{1-\delta^2}}\sin(\sqrt{1-\delta^2}t) \\ -\frac{1}{\sqrt{1-\delta^2}}\sin(\sqrt{1-\delta^2}t) & \cos(\sqrt{1-\delta^2}t) + \frac{\delta}{\sqrt{1-\delta^2}}\sin(\sqrt{1-\delta^2}t) \end{bmatrix}.$$

In Figure 10.2 we plot the parameterised curves in (x_1, x_2)-space in what we shall in Section 5.5 call "phase portraits. Without going through the details of the analysis, we shall simply make the following observations.

1. $\delta > 0$: Here we see that $\boldsymbol{x}_0 = (0, 0)$ is asymptotically stable.
2. $\delta = 0$: Here we see that $\boldsymbol{x}_0 = (0, 0)$ is stable, but not asymptotically stable.

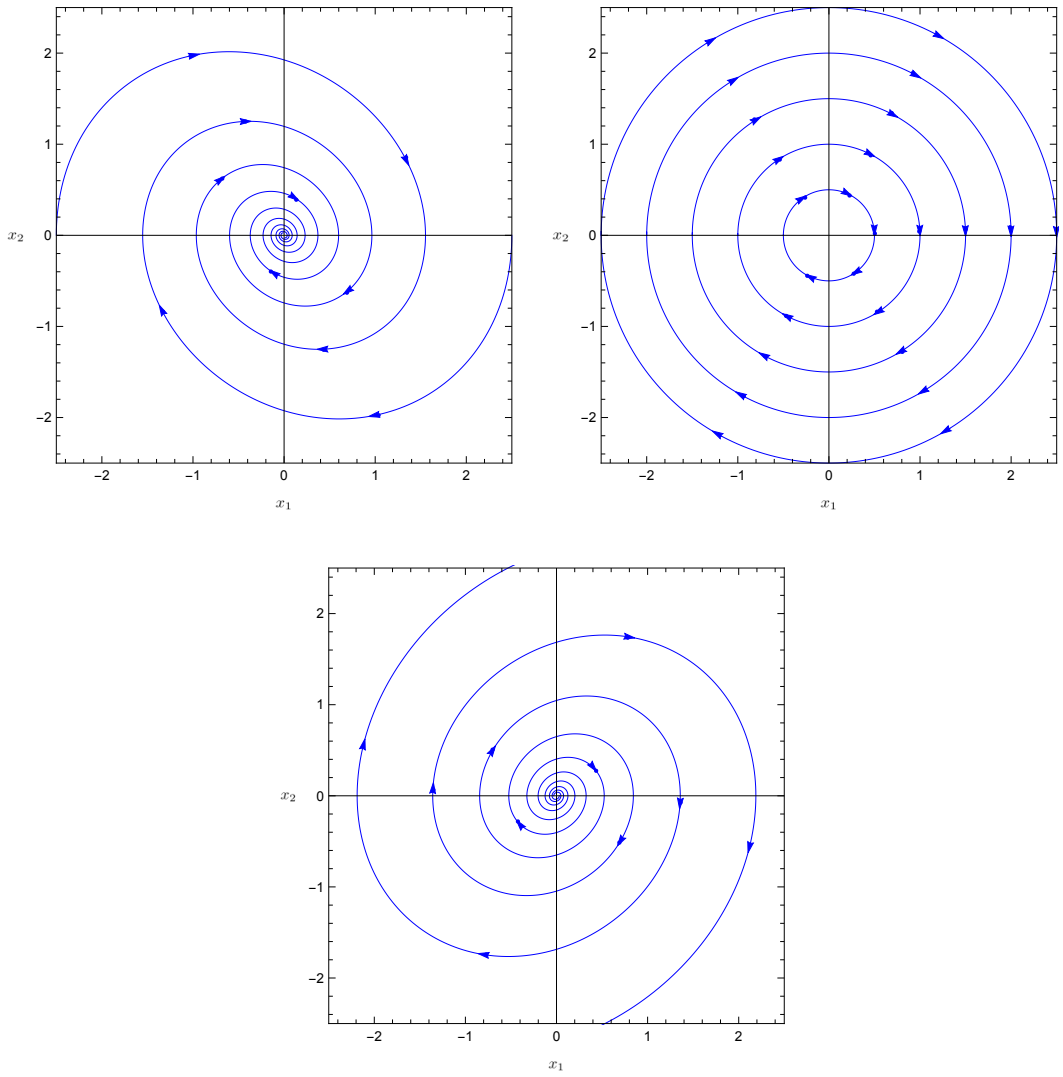


Figure 10.2 Phase portraits for $\widehat{F}(t,(x_1,x_2)) = (x_2, -x_1 - \delta x_2)$ for $\delta < 0$ (top left), $\delta = 0$ (top right), and $\delta > 0$ (bottom)

3. $\delta < 0$: Here we see that $\boldsymbol{x}_0 = (0,0)$ is unstable.

One can look at the behaviour of solutions in Figure 10.2 to convince oneself of the validity of these conclusions. •

The definitions we give in Definition 10.2.4 are "local.[2]" This means that they only give conclusions about the behaviour of solutions nearby the reference solution. Our preceding two examples might give one the impression that they hold globally, but this is not the case, as we illustrate in the next two examples.

[2]Indeed, the definitions we give are often prefixed by "local."

10.2.14 Example (Stable does not mean "globally stable" I) Here we consider the ordinary differential equation F with state space $U = \mathbb{R}$ and right-hand side $\widehat{F}(t, x) = x - x^3$. We will look at the stability of the equilibria for this differential equation. According to Proposition 5.1.5, a state $x_0 \in \mathbb{R}$ is an equilibrium state if and only if $x_0 - x_0^3 = 0$, which gives the three equilibria $x_- = -1$, $x_0 = 0$, and $x_+ = 0$. We shall subsequently see how to rigorously prove the stability of these three equilibria, but here we shall argue heuristically. In Figure 10.3 we graph the right-hand side as a

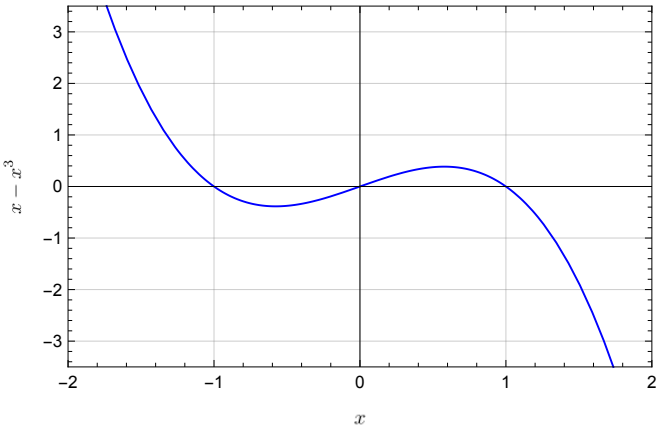


Figure 10.3 The right-hand side $x - x^3$

function of x. From this graph, we make the following conclusions.

1. x_0 is unstable: We see that, when $x > x_0 = 0$ and x is nearby $x_0 = 0$, that $\widehat{F}(t, x) > 0$, Therefore, if $\xi(t_0) > 0$ and is nearby 0, then $\xi(t)$ will become "more positive." In similar manner, if $\xi(t_0) < 0$ and is nearby 0, then $\xi(t)$ will become "more negative." Thus all solutions nearby 0 "move away" from 0.
2. $x_\pm$ are asymptotically stable: Here the opposite phenomenon occurs as compared to x_0. When $x > x_\pm$ and x is nearby $x_\pm$, then $\widehat{F}(t, x) < 0$. Therefore, if $\xi(t_0) > x_\pm$ and is nearby $x_\pm$, then $\xi(t)$ will "move towards" $x_\pm$. In similar manner, if $\xi(t_0) < x_\pm$ and is nearby $x_\pm$, then $\xi(t)$ will again "move towards" $x_\pm$.

The point is that our conclusions about stability for all three equilibria hold only for initial conditions nearby the equilibria. Moreover, the stability is different for different equilibria. •

10.2.15 Example (Stable does not mean "globally stable" II) The example here illustrates a similar phenomenon as the preceding example, but does so while producing some useful pictures. The ordinary differential equation we consider has state space $U = \mathbb{R}^2$ with right-hand side $\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = (x_2, x_1 - x_1^3 - \frac{1}{2}x_2)$. Thus solutions

$\xi\colon \mathbb{T} \to \mathbb{R}^2$ satisfy

$$\begin{aligned}\dot{\xi}_1(t) &= \xi_2(t)\\ \dot{\xi}_2(t) &= \xi_1(t) - \xi_1(t)^3 - \tfrac{1}{2}\xi_2(t).\end{aligned}$$

We will consider the stability of the equilibria for $\boldsymbol{F}$. By Proposition 5.1.5, an equilibrium $\boldsymbol{x}_0 = (x_{01}, x_{02})$ will satisfy

$$\begin{aligned}0 &= x_{02},\\ 0 &= x_{01} - x_{01}^3 - \frac{1}{2}x_{01},\end{aligned}$$

which gives the three equilibrium points $\boldsymbol{x}_0 = (0,0)$, $\boldsymbol{x}_- = (-1,0)$, and $\boldsymbol{x}_+ = (0,1)$. In Figure 10.4 we show a few parameterised solutions for $\boldsymbol{F}$ in the (x_1, x_2)-plane.

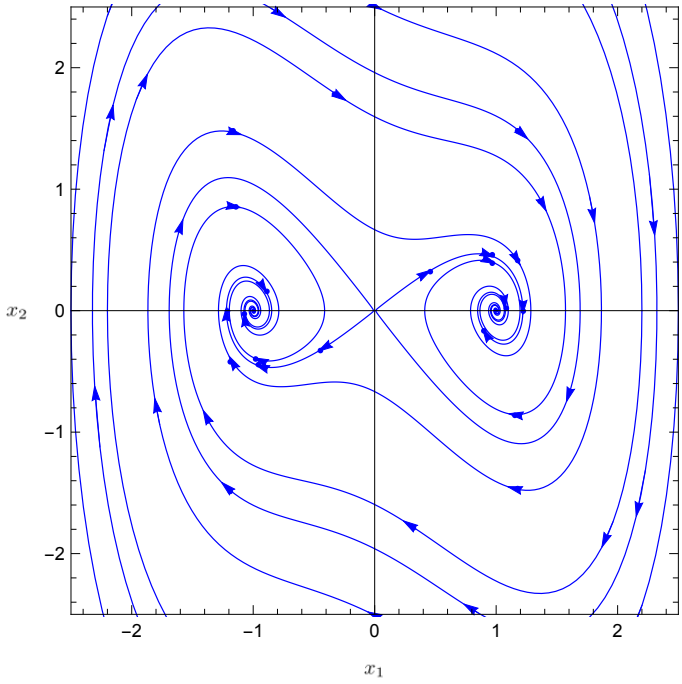


Figure 10.4 Phase portrait for $\widehat{F}(t,(x_1,x_2)) = (x_2, x_1 - x_1^3 - \frac{1}{2}x_2)$

From this figure we deduce that $\boldsymbol{x}_0$ is unstable and $\boldsymbol{x}_\pm$ is asymptotically stable. •

The reader will have noticed that "stable" is included in the definition of "asymptotically stable." It seems like this might be redundant, but it is not as the next example indicates.

10.2.16 Example (Why "stable" is part of the definition of "asymptotically stable") We work with the ordinary differential equation $\boldsymbol{F}$ with state space $\mathbb{R}^2$ and with

$$\widehat{\boldsymbol{F}}(t,(x_1,x_2)) = \left(\frac{x_1^2(x_2-x_1)+x_2^5}{(x_1^2+x_2^2)(1+(x_1^2+x_2^2)^2)}, \frac{x_2^2(x_2-2x_1)}{(x_1^2+x_2^2)(1+(x_1^2+x_2^2)^2)}\right)$$

This solutions $\xi\colon \mathbb{T} \to \mathbb{R}^2$ for $\boldsymbol{F}$ satisfy

$$\begin{aligned}\dot{\xi}_1(t) &= \frac{\xi_1(t)^2(\xi_2(t)-\xi_1(t))+\xi_2(t)^5}{(\xi_1(t)^2+\xi_2(t)^2)(1+(\xi_1(t)^2+\xi_2(t)^2)^2)}\\ \dot{\xi}_2(t) &= \frac{\xi_2(t)^2(\xi_2(t)-2\xi_1(t))}{(\xi_1(t)^2+\xi_2(t)^2)(1+(\xi_1(t)^2+\xi_2(t)^2)^2)}.\end{aligned}$$

We are interested in the stability of the equilibrium point $\boldsymbol{x}_0 = (0,0)$. In Figure 10.5 we depict the phase portrait for the equation. From the phase portrait, we can

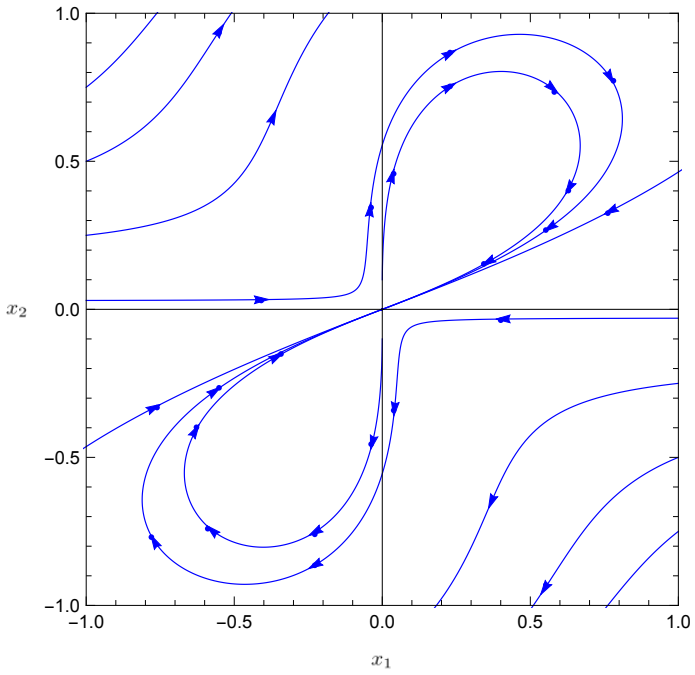


Figure 10.5 Phase portrait for $\widehat{F}(t,x) = \left(\frac{x_1^2(x_2-x_1)+x_2^5}{(x_1^2+x_2^2)(1+(x_1^2+x_2^2)^2)}, \frac{x_2^2(x_2-2x_1)}{(x_1^2+x_2^2)(1+(x_1^2+x_2^2)^2)}\right)$

reasonable say that (1) for any initial condition $\xi(t_0) \in \mathbb{R}^2$, we have $\lim_{t\to\infty} \xi(t) = (0,0)$ and (2) $\boldsymbol{x}_0 = (0,0)$ is not stable. The former can be seen straightaway from Figure 10.5. For the latter, we note that, for any $\epsilon \in \mathbb{R}_{>0}$, no matter how small we choose δ, there is an initial condition satisfying $\|\xi(t_0) - \boldsymbol{x}_0\| < \delta$ for which the

corresponding solution leaves the ball of radius ϵ centred at x_0. Thus stability is required as part of the definition of asymptotic stability in order to rule out this "large deviation" behaviour.[3] •

10.2.17 Example (Asymptotically stable versus exponentially stable) In some of our examples above where an equilibrium is asymptotically stable, it is also exponentially stable. However, this need not be the case always. To illustrate this, we consider the ordinary differential equation with state space $U = \mathbb{R}$ and right-hand side $\widehat{F}(t, x) = -x^3$. In this case, we can argue as in Example 10.2.14 that the equilibrium state at $x_0 = 0$ is asymptotically stable. Let us show that it is not exponentially stable. For $\xi(t_0) \in \mathbb{R}$, we can use the technique of Section 4.1.1 to obtain the solution with this initial condition as

$$\xi(t) = \operatorname{sign}(\xi(t_0)) \left(\frac{1 + 2(t - t_0)\xi(t_0)^2}{\xi(t_0)^2} \right)^{-1/2},$$

where $\operatorname{sign}\colon \mathbb{R} \to \{-1, 0, 1\}$ returns the sign of a real number. The observation we make is that, as $t \to \infty$, $\xi(t)$ decays to zero like $(t - t_0)^{-1/2}$, which prohibits exponential stability. •

10.2.18 Example (Stable versus orbitally stable) As we saw in Proposition 10.2.5, one cannot distinguish between "stable" and "orbitally stable" for equilibria. Therefore, necessarily, if we wish to consider a distinction between these sorts of stability, we need to work with a nonequilibrium solution. The example we give is one that is easily imagined, and we do not rigorously prove our assertions.

We consider the motion of a simple pendulum. This can be thought of as a first-order system of ordinary differential equations with state space $U = \mathbb{R}^2$ and with right-hand side

$$\widehat{\boldsymbol{F}}(t, (x_1, x_2)) = \left(x_2, -\frac{a_g}{\ell} \sin(x_1) \right).$$

Here a_g is acceleration due to gravity and ℓ is the length of the pendulum. Solutions $\boldsymbol{\xi}\colon \mathbb{T} \to \mathbb{R}^2$ satisfy

$$\begin{aligned} \dot{\xi}_1(t) &= \xi_2(t) \\ \dot{\xi}_2(t) &= -\frac{a_g}{\ell} \sin(x_1). \end{aligned}$$

Let us make some (mathematically unproved, but physically "obvious") observations about this equation.

[3] One very often sees the following definition.

Definition A solution ξ_0 of an ordinary differential equation is ***attractive*** if there exists $\delta \in \mathbb{R}_{>0}$ such that, for any ϵ, there exists $T \in \mathbb{R}_{>0}$ for which, if $\|\boldsymbol{\xi}(t_0) - \boldsymbol{\xi}_0(t_0)\| < \delta$, then $\|\boldsymbol{\xi}(t) - \boldsymbol{\xi}_0(t)\| < \epsilon$ for $t \geq t_0 + T$. •

One can then say that "asymptotic stability" means "stable" and "attractive."

1. For small oscillations of the pendulum, the period of the oscillation is $2\pi\sqrt{\frac{\ell}{a_g}}$.
2. As the amplitude of the oscillation becomes large (approaching π), the period becomes large. Indeed, for oscillations with amplitude exactly π, the period is "∞." Let us be clear what this means. There is a motion of the pendulum where, at "$t = -\infty$," the pendulum is upright at rest, and then begins to fall. It will fall and then swing up to the upright configuration at rest, getting there at "$t = \infty$."
3. For amplitudes between 0 and π, the period will grow monotonically from $2\pi\sqrt{\frac{\ell}{a_g}}$ to ∞. There is, in fact, a precise formula for this, and it is

$$T(\theta_0) = 4\sqrt{\frac{\ell}{a_g}}\int_0^{\pi/2}\frac{1}{(1-\sin^2(\frac{\theta_0}{2})\sin^2(\phi))^{1/2}}\,\mathrm{d}\phi,$$

where θ_0 is the amplitude of the oscillation. This integral, while not expressible in terms of anything you know about, *is* expressible in terms of what is known as an "elliptic function." The formula itself can be derived using conservation of energy. In Figure 10.6 we plot the period of oscillation versus the amplitude.

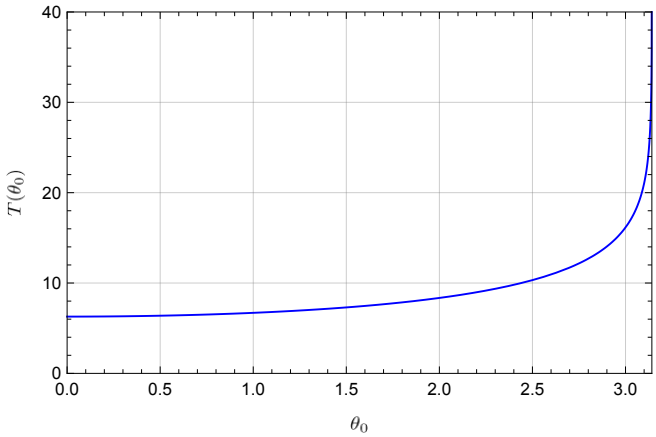


Figure 10.6 Normalised (by $\frac{\ell}{a_g}$) period of a pendulum oscillation as a function of the amplitude

Now let us make use of the preceding observations. We will consider the stability of some nontrivial periodic motion of the pendulum with amplitude between 0 and π. We claim that such a solution is orbitally stable, but not stable. In Figure 10.7 we show a periodic motion of the pendulum as a parameterised curve in the (x_1, x_2)-plane. In the figure we plot three solutions. The middle of the three solutions is the solution ξ_0 whose stability we are referencing. It has initial condition θ_0 and is defined on the time interval $[0, T(\theta_0)]$. The other two solutions have

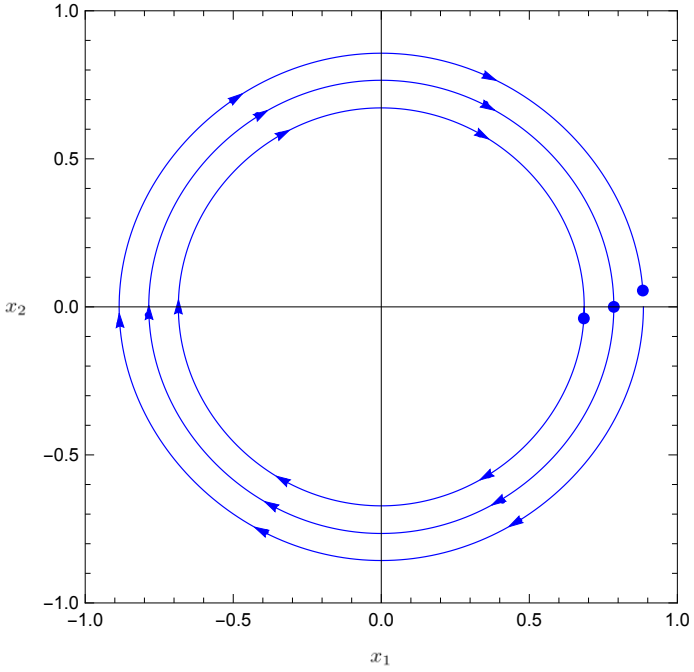


Figure 10.7 Orbital stability, but not stability, of the nontrivial periodic motions of a pendulum; the middle curve is the nominal solution whose stability is being determined

nearby initial conditions, and are defined on the same time interval, and a dot is placed at the final point of the solution. We make the following observations.

1. ξ_0 *is not stable:* The reasoning here is this. In Figure 10.7 we see that the periodic solutions nearby ξ_0 do not undergo exactly one period in the time it takes ξ_0 to undergo exactly one period; the inner solution travels more than one period and the outer solution travels less than one period. Now imagine letting the trajectory ξ_0 undergo an increasing number of periods. The inner and outer solutions will drift further and further from ξ_0 when compared at the same times. This prohibits stability of ξ_0 since nearby initial conditions will produce solutions that are eventually not close.
2. ξ_0 *is orbitally stable:* The reasoning here is this. While solutions with nearby initial conditions will drift apart in time, the solutions themselves remain close in the sense that any point on one solution is nearby some point (not at the same time) on the other solution. More viscerally, the images of solutions for nearby initial solutions are close. •

10.2.19 Example (Stable versus uniformly stable I) Here we take the linear homoge-

neous ordinary differential equation F in $\mathsf{V} = \mathbb{R}$ defined by the right-hand side

$$\widehat{F}(t, x) = -\frac{x}{1+t}$$

for $t \in \mathbb{T} = [0, \infty)$. We will consider the stability of the equilibrium point $x_0 = 0$. We can explicitly solve this ordinary differential equation (for example, using the method of Section 4.1.1) to give

$$\xi(t) = \frac{\xi(t_0)(1+t_0)}{1+t}.$$

From this we can make the following observations.

1. $x_0 = 0$ *is asymptotically stable:* This follows since, for any initial condition $\xi(t_0)$, we have $\lim_{t\to\infty} \xi(t) = 0$.
2. $x_0 = 0$ *is uniformly stable:* Let $\epsilon \in \mathbb{R}_{>0}$ and take $\delta = \epsilon$. If $|\xi(t_0) - 0| < \delta$, then

   $$|\xi(t)| \leq |\xi(t_0)| < \epsilon$$

   for $t \geq t_0$. This gives the desired uniform stability.
3. $x_0 = 0$ *is not uniformly asymptotically stable:* We must show that, for every $\delta \in \mathbb{R}_{>0}$ and $T \in \mathbb{R}_{>0}$, there exists $\epsilon \in \mathbb{R}_{>0}$, $t_0 \in \mathbb{T}$, and $x \in \mathbb{R}$ satisfying $|x - 0| < \delta$, such that the solution $\xi \colon \mathbb{T} \to \mathbb{R}$ to the initial value problem

   $$\dot{\xi}(t) = -\frac{\xi(t)}{1+t}, \quad \xi(t_0) = x,$$

   satisfies $|\xi(t_0 + T)| \geq \epsilon$. We take $x = \frac{\delta}{2}$, $\epsilon = 1$, $T \in \mathbb{R}_{>0}$, and $t_0 \in \mathbb{T}$ such that

   $$\frac{1+t_0}{1+t_0+T} \geq \frac{2}{\delta};$$

   this is possible since $\lim_{t_0\to\infty} \frac{1+t_0}{1+t_0+T} = 1$ for any $T \in \mathbb{R}_{>0}$. Now let $x \in \mathbb{R}$ satisfy $|x - 0| < \delta$, and let $\xi \colon \mathbb{T} \to \mathbb{R}$ be the solution to the initial value problem

   $$\dot{\xi}(t) = -\frac{\xi(t)}{1+t}, \quad \xi(t_0) = x.$$

   Then

   $$|\xi(t_0 + T)| = \left|\frac{x(1+t_0)}{1+t_0+T}\right| = \frac{\delta}{2}\left|\frac{1+t_0}{1+t_0+T}\right| \geq 1,$$

   which gives the desired lack uniform asymptotic convergence. •

10.2.20 Example (Stable versus uniformly stable II) We again consider a linear homogeneous ordinary differential equation in $\mathbb{R}$, this one with right-hand side

$$\widehat{F}(t, x) = \sin(\ln(t)) + \cos(\ln(t)) - \alpha$$

for some $\alpha \in (1, \sqrt{2})$. Here we consider $\mathbb{T} = (0, \infty)$. Again we consider stability of the equilibrium point at $x_0 = 0$. In this case, an application of the method of Section 4.1.1 gives the solution

$$\xi(t) = \mathrm{e}^{-\alpha(t-t_0)+t\sin(\ln(t))-t_0\sin(\ln(t_0))}\xi(t_0).$$

We make the following observations.

1. $x_0 = 0$ *is asymptotically stable:* Here we note that, since

$$\lim_{t\to\infty}(-\alpha(t - t_0) + t\sin(\ln(t)) - t_0\sin(\ln(t_0))) = -\infty$$

since $\alpha > 1$, we must have $\lim_{t\to\infty}\xi(t) = 0$ for any initial condition $\xi(t_0)$. This gives asymptotic stability. In fact, we can refine this conclusion a little.

2. $x_0 = 0$ *is not uniformly stable:* This is more difficult to prove. We choose $\beta \in (\alpha, \sqrt{2})$ and $\theta_1 \in (0, \frac{\pi}{4})$ and $\theta_2 \in (\frac{\pi}{4}, \frac{\pi}{2})$ such that

$$\sin\theta + \cos\theta \geq \beta, \qquad \theta \in [\theta_1, \theta_2].^{4}$$

Then, for $j \in \mathbb{Z}_{>0}$, define

$$t_j = \mathrm{e}^{2j\pi+\theta_2}, \quad t_{0,j} = \mathrm{e}^{2j\pi+\theta_1},$$

and compute, for $j \in \mathbb{Z}_{>0}$,

$$\begin{aligned}\int_{t_{0,j}}^{t_j}(\sin(\ln(t)) + \cos(\ln(t)) - \alpha)\,\mathrm{d}t &= \int_{2j\pi+\theta_1}^{2j\pi+\theta_2}(\sin\theta + \cos\theta - \alpha)\mathrm{e}^{2j\pi+\theta}\,\mathrm{d}\theta\\ &= \int_{\theta_1}^{\theta_2}(\sin\theta + \cos\theta - \alpha)\mathrm{e}^{2j\pi+\theta}\,\mathrm{d}\theta\\ &\geq (\beta - \alpha)\mathrm{e}^{2j\pi}\int_{\theta_1}^{\theta_2}\mathrm{e}^{\theta}\,\mathrm{d}\theta\\ &= (\beta - \alpha)\mathrm{e}^{2j\pi}(\mathrm{e}^{\theta_2} - \mathrm{e}^{\theta_1}),\end{aligned}$$

[4] To see why this is possible, first note that

$$\sqrt{2}\cos(\theta - \tfrac{\pi}{4}) = \sin\theta\sin\tfrac{\pi}{4} + \cos\theta\cos\tfrac{\pi}{4} = \sin\theta + \cos\theta,$$

using standard trigonometric identities. Then note that the function

$$\theta \mapsto \sqrt{2}\cos(\theta - \tfrac{\pi}{4})$$

has a local maximum at $\theta = \frac{\pi}{4}$ with value $\sqrt{2}$. Thus, since $\alpha < \beta < \sqrt{2}$, we can choose $\theta_1 < \frac{\pi}{4}$ and $\theta_2 > \frac{\pi}{4}$ sufficiently close to $\frac{\pi}{4}$ to ensure that $\sin\theta + \cos\theta \geq \beta$ for $\theta \in [\theta_1, \theta_2]$.

where we have used the change of variable $t = \mathrm{e}^{2j\pi+\theta}$ in the second line. Note, then, that

$$\lim_{j\to\infty} \int_{t_{0,j}}^{t_j} (\sin(\ln(t)) + \cos(\ln(t)) - \alpha)\,\mathrm{d}t = \infty.$$

Now, using this fact, we claim that $x_0 = 0$ is not uniformly stable. We must show that there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for every $\delta \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$, $t_0 \in \mathbb{T}$, and $x \in \mathbb{R}$ satisfying $|x - 0| < \delta$ and for which the solution to the initial value problem

$$\dot{\xi}(t) = (\sin(\ln(t)) + \cos(\ln(t)) - \alpha)\xi(t), \quad \xi(t_0) = x,$$

satisfies $|\xi(t_0 + T) - 0| \geq \epsilon$. We take $\epsilon = 1$. Let $\delta \in \mathbb{R}_{>0}$ and $x = \frac{\delta}{2}$. Let $j \in \mathbb{Z}_{>0}$ be sufficiently large that

$$\int_{t_{0,j}}^{t_j} (\sin(\ln(t)) + \cos(\ln(t)) - \alpha)\,\mathrm{d}t \geq \frac{2}{\delta}.$$

Then take $t_0 = t_{0,j}$ and $T = t_j$. We then have

$$|\xi(t_0 + T)| = \left| \int_{t_{0,j}}^{t_j} (\sin(\ln(t)) + \cos(\ln(t)) - \alpha)\,\mathrm{d}t \right| |x| \geq 1,$$

giving the desired absence of uniform stability. •

While the preceding examples do not cover all of the possible gaps in the stability definitions of Definition 10.2.4, they do hopefully sufficiently illustrate the essence of the difference in the various definitions that a reader can have a picture in their mind of these differences as we proceed to study stability in more detail in the sequel.

Exercises

10.2.1 Let us consider the system of ordinary differential equations $\boldsymbol{F}$ with state-space $\mathbb{R}^2$ defined by the right-hand side

$$\begin{aligned} \widehat{F}\colon \mathbb{R} \times \mathbb{R}^2 &\to \mathbb{R}^2 \\ (t, (x_1, x_2)) &\mapsto (1, 0). \end{aligned}$$

Answer the following questions.

(a) Show that

$$\begin{aligned} \xi_0\colon \mathbb{R} &\to \mathbb{R}^2 \\ t &\mapsto (t, 0) \end{aligned}$$

is the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(0) = \boldsymbol{0}.$$

(b) Show that the solution ξ_0 is stable but not asymptotically stable.

Now consider a change of coordinates from $(x_1, x_2) \in \mathbb{R}^2$ to $(y_1, y_2) \in \mathbb{R}^2$ defined by

$$y_1 = x_1, \quad y_2 = \mathrm{e}^{x_1} x_2,$$

and let $\boldsymbol{G}$ be the ordinary differential equation $\boldsymbol{F}$, represented in these coordinates.

(c) Use the Chain Rule to compute $\dot{y}_1$ and $\dot{y}_2$,

$$\dot{y}_1(t) = \frac{\partial y_1}{\partial x_1}\dot{x}_1(t) + \frac{\partial y_1}{\partial x_2}\dot{x}_2(t),$$
$$\dot{y}_2(t) = \frac{\partial y_2}{\partial x_1}\dot{x}_1(t) + \frac{\partial y_2}{\partial x_2}\dot{x}_2(t),$$

and so give the right-hand side $\widehat{G}$ for $\boldsymbol{G}$.

Hint: *Write everything in terms of the coordinates* $(\mathrm{y}_1, \mathrm{y}_2)$.

(d) Show that the solution ξ_0 is mapped, under the change of coordinates, to the solution $\eta_0 \colon \mathbb{R} \to \mathbb{R}^2$ given by $\eta_0(t) = (t, 0)$.

(e) Show that η_0 is not stable.

Now consider a change of coordinates from $(x_1, x_2) \in \mathbb{R}^2$ to $(z_1, z_2) \in \mathbb{R}^2$ defined by

$$z_1 = x_1, \quad z_2 = \mathrm{e}^{-x_1} x_2,$$

and let $\boldsymbol{H}$ be the ordinary differential equation $\boldsymbol{F}$, represented in these coordinates.

(f) Use the Chain Rule to compute $\dot{z}_1$ and $\dot{z}_2$,

$$\dot{z}_1(t) = \frac{\partial z_1}{\partial x_1}\dot{x}_1(t) + \frac{\partial z_1}{\partial x_2}\dot{x}_2(t),$$
$$\dot{z}_2(t) = \frac{\partial z_2}{\partial x_1}\dot{x}_1(t) + \frac{\partial z_2}{\partial x_2}\dot{x}_2(t),$$

and so give the right-hand side $\widehat{H}$ for $\boldsymbol{H}$.

Hint: *Write everything in terms of the coordinates* $(\mathrm{z}_1, \mathrm{z}_2)$.

(g) Show that the solution ξ_0 is mapped, under the change of coordinates, to the solution $\zeta_0 \colon \mathbb{R} \to \mathbb{R}^2$ given by $\zeta_0(t) = (t, 0)$.

(h) Show that ζ_0 is asymptotically stable.

Section 10.3

Stability of linear ordinary differential and difference equations

In this section we devote ourselves specially to the theory of stability for linear systems. We shall see that, for linear systems, there are a few natural places where one can refine the general definitions of stability from Definition 10.2.4, taking advantage of the linearity of the dynamics. Moreover, there are also equivalent characterisations of stability that hold for linear equations that do not hold in general.

As we did in Chapter 5 when dealing with linear systems, we shall work with linear systems whose state space is a finite-dimensional vector space V. Our stability definitions from Definition 10.2.4 all involve the measure of distance provided by the Euclidean norm on $\mathbb{R}^n$. An abstract vector space does not have a natural norm, but one can always be provided by, for example, choosing a basis $\mathscr{B} = \{e_1, \dots, e_n\}$ and then defining $\|v\|_{\mathscr{B}} = \|(v_1, \dots, v_n)\|$, where $v = v_1 e_1 + \cdots + v_n e_n$. The fact of the matter is that nothing we do depends in any way on the choice of this norm,[5] and so we shall simply use the symbol "$\|\cdot\|$" to represent some choice of norm, possibly arising from the Euclidean norm by a choice of basis as described above. For readers following the "all vector spaces are $\mathbb{R}^n$" path, this is not anything of concern so you can resume sleeping.

We proceed in a manner contrary to our approach in Sections 4.2, 4.3, 5.2, and 5.3, and first consider in Section 10.3.1 equations with constant coefficients. The rationale is that, for equations with constant coefficients, there are easily understandable characterisations for all of the various sorts of stability. When we turn in Section 10.3.2 to general equations, the constant coefficient characterisations give us something with which to compare. Much of what can be said for the stability of linear equations with constant coefficients has to do with the roots of the characteristic polynomial of the linear transformation associated to the equation.

10.3.1 Equations with constant coefficients

We shall study the stability of systems of linear homogeneous ordinary differential equations F with constant coefficients in a finite-dimensional $\mathbb{R}$-vector space

[5] The "big fact" here is that if we have two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ for a vector space V, then there exists $C \in \mathbb{R}_{>0}$ such that

$$C\|v\|_2 \leq \|v\|_1 \leq C^{-1}\|v\|_2, \qquad v \in \mathsf{V}.$$

Thus, if a reader goes through our definitions where a norm is used, she will see that using a different norm will only have the effect of change constants in the definition, while not materially altering the meaning of the definition.

V. Such an equation will have right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V}\\ (t, x) &\mapsto \mathsf{A}(x)\end{aligned}$$

for $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$.

First we observe that the general stability definitions of Definition 10.2.7 for linear homogeneous ordinary differential equations collapse.

10.3.1 Proposition (Collapsing of stability definitions for linear homogeneous equations with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *and with right-hand side* $\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}) = \mathsf{A}(\mathrm{x})$ *for* $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$*. Suppose that* $\sup \mathbb{T} = \infty$*. Then* F *is GS (resp. GAS, GES) if and only if it is UGS (resp. UGAS, UGES). Moreover,* F *is GAS if and only if it is GES.*

Proof The first assertion follows from Proposition 10.2.5 and the second follows from Theorem 10.2.10. ■

Now we turn to providing a useful characterisation of stability for linear homogeneous ordinary differential equations with constant coefficients. To do this we first make a definition.

10.3.2 Definition (Spectrum of a linear transformation) Let V be a finite-dimensional $\mathbb{R}$-vector space and let $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$. The ***spectrum*** of A is the set

$$\mathrm{spec}(\mathsf{A}) = \{\lambda \in \mathbb{C} \mid \lambda \text{ is an eigenvalue for } \mathsf{A}_{\mathbb{C}}\}$$

of eigenvalues of the complexification of A. •

Our characterisations of stability will be given in terms of the location of spec(A). It will be convenient to introduce the following notation:

$$\begin{aligned}\mathbb{C}_- &= \{z \in \mathbb{C} \mid \mathrm{Re}(z) < 0\}, \quad \mathbb{C}_+ = \{z \in \mathbb{C} \mid \mathrm{Re}(z) > 0\},\\ \overline{\mathbb{C}}_- &= \{z \in \mathbb{C} \mid \mathrm{Re}(z) \leq 0\}, \quad \overline{\mathbb{C}}_+ = \{z \in \mathbb{C} \mid \mathrm{Re}(z) \geq 0\},\\ \mathrm{i}\mathbb{R} &= \{z \in \mathbb{C} \mid \mathrm{Re}(z) = 0\}.\end{aligned}$$

With this notation, we state the following theorem, which is the main result of this section.

10.3.3 Theorem (Stability of systems of linear homogeneous ordinary differential equations with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional vector space* V *with constant coefficients and with* $\widehat{\mathrm{F}}(\mathrm{t}, \mathrm{x}) = \mathsf{A}(\mathrm{x})$ *for* $\mathsf{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$*. The following statements hold.*

(i) F *unstable if* $\mathrm{spec}(\mathsf{A}) \cap \mathbb{C}_+ \neq \emptyset$*.*

(ii) F *is GAS if* $\mathrm{spec}(\mathsf{A}) \subseteq \mathbb{C}_-$*.*

(iii) F *is GS if* $\operatorname{spec}(\mathsf{A}) \cap \mathbb{C}_+ = \varnothing$ *and if* $\mathrm{m_g}(\lambda, \mathsf{A}) = \mathrm{m_a}(\lambda, \mathsf{A})$ *for* $\lambda \in \operatorname{spec}(\mathsf{A}) \cap (\mathrm{i}\mathbb{R})$.

(iv) F *is unstable if* $\mathrm{m_g}(\lambda, \mathsf{A}) < \mathrm{m_a}(\lambda, \mathsf{A})$ *for* $\lambda \in \operatorname{spec}(\mathsf{A}) \cap (\mathrm{i}\mathbb{R})$.

Proof (i) In this case there is an eigenvalue $\sigma + \mathrm{i}\omega \in \mathbb{C}_+$ and a corresponding eigenvector $u + \mathrm{i}v \in \mathsf{V}_{\mathbb{C}}$ which gives rise to real solutions

$$\xi_1(t) = \mathrm{e}^{\sigma t}(\cos(\omega t)u - \sin(\omega t)v), \quad \xi_2(t) = \mathrm{e}^{\sigma t}(\sin(\omega t)u + \cos(\omega t)v).$$

Clearly these solutions are unbounded as $t \to \infty$ since $\sigma > 0$.

(ii) If all eigenvalues lie in $\mathbb{C}_-$, then any solution of F will be a linear combination of n linearly independent vector functions of the form

$$t^k \mathrm{e}^{-\alpha t}u \quad \text{or} \quad t^k \mathrm{e}^{-\sigma t}(\cos(\omega t)u - \sin(\omega t)v) \quad \text{or} \quad t^k \mathrm{e}^{-\sigma t}(\sin(\omega t)u + \cos(\omega t)v) \tag{10.3}$$

for $\alpha, \sigma > 0$. Note that all such functions tend in length to zero as $t \to \infty$. Suppose that we have a collection $\xi_1, \dots, \xi_n$ of such vector functions. Then, for any solution ξ we have, for some constants $c_1, \dots, c_n$,

$$\begin{aligned} \lim_{t\to\infty} \|\xi(t)\| &= \lim_{t\to\infty} \|c_1\xi_1(t) + \cdots + c_n\xi_n(t)\| \\ &\le |c_1| \lim_{t\to\infty} \|\xi_1(t)\| + \cdots + |c_n| \lim_{t\to\infty} \|\xi_n(t)\| \\ &= 0, \end{aligned}$$

where we have used the triangle inequality, and the fact that the solutions $\xi_1, \dots, \xi_n$ all tend to zero as $t \to \infty$.

(iii) If $\operatorname{spec}(\mathsf{A}) \cap \mathbb{C}_+ = \varnothing$ and if, further, $\operatorname{spec}(\mathsf{A}) \subseteq \mathbb{C}_-$, then we are in case (ii), so F is GAS, and so GS. Thus we need only concern ourselves with the case when we have eigenvalues on the imaginary axis. In this case, provided that all such eigenvalues have equal geometric and algebraic multiplicities, all solutions will be linear combinations of functions like those in (10.3) or functions like

$$\sin(\omega t)u \quad \text{or} \quad \cos(\omega t)u. \tag{10.4}$$

Let $\xi_1, \dots, \xi_\ell$ be ℓ linearly independent functions of the form (10.3), and let $\xi_{\ell+1}, \dots, \xi_n$ be linearly independent functions of the form (10.4), so that $\xi_1, \dots, \xi_n$ forms a set of linearly independent solutions for F. Thus we will have, for some constants $c_1, \dots, c_n$,

$$\begin{aligned} \limsup_{t\to\infty} \|\xi(t)\| &= \limsup_{t\to\infty} \|c_1\xi_1(t) + \cdots + c_n\xi_n(t)\| \\ &\le |c_1| \limsup_{t\to\infty} \|\xi_1(t)\| + \cdots + |c_\ell| \limsup_{t\to\infty} \|\xi_\ell(t)\| + \\ &\quad |c_{\ell+1}| \limsup_{t\to\infty} \|\xi_{\ell+1}(t)\| + \cdots + |c_n| \limsup_{t\to\infty} \|\xi_n(t)\| \\ &= |c_{\ell+1}| \limsup_{t\to\infty} \|\xi_{\ell+1}(t)\| + \cdots + |c_n| \limsup_{t\to\infty} \|\xi_n(t)\|. \end{aligned}$$

Since each of the terms $\|\xi_{\ell+1}(t)\|, \dots, \|\xi_n(t)\|$ are bounded as functions of t, their lim sup's will exist, which is what we wish to show.

(iv) If A has an eigenvalue $\lambda = \mathrm{i}\omega$ on the imaginary axis for which $m_{\mathrm{g}}(\lambda, \mathsf{A}) < \mathit{algmult}(\lambda, \mathsf{A})$, then there will be solutions for F that are linear combinations of vector functions of the form $t^k \sin(\omega t)u$ or $t^k \cos(\omega t)v$. Such functions are unbounded as $t \to \infty$, and so F is unstable. ∎

10.3.4 Remarks (Stability and eigenvalues)

1. A linear mapping A is ***Hurwitz*** if $\operatorname{spec}(\mathsf{A}) \subseteq \mathbb{C}_-$. Thus A is Hurwitz if and only if F is GAS.
2. We see that stability is almost completely determined by the eigenvalues of A. Indeed, one says that F is ***spectrally stable*** if A has no eigenvalues in $\mathbb{C}_+$. It is only in the case where there are repeated eigenvalues on the imaginary axis that one gets to distinguish spectral stability from stability. •

The notion of stability for systems of linear homogeneous ordinary differential equations with constant coefficients is, in principle, an easy one to check, as we see from an example.

10.3.5 Example (Stability of system of linear homogeneous ordinary differential equations with constant coefficients) We look at a system of linear homogeneous ordinary differential equations $\boldsymbol{F}$ in $\mathbb{R}^2$ with constant coefficients, and determined by the 2×2-matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix}.$$

The eigenvalues of $\boldsymbol{A}$ are the roots of the characteristic polynomial $P_A = X^2 + aX + b$, and these are

$$-\tfrac{a}{2} \pm \tfrac{1}{2}\sqrt{a^2 - 4b}.$$

The situation with the eigenvalue placement can be broken into cases.

1. $a = 0$ and $b = 0$: In this case there is a repeated zero eigenvalue. Thus we have spectral stability, but we need to look at eigenvectors to determine stability. One readily verifies that there is only one linearly independent eigenvector for the zero eigenvalue, so the system is unstable.
2. $a = 0$ and $b > 0$: In this case the eigenvalues are purely imaginary. Since the roots are also distinct, they will have equal algebraic and geometric multiplicity. Thus the system is GS, but not GAS.
3. $a = 0$ and $b < 0$: In this case both roots are real, and one will be positive. Thus the system is unstable.
4. $a > 0$ and $b = 0$: There will be one zero eigenvalue if $b = 0$. If $a > 0$ the other root will be real and negative. In this case then, we have a root on the imaginary axis. Since it is distinct, the system will be GS, but not GAS.
5. $a > 0$ and $b > 0$: One may readily ascertain (in Section 10.4 we'll see an easy way to do this) that all eigenvalues are in $\mathbb{C}_-$ if $a > 0$ and $b > 0$. Thus when a and b are strictly positive, the system is GAS.
6. $a > 0$ and $b < 0$: In this case both eigenvalues are real, one being positive and the other negative. Thus the system is unstable.
7. $a < 0$ and $b = 0$: We have one zero eigenvalue. The other, however, will be real and positive, and so the system is unstable.

8. $a < 0$ and $b > 0$: We play a little trick here. If s_0 is a root of $s^2 + as + b$ with $a, b < 0$, then $-s_0$ is clearly also a root of $s^2 - as + b$. From the previous case, we know that $-s_0 \in \mathbb{C}_-$, which means that $s_0 \in \mathbb{C}_+$. So in this case all eigenvalues are in $\mathbb{C}_+$, and so we have instability.
9. $a < 0$ and $b < 0$: In this case we are guaranteed that all eigenvalues are real, and furthermore it is easy to see that one eigenvalue will be positive, and the other negative. Thus the system will be unstable. •

10.3.2 Equations with time-varying coefficients

We work in this section with a system F of linear homogeneous ordinary differential equations with right-hand side

$$\begin{aligned} \widehat{F}\colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V} \\ (t, x) &\mapsto \mathsf{A}(t)(x) \end{aligned} \tag{10.5}$$

for some function $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$.

unstable-nopoles.pdf for an unstable system with no poles

Exercises

10.3.1

In the next exercise we shall make use of a norm $|||\cdot|||$ on the set $\mathrm{L}(\mathsf{V}; \mathsf{V})$ of linear transformations induced by a norm $\|\cdot\|$ on V. The norm is defined by

$$|||L||| = \sup\left\{ \frac{\|L(v)\|}{\|v\|} \;\middle|\; v \in \mathsf{V} \setminus \{0\} \right\},$$

for $L \in \mathrm{L}(\mathsf{V}; \mathsf{V})$. It is easy to show that this is, in fact, a norm and we refer the reader to the references for this.

10.3.2 Let F be a system of linear homogeneous ordinary differential equations in an n-dimensional $\mathbb{R}$-vector space V and with right-hand side $\widehat{F}(t, x) = \mathsf{A}(t)(x)$ for a continuous map $\mathsf{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$. Suppose that $\sup \mathbb{T} = \infty$.

(a) Show that F is stable if and only if, for every $t_0 \in \mathbb{T}$, there exists $C \in \mathbb{R}_{>0}$ such that $|||\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)||| \leq C$ for $t \geq t_0$.
(b) Show that F is asymptotically stable if and only if, for every $t_0 \in \mathbb{T}$ and $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that $|||\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)||| < \epsilon$ for $t \geq t_0 + T$.
(c) Show that F is exponentially stable if and only if, for every $t_0 \in \mathbb{T}$, there exist $M, \sigma \in \mathbb{R}_{>0}$ such that $|||\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)||| \leq M\mathrm{e}^{-\sigma(t-t_0)}$ for $t \geq t_0$.
(d) Show that F is uniformly stable if and only if there exists $C \in \mathbb{R}_{>0}$ such that, for every $t_0 \in \mathbb{T}$, $|||\Phi^{\mathrm{c}}_{\mathsf{A}}(t, t_0)||| \leq C$ for $t \geq t_0$.
(e) Show that F is uniformly asymptotically stable if and only if,

1. there exists $C \in \mathbb{R}_{>0}$ such that, for every $t_0 \in \mathbb{T}$, $|||\Phi^c_A(t, t_0)||| \leq C$ for $t \geq t_0$ and
2. for every $\epsilon \in \mathbb{R}_{>0}$, there exists $T \in \mathbb{R}_{>0}$ such that, for every $t_0 \in \mathbb{T}$, $|||\Phi^c_A(t, t_0)||| < \epsilon$ for $t \geq t_0 + T$.

(f) Show that F is exponentially stable if and only if there exist $M, \sigma \in \mathbb{R}_{>0}$ such that, for every $t_0 \in \mathbb{T}$, $|||\Phi^c_A(t, t_0)||| \leq Me^{-\sigma(t-t_0)}$ for $t \geq t_0$.

Section 10.4

Hurwitz polynomials

From Theorem 10.3.3 we see that it is important to be able to determine when the roots of a polynomial lie in the negative half-plane. However, checking that such a condition holds may not be so easy; one should regard the problem of computing the roots of a polynomial as being impossible for polynomials of degree 5 or more, and annoyingly complicated for polynomials of degree 3 or 4. However, one may establish conditions on the coefficients of a polynomial. In this section, we present three methods for doing exactly this. We also look at a test for the roots to lie in $\mathbb{C}_-$ when we only approximately know the coefficients of the polynomial. We shall generally say that a polynomial all of whose roots lie in $\mathbb{C}_-$ is ***Hurwitz***.

10.4.1 The Routh criterion

For the method of Routh, we construct an array involving the coefficients of the polynomial in question. The array is constructed inductively, starting with the first two rows. Thus suppose one has two collections $a_{11}, a_{12}, \dots$ and $a_{21}, a_{22}, \dots$ of numbers. In practice, this is a finite collection, but let us suppose the length of each collection to be indeterminate for convenience. Now construct a third row of numbers $a_{31}, a_{32}, \dots$ by defining $a_{3k} = a_{21}a_{1,k+1} - a_{11}a_{2,k+1}$. Thus a_{3k} is minus the determinant of the matrix $\begin{bmatrix} a_{11} & a_{1,k+1} \\ a_{21} & a_{2,k+1} \end{bmatrix}$. In practice, one writes this down as follows:

$$\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1k} & \cdots \\ a_{21} & a_{22} & \cdots & a_{2k} & \cdots \\ a_{21}a_{12} - a_{11}a_{22} & a_{21}a_{13} - a_{11}a_{23} & \cdots & a_{21}a_{1,k+1} - a_{11}a_{2,k+1} & \cdots \end{array}$$

One may now proceed in this way, using the second and third row to construct a fourth row, the third and fourth row to construct a fifth row, and so on. To see how to apply this to a given polynomial $P \in \mathbb{R}[X]$. Define two polynomials $P_+, P_- \in \mathbb{R}[X]$ as the even and odd part of P. To be clear about this, if

$$P = p_0 + p_1X + p_2X^2 + p_3X^3 + \cdots + p_{n-1}X^{n-1} + p_nX^n,$$

then

$$P_+ = p_0 + p_2X + p_4X^2 + \dots, \quad P_- = p_1 + p_3X + p_5X^2 + \dots.$$

Note then that $P(X) = P_+(X^2) + XP_-(X^2)$. Let $R(P)$ be the array constructed as above, with the first two rows being comprised of the coefficients of P_+ and P_-, respectively, starting with the coefficients of lowest powers of X, and increasing to

higher powers of X. Thus the first three rows of $R(P)$ are

$$\begin{array}{rll} p_0 & p_2 & \cdots \\ p_1 & p_3 & \cdots \\ p_1p_2 - p_0p_3 & p_1p_4 - p_0p_5 & \cdots \\ \vdots & \vdots & \vdots \end{array} \qquad \begin{array}{rl} p_{2k} & \cdots \\ p_{2k+1} & \cdots \\ p_1p_{2k+2} - p_0p_{2k+3} & \cdots \\ \vdots & \vdots \end{array}$$

In making this construction, a zero is inserted whenever an operation is undefined. It is readily determined that the first column of $R(P)$ has at most $n+1$ nonzero components. The ***Routh array*** is then the first column of the first $n+1$ rows.

With this as setup, we may now state a criterion for determining whether a polynomial is Hurwitz.

10.4.1 Theorem (Routh's criterion) *A polynomial*

$$\mathrm{P} = \mathrm{X}^{\mathrm{n}} + \mathrm{p}_{\mathrm{n}-1}\mathrm{X}^{\mathrm{n}-1} + \cdots + \mathrm{p}_1\mathrm{X} + \mathrm{p}_0 \in \mathbb{R}[\mathrm{X}]$$

is Hurwitz if and only if all elements of the Routh array corresponding to R(P) *are positive.*

Proof Let us construct a sequence of polynomials as follows. We let $P_0 = P_+$ and $P_1 = P_-$ and let

$$P_2(X) = X^{-1}\Big(P_1(0)P_0(X) - P_0(0)P_1(X)\Big).$$

Note that the constant coefficient of $P_1(0)P_0(X) - P_0(0)P_1(X)$ is zero, so this does indeed define P_2 as a polynomial. Now inductively define

$$P_k(X) = X^{-1}\Big(P_{k-1}(0)P_{k-2}(X) - P_{k-2}(0)P_{k-1}(X)\Big)$$

for $k \geq 3$. With this notation, we have the following lemma that describes the statement of the theorem.

1 Lemma *The* (k + 1)*st row of* R(P) *consists of the coefficients of* $\mathrm{P_k}$ *with the constant coefficient in the first column. Thus the hypothesis of the theorem is equivalent to the condition that* $\mathrm{P}_0(0), \mathrm{P}_1(0), \ldots, \mathrm{P}_{\mathrm{n}}(0)$ *all be positive.*

Proof We have $P_0(0) = p_0$, $P_1(0) = p_1$, and $P_2(0) = p_1p_2 - p_0p_3$, directly from the definitions. Thus the lemma holds for $k \in \{0, 1, 2\}$. Now suppose that the lemma holds for $k \geq 3$. Thus the kth and the $(k+1)$st rows of $R(P)$ are the coefficients of the polynomials

$$P_{k-1}(X) = p_{k-1,0} + p_{k-1,1}X + \cdots$$

and

$$P_k(X) = p_{k,0} + p_{k,1}X + \cdots,$$

respectively. Using the definition of P_{k+1} we see that $P_{k+1}(0) = p_{k,0}p_{k-1,1} - p_{k-1,0}p_{k,1}$. However, this is exactly the term as it would appear in first column of the $(k+2)$nd row of $R(P)$. ▼

Now note that $P(X) = P_0(X^2) + XP_1(X^2)$ and define $Q \in \mathbb{R}[X]$ by $Q(X) = P_1(X^2) + XP_2(X^2)$. One may readily verify that $\deg(Q) \leq n-1$. Indeed, in the proof of Theorem 10.4.3, a formula for Q will be given. The following lemma is key to the proof. Let us suppose for the moment that p_n is not equal to 1.

2 Lemma *The following statements are equivalent:*

(i) P *is Hurwitz and* $\mathrm{p_n} > 0$*;*

(ii) Q *is Hurwitz,* $\mathrm{q_{n-1}} > 0$*, and* $\mathrm{P}(0) > 0$*.*

Proof We have already noted that $P(X) = P_0(X^2) + XP_1(X^2)$. We may also compute

$$Q(X) = P_1(X^2) + X^{-1}\Big(P_1(0)P_0(X^2) - P_0(0)P_1(X^2)\Big). \tag{10.6}$$

For $\lambda \in [0,1]$ define $Q_\lambda(X) = (1-\lambda)P(X) + \lambda Q(X)$, and compute

$$Q_\lambda(X) = \Big((1-\lambda) + X^{-1}\lambda P_1(0)\Big)P_0(X^2) + \Big((1-\lambda)X + \lambda - X^{-1}\lambda P_0(0)\Big)P_1(X^2).$$

The polynomials $P_0(X^2)$ and $P_1(X^2)$ are even, so that when evaluated on the imaginary axis they are real. Now we claim that the roots of Q_λ that lie on the imaginary axis are independent of λ, provided that $P(0) > 0$ and $Q(0) > 0$. First note that, if $P(0) > 0$ and $Q(0) > 0$, then 0 is not a root of Q_λ. Now, if $\mathrm{i}\omega_0$ is a nonzero imaginary root, then we must have

$$\Big((1-\lambda) - \mathrm{i}\omega_0^{-1}\lambda P_1(0)\Big)P_0(-\omega_0^2) + \Big((1-\lambda)i\omega_0 + \lambda + \mathrm{i}\omega_0^{-1}\lambda P_0(0)\Big)P_1(-\omega_0^2) = 0.$$

Balancing real and imaginary parts of this equation gives

$$\begin{aligned} &(1-\lambda)P_0(-\omega_0^2) + \lambda P_1(-\omega_0^2) = 0 \\ &\lambda\omega_0^{-1}\Big(P_0(0)P_1(-\omega_0^2) - P_1(0)P_0(-\omega_0^2)\Big) + \omega_0(1-\lambda)P_1(-\omega_0^2). \end{aligned} \tag{10.7}$$

If we think of this as a homogeneous linear equation in $P_0(-\omega_0^2)$ and $P_1(\omega_0^2)$, one determines that the determinant of the coefficient matrix is

$$\omega_0^{-1}\Big((1-\lambda)^2\omega_0^2 + \lambda((1-\lambda)P_0(0) + \lambda P_1(0))\Big).$$

This expression is positive for $\lambda \in [0,1]$ since $P(0), Q(0) > 0$ implies that $P_0(0), P_1(0) > 0$. To summarise, we have shown that, provided that $P(0) > 0$ and $Q(0) > 0$, all imaginary axis roots $\mathrm{i}\omega_0$ of Q_λ satisfy $P_0(-\omega_0^2) = 0$ and $P_1(-\omega_0^2) = 0$. In particular, the imaginary axis roots of Q_λ are independent of $\lambda \in [0,1]$ in this case.

(i) $\Longrightarrow$ (ii) For $\lambda \in [0,1]$ let

$$N(\lambda) = \begin{cases} n, & \lambda \in [0,1) \\ n-1, & \lambda = 1. \end{cases}$$

Thus $N(\lambda)$ is the number of roots of Q_λ. Now let

$$Z_\lambda = \{z_{\lambda,i} \mid \ i \in \{1, \dots, N(\lambda)\}\}$$

be the set of roots of Q_λ. Since P is Hurwitz, $Z_0 \subseteq \mathbb{C}_-$. Our previous computations then show that $Z_\lambda \cap \mathrm{i}\mathbb{R} = \varnothing$ for $\lambda \in [0,1]$. Now, if $Q = Q_1$ were to have a root in $\overline{\mathbb{C}}_+$, this would mean that, for some value of λ, one of the roots of Q_λ would have to lie on the imaginary axis, using the (nontrivial) fact that the roots of a polynomial are continuous functions of its coefficients. This then shows that all roots of Q must lie in

$\mathbb{C}_-$. That $P(0) > 0$ is a consequence of Exercise 10.4.1 and P being Hurwitz. One may check that $q_{n-1} = p_1 \cdots p_n$, so that $q_{n-1} > 0$ follows from Exercise 10.4.1 and $p_n > 0$.

(ii) $\implies$ (i) Let us adopt the notation $N(\lambda)$ and Z_λ from the previous part of the proof. Since Q is Hurwitz, $Z_1 \subseteq \mathbb{C}_-$. Furthermore, since $Z_\lambda \cap i\mathbb{R} = \emptyset$, it follows that, for $\lambda \in [0,1]$, the number of roots of Q_λ within $\mathbb{C}_-$ must equal $n-1$ as $\deg(Q) = n-1$. In particular, P can have at most one root in $\mathbb{C}_+$. This root, then, must be real, and let us denote it by $z_0 > 0$. Thus $P(X) = \tilde{P}(X)(X - z_0)$ where $\tilde{P}$ is Hurwitz. By Exercise 10.4.1 it follows that all coefficients of $\tilde{P}$ are positive. If we write

$$\tilde{P} = \tilde{p}_{n-1}X^{n-1} + \tilde{p}_{n-2}X^{n-2} + \cdots + \tilde{p}_1 X + \tilde{p}_0,$$

then

$$P(X) = \tilde{p}_{n-1}X^n + (\tilde{p}_{n-2} - z_0\tilde{p}_{n-1})X^{n-1} + \cdots + (\tilde{p}_0 - z_0\tilde{p}_1)X - \tilde{p}_0 z_0.$$

Thus the existence of a root $z_0 \in \mathbb{C}_+$ contradicts the fact that $P(0) > 0$. Note that we have also shown that $p_n > 0$. ▼

Now we proceed with the proof proper. First suppose that P is Hurwitz. By successive applications of Lemma 2, it follows that the polynomials

$$Q_k(X) = P_k(X^2) + XP_{k+1}(X^2), \quad k \in \{1, \ldots, n\},$$

are Hurwitz and that $\deg(Q_k) = n - k$, $k \in \{1, \ldots, n\}$. What's more, the coefficient of X^{n-k} is positive in Q_k. Now, by Exercise 10.4.1, we have $P_0(0) > 0$ and $P_1(0) > 0$. Now suppose that $P_0(0), P_1(0), \ldots, P_k(0)$ are all positive. Since Q_k is Hurwitz with the coefficient of the highest power of X being positive, from Exercise 10.4.1 it follows that the coefficient of X in Q_k should be positive. However, this coefficient is exactly $P_{k+1}(0)$. Thus we have shown that $P_k(0) > 0$ for $k = 0, 1, \ldots, n$. From Lemma 1 it follows that the elements of the Routh array are positive.

Now suppose that one element of the Routh array is nonpositive and that P is Hurwitz. By Lemma 2, we may suppose that $P_{k_0}(0) \leq 0$ for some $k_0 \in \{2, 3, \ldots, n\}$. Furthermore, since P is Hurwitz, as above the polynomials Q_k, $k \in \{1, \ldots, n\}$, must also be Hurwitz, with $\deg(Q_k) = n - k$ where the coefficient of X^{n-k} in Q_k is positive. In particular, by Exercise 10.4.1, all coefficients of Q_{k_0-1} are positive. However, since $Q_{k_0-1}(X) = P_{k_0-1}(X^2) + XP_{k_0}(X^2)$ it follows that the coefficient of X in Q_{k_0-1} is negative, and hence we arrive at a contradiction, and the theorem follows. ■

The Routh criterion is simple to apply, and we illustrate it in the simple case of a degree two polynomial.

10.4.2 Example (The Routh criterion) Let us apply the criteria to the simplest nontrivial example possible: $P = X^2 + aX + b$. We compute the Routh table to be

$$R(P) = \begin{matrix} b & 1 \\ a & 0. \\ a & 0 \end{matrix}$$

Thus the Routh array is $\begin{bmatrix} b & a & a \end{bmatrix}$, and its entries are all positive if and only if $a, b > 0$. Let us see how this compares to what we know doing the calculations "by hand."

The roots of P are $r_1 = -\frac{a}{2} + \frac{1}{2}\sqrt{a^2 - 4b}$ and $r_2 = -\frac{a}{2} - \frac{1}{2}\sqrt{a^2 - 4b}$. Let us consider the various cases.

1. If $a^2 - 4b < 0$, then the roots are complex with nonzero imaginary part, and with real part $-a$. Thus the roots in this case lie in the negative half-plane if and only if $a > 0$. We also have $b > \frac{a^2}{4}$ and so $b > 0$, and hence $ab > 0$ as in the Routh criterion.
2. If $a^2 - 4b = 0$, then the roots are both $-a$, and so lie in the negative half-plane if and only if $a > 0$. In this case $b = \frac{a^2}{4}$ and so $b > 0$. Thus $ab > 0$ as predicted.
3. Finally we have the case when $a^2 - 4b > 0$. We have two subcases.
   (a) When $a > 0$, then we have negative half-plane roots if and only if $a^2 - 4b < a^2$ which means that $b > 0$. Therefore, we have negative half-plane roots if and only $a > 0$ and $ab > 0$.
   (b) When $a < 0$, then we will never have all negative half-plane roots since $-a + \sqrt{a^2 - 4b}$ is always positive.

So we see that the Routh criterion provides a very simple encapsulation of the necessary and sufficient conditions for all roots to lie in the negative half-plane, even for this simple example. •

10.4.2 The Hurwitz criterion

We consider in this section another test for a polynomial to be Hurwitz. The key ingredient in the Hurwitz construction we consider is a matrix formed from the coefficients of a polynomial

$$P = X^n + p_{n-1}X^{n-1} + \cdots + p_1 X + p_0 \in \mathbb{R}[X].$$

We denote the ***Hurwitz matrix*** by $\boldsymbol{H}(P) \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and define it by

$$\boldsymbol{H}(P) = \begin{bmatrix} p_{n-1} & 1 & 0 & 0 & \cdots & 0 \\ p_{n-3} & p_{n-2} & p_{n-1} & 1 & \cdots & 0 \\ p_{n-5} & p_{n-4} & p_{n-3} & p_{n-2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & p_0 \end{bmatrix}.$$

Any terms in this matrix that are not defined are taken to be zero. Of course, we also take $p_n = 1$. Now define $\boldsymbol{H}(P)_k \in \mathrm{L}(\mathbb{R}^k; \mathbb{R}^k)$, $k \in \{1, \ldots, n\}$, to be the matrix of elements $\boldsymbol{H}(P)_{ij}$, $i, j \in \{1, \ldots, k\}$. Thus $\boldsymbol{H}(P)_k$ is the matrix formed by taking the "upper left $k \times k$ block from $\boldsymbol{H}(P)$." Also define $\Delta_k = \det \boldsymbol{H}(P)_k$.

With this notation, the Hurwitz criterion is as follows.

10.4.3 Theorem (Hurwitz's criterion) *A polynomial*

$$\mathrm{P} = \mathrm{X}^{\mathrm{n}} + \mathrm{p}_{\mathrm{n}-1}\mathrm{X}^{\mathrm{n}-1} + \cdots + \mathrm{p}_1\mathrm{X} + \mathrm{p}_0 \in \mathbb{R}[\mathrm{X}]$$

is Hurwitz if and only if the n ***Hurwitz determinants*** $\Delta_1, \ldots, \Delta_{\mathrm{n}}$ *are positive.*

Proof Let us begin by resuming with the notation from the proof of Theorem 10.4.1. In particular, we recall the definition of $Q(X) = P_1(X^2) + XP_2(X^2)$. We wish to compute $\boldsymbol{H}(Q)$, so we need to compute Q in terms of the coefficients of P. A computation using the definition of Q and P_2 gives

$$Q(X) = p_1 + (p_1p_2 - p_0p_3)X + p_3X^2 + (p_1p_4 - p_0p_5)X^3 + \cdots.$$

One can then see that, when n is even, we have

$$\boldsymbol{H}(Q) = \begin{bmatrix} p_{n-1} & p_1p_n & 0 & 0 & \cdots & 0 & 0 \\ p_{n-3} & p_1p_{n-2} - p_0p_{n-1} & p_{n-1} & p_1p_n & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & p_1p_2 - p_0p_3 & p_3 \\ 0 & 0 & 0 & 0 & \cdots & 0 & p_1 \end{bmatrix}$$

and, when n is odd, we have

$$\boldsymbol{H}(Q) = \begin{bmatrix} p_1p_{n-1} - p_0p_n & p_n & 0 & 0 & \cdots & 0 & 0 \\ p_1p_{n-3} - p_0p_{n-2} & p_{n-2} & p_1p_{n-1} - p_0p_n & p_n & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & p_1p_2 - p_0p_3 & p_3 \\ 0 & 0 & 0 & 0 & \cdots & 0 & p_1 \end{bmatrix}.$$

Now define $\boldsymbol{T} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ by

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & p_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -p_0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -p_0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}$$

when n is even and by

$$\boldsymbol{T} = \begin{bmatrix} p_1 & 0 & \cdots & 0 & 0 & 0 \\ -p_0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & p_1 & 0 & 0 \\ 0 & 0 & \cdots & -p_0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}$$

when n is odd. One then verifies by direct calculation that

$$\boldsymbol{H}(P)\boldsymbol{T} = \begin{bmatrix} & & & \vdots \\ & \boldsymbol{H}(Q) & & p_4 \\ & & & p_2 \\ 0 & \cdots & 0 & p_0 \end{bmatrix}. \tag{10.8}$$

We now let $\Delta_1, \dots, \Delta_n$ be the determinants defined above and let $\tilde{\Delta}_1, \dots, \tilde{\Delta}_{n-1}$ be the similar determinants corresponding to $\boldsymbol{H}(Q)$. A straightforward computation using (10.8) gives the following relationships between the Δ's and the $\tilde{\Delta}$'s:

$$\begin{aligned} \Delta_1 &= p_1 \\ \Delta_{k+1} &= \begin{cases} p_1^{-\lfloor \frac{k}{2} \rfloor} \tilde{\Delta}_k, & k \text{ even} \\ p_1^{-\lceil \frac{k}{2} \rceil} \tilde{\Delta}_k, & k \text{ odd} \end{cases}, \qquad k = 1, \dots, n-1, \end{aligned} \tag{10.9}$$

where $\lfloor x \rfloor$ gives the greatest integer less than or equal to x and $\lceil x \rceil$ gives the smallest integer greater than or equal to x.

With this background notation, let us proceed with the proof, first supposing that P is Hurwitz. In this case, by Exercise 10.4.1, it follows that $p_1 > 0$ so that $\Delta_1 > 0$. By Lemma 2 of Theorem 10.4.1, it also follows that Q is Hurwitz. Thus $\tilde{\Delta}_1 > 0$. A trivial induction argument on $n = \deg(P)$ then shows that $\Delta_2, \dots, \Delta_n > 0$.

Now suppose that one of $\Delta_1, \dots, \Delta_n$ is nonpositive and that P is Hurwitz. Since Q is then Hurwitz by Lemma 2 of Theorem 10.4.1, we readily arrive at a contradiction, and this completes the proof. ■

The Hurwitz criterion is simple to apply, and we illustrate it in the simple case of a degree two polynomial.

10.4.4 Example (The Hurwitz criterion) Let us apply the criteria to our simple example of $P = X^2 + aX + b$. We then have

$$\boldsymbol{H}(P) = \begin{bmatrix} a & 1 \\ 0 & b \end{bmatrix}$$

We then compute $\Delta_1 = a$ and $\Delta_2 = ab$. Thus $\Delta_1, \Delta_2 > 0$ if and only if $a, b > 0$. This agrees with our application of the Routh method to the same polynomial in Example 10.4.2. •

10.4.3 The Hermite criterion

We next look at a manner of determining whether a polynomial is Hurwitz which makes contact with the Lyapunov methods of Section 10.7.4. Let us consider, as usual, a monic polynomial of degree n:

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0.$$

Corresponding to such a polynomial, we construct its ***Hermite matrix*** as the $n \times n$ matrix $\boldsymbol{P}(P)$ given by

$$\boldsymbol{P}(P)_{ij} = \begin{cases} \sum_{k=1}^{i} (-1)^{k+i} p_{n-k+1} p_{n-i-j+k}, & j \geq i,\ i+j \text{ even} \\ \boldsymbol{P}(P)_{ji}, & j < i,\ i+j \text{ even} \\ 0, & i+j \text{ odd}. \end{cases}$$

As usual, in this formula we take $p_i = 0$ for $i < 0$. One can get an idea of how this matrix is formed by looking at its appearance for small values of n. For $n = 2$ we have

$$\boldsymbol{P}(P) = \begin{bmatrix} p_1p_2 & 0 \\ 0 & p_0p_1 \end{bmatrix},$$

for $n = 3$ we have

$$\boldsymbol{P}(P) = \begin{bmatrix} p_2p_3 & 0 & p_0p_3 \\ 0 & p_1p_2 - p_0p_3 & 0 \\ p_0p_3 & 0 & p_0p_1 \end{bmatrix},$$

and for $n = 4$ we have

$$\boldsymbol{P}(P) = \begin{bmatrix} p_3p_4 & 0 & p_1p_4 & 0 \\ 0 & p_2p_3 - p_1p_4 & 0 & p_0p_3 \\ p_1p_4 & 0 & p_1p_2 - p_0p_3 & 0 \\ 0 & p_0p_3 & 0 & p_0p_1 \end{bmatrix}.$$

The following theorem gives necessary and sufficient conditions for P to be Hurwitz based on its Hermite matrix.

10.4.5 Theorem (Hermite's criterion) *A polynomial*

$$\mathrm{P(s)} = \mathrm{s^n} + \mathrm{p_{n-1}s^{n-1}} + \cdots + \mathrm{p_1 s} + \mathrm{p_0} \in \mathbb{R}[\mathrm{s}]$$

is Hurwitz if and only if $\mathbf{P}(\mathrm{P})$ *is positive-definite.*

Proof Let

$$\boldsymbol{A}(P) = \begin{bmatrix} -p_{n-1} & -p_{n-2} & \cdots & -p_1 & -p_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \boldsymbol{b}(P) = \begin{bmatrix} p_{n-1} \\ 0 \\ p_{n-3} \\ 0 \\ \vdots \end{bmatrix}.$$

An unenjoyable computation gives

$$\boldsymbol{P}(P)\boldsymbol{A}(P) + \boldsymbol{A}(P)^T\boldsymbol{P}(P) = -\boldsymbol{b}(P)\boldsymbol{b}(P)^T.$$

First suppose that $\boldsymbol{P}(P)$ is positive-definite. By Theorem 10.7.9(i), since $\boldsymbol{b}(P)\boldsymbol{b}(P)^T$ is positive-semidefinite, $\boldsymbol{A}(P)$ is Hurwitz. Conversely, if $\boldsymbol{A}(P)$ is Hurwitz, then there is only one symmetric $\boldsymbol{P}$ so that

$$\boldsymbol{P}\boldsymbol{A}(P) + \boldsymbol{A}(P)^T\boldsymbol{P} = -\boldsymbol{b}(P)\boldsymbol{b}(P)^T,$$

this by Theorem 10.10.6(i). Since $\boldsymbol{P}(P)$ satisfies this relation even when $\boldsymbol{A}(P)$ is not Hurwitz, it follows that $\boldsymbol{P}(P)$ is positive-definite. The theorem now follows since the characteristic polynomial of $\boldsymbol{A}(P)$ is P. ■

Let us apply this theorem to our favourite example.

10.4.6 Example (Hermite's criterion) We consider the polynomial $P(s) = s^2 + as + b$ which has the Hermite matrix

$$\boldsymbol{P}(P) = \begin{bmatrix} a & 0 \\ 0 & ab \end{bmatrix}.$$

Since this matrix is diagonal, it is positive-definite if and only if the diagonal entries are zero. Thus we recover the by now well established condition that $a, b > 0$. •

The Hermite criterion, Theorem 10.4.5, does indeed record necessary and sufficient conditions for a polynomial to be Hurwitz. However, it is more computationally demanding than it needs to be, especially for large polynomials. Part of the problem is that the Hermite matrix contains so many zero entries. To get conditions involving smaller matrices leads to the so-called ***reduced Hermite criterion*** which we now discuss. Given a degree n polynomial P with its Hermite matrix $\boldsymbol{P}(P)$, we define ***reduced Hermite matrices*** $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$ as follows:

1. $\boldsymbol{C}(P)$ is obtained by removing the even numbered rows and columns of $\boldsymbol{P}(P)$ and
2. $\boldsymbol{D}(P)$ is obtained by removing the odd numbered rows and columns of $\boldsymbol{P}(P)$.

Thus, if n is even, $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$ are $\frac{n}{2} \times \frac{n}{2}$, and if n is odd, $\boldsymbol{C}(P)$ is $\frac{n+1}{2} \times \frac{n+1}{2}$ and $\boldsymbol{D}(P)$ is $\frac{n-1}{2} \times \frac{n-1}{2}$. Let us record a few of these matrices for small values of n. For $n = 2$ we have

$$\boldsymbol{C}(P) = \begin{bmatrix} p_1 p_2 \end{bmatrix}, \quad \boldsymbol{D}(P) = \begin{bmatrix} p_0 p_1 \end{bmatrix},$$

for $n = 3$ we have

$$\boldsymbol{C}(P) = \begin{bmatrix} p_2 p_3 & p_0 p_3 \\ p_0 p_3 & p_0 p_1 \end{bmatrix}, \quad \boldsymbol{D}(P) = \begin{bmatrix} p_1 p_2 - p_0 p_3 \end{bmatrix},$$

and for $n = 4$ we have

$$\boldsymbol{C}(P) = \begin{bmatrix} p_3 p_4 & p_1 p_4 \\ p_1 p_4 & p_1 p_2 - p_0 p_3 \end{bmatrix}, \quad \boldsymbol{D}(P) = \begin{bmatrix} p_2 p_3 - p_1 p_4 & p_0 p_3 \\ p_0 p_3 & p_0 p_1 \end{bmatrix}.$$

Let us record a useful property of the matrices $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$, noting that they are symmetric.

10.4.7 Lemma (A property of reduced Hermite matrices) P(P) *is positive-definite if and only if both* C(P) *and* D(P) *are positive-definite.*

Proof For $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, denote $\boldsymbol{x}_{\text{odd}} = (x_1, x_3, \dots)$ and $\boldsymbol{x}_{\text{even}} = (x_2, x_4, \dots)$. A simple computation then gives

$$\boldsymbol{x}^T \boldsymbol{P}(P) \boldsymbol{x} = \boldsymbol{x}_{\text{odd}}^T \boldsymbol{C}(P) \boldsymbol{x}_{\text{odd}} + \boldsymbol{x}_{\text{even}}^T \boldsymbol{D}(P) \boldsymbol{x}_{\text{even}}. \tag{10.10}$$

Clearly, if $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$ are both positive-definite, then so too is $\boldsymbol{P}(P)$. Conversely, suppose that one of $\boldsymbol{C}(P)$ or $\boldsymbol{D}(P)$, say $\boldsymbol{C}(P)$, is not positive-definite. Thus there exists $\boldsymbol{x} \in \mathbb{R}^n$ so that $\boldsymbol{x}_{\text{odd}} \neq \boldsymbol{0}$ and $\boldsymbol{x}_{\text{even}} = \boldsymbol{0}$, and for which

$$\boldsymbol{x}_{\text{odd}}^T \boldsymbol{C}(P) \boldsymbol{x}_{\text{odd}} \leq 0.$$

From (10.10), it now follows that $\boldsymbol{P}(P)$ is not positive-definite. ■

The Hermite criterion then tells us that P is Hurwitz if and only if both $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$ are positive-definite. The remarkable fact is that we need only check one of these matrices for definiteness, and this is recorded in the following theorem.

10.4.8 Theorem (Reduced Hermite criterion) *A polynomial*

$$\mathrm{P} = \mathrm{X}^{\mathrm{n}} + \mathrm{p}_{\mathrm{n}-1}\mathrm{X}^{\mathrm{n}-1} + \cdots + \mathrm{p}_1\mathrm{X} + \mathrm{p}_0 \in \mathbb{R}[\mathrm{X}]$$

is Hurwitz if and only if any one of the following conditions holds:

(i) $\mathrm{p}_{2\mathrm{k}} > 0$, $\mathrm{k} \in \{0, 1, \ldots, \lfloor \frac{\mathrm{n}-1}{2} \rfloor\}$ *and* $\mathbf{C}(\mathrm{P})$ *is positive-definite;*

(ii) $\mathrm{p}_{2\mathrm{k}} > 0$, $\mathrm{k} \in \{0, 1, \ldots, \lfloor \frac{\mathrm{n}-1}{2} \rfloor\}$ *and* $\mathbf{D}(\mathrm{P})$ *is positive-definite;*

(iii) $\mathrm{p}_0 > 0$, $\mathrm{p}_{2\mathrm{k}+1} > 0$, $\mathrm{k} \in \{0, 1, \ldots, \lfloor \frac{\mathrm{n}-2}{2} \rfloor\}$ *and* $\mathbf{C}(\mathrm{P})$ *is positive-definite;*

(iv) $\mathrm{p}_0 > 0$, $\mathrm{p}_{2\mathrm{k}+1} > 0$, $\mathrm{k} \in \{0, 1, \ldots, \lfloor \frac{\mathrm{n}-2}{2} \rfloor\}$ *and* $\mathbf{D}(\mathrm{P})$ *is positive-definite.*

Proof First suppose that P is Hurwitz. Then all coefficients are positive (see Exercise 10.4.1) and $\boldsymbol{P}(P)$ is positive-definite by Theorem 10.4.5. This implies that $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$ are positive-definite by Lemma 10.4.7, and thus conditions (i)–(iv) hold. For the converse assertion, the cases when n is even or odd are best treated separately. This gives eight cases to look at. As certain of them are quite similar in flavour, we only give details the first time an argument is encountered.

Case 1: We assume (i) and that n is even. Denote

$$\boldsymbol{A}_1(P) = \begin{bmatrix} -\frac{p_{n-2}}{p_n} & -\frac{p_{n-4}}{p_n} & \cdots & -\frac{p_2}{p_n} & -\frac{p_0}{p_n} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

A calculation then gives $\boldsymbol{C}(P)\boldsymbol{A}_1(P) = -\boldsymbol{D}(P)$. Since $\boldsymbol{C}(P)$ is positive-definite, there exists an orthogonal matrix $\boldsymbol{R}$ so that $\boldsymbol{R}\boldsymbol{C}(P)\boldsymbol{R}^T = \boldsymbol{\Delta}$, where $\boldsymbol{\Delta}$ is diagonal with strictly positive diagonal entries. Let $\boldsymbol{\Delta}^{1/2}$ denote the diagonal matrix whose diagonal entries are the square roots of those of $\boldsymbol{\Delta}$. Now denote $\boldsymbol{C}(P)^{1/2} = \boldsymbol{R}^T\boldsymbol{\Delta}^{1/2}\boldsymbol{R}$, noting that $\boldsymbol{C}(P)^{1/2}\boldsymbol{C}(P)^{1/2} = \boldsymbol{C}(P)$. Also note that $\boldsymbol{C}(P)^{1/2}$ is invertible, and we shall denote its inverse by $\boldsymbol{C}(P)^{-1/2}$. Note that this inverse is also positive-definite. This then gives

$$\boldsymbol{C}(P)^{1/2}\boldsymbol{A}_1(P)\boldsymbol{C}(P)^{-1/2} = -\boldsymbol{C}(P)^{-1/2}\boldsymbol{D}(P)\boldsymbol{C}(P)^{-1/2}. \tag{10.11}$$

The matrix on the right is symmetric, so this shows that $\boldsymbol{A}_1(P)$ is similar to a symmetric matrix, allowing us to deduce that $\boldsymbol{A}_1(P)$ has real eigenvalues. These eigenvalues are also roots of the characteristic polynomial

$$s^{n/2} + \frac{p_{n-2}}{p_n}s^{n/2-1} + \cdots + \frac{p_2}{p_n}s + \frac{p_0}{p_n}.$$

Our assumption (i) ensures that is s is real and nonnegative, the value of the characteristic polynomial is positive. From this we deduce that all eigenvalues of $\boldsymbol{A}_1(P)$ are negative. From (10.11) it now follows that $\boldsymbol{D}(P)$ is positive-definite, and so P is Hurwitz by Lemma 10.4.7 and Theorem 10.4.5.

Case 2: We assume (ii) and that n is even. Consider the polynomial $P^{-1}(s) = s^n P(\frac{1}{s})$. Clearly the roots of P^{-1} are the reciprocals of those for P. Thus P^{-1} is Hurwitz if and only if P is Hurwitz (see Exercise 10.4.2). Also, the coefficients for P^{-1} are obtained by inverting those for P. Using this facts, one can see that $\boldsymbol{C}(P^{-1})$ is obtained from $\boldsymbol{D}(P)$ by reversing the rows and columns, and that $\boldsymbol{D}(P^{-1})$ is obtained from $\boldsymbol{C}(P)$ by reversing the rows and columns. One can then show that P^{-1} is Hurwitz just as in Case 1, and from this it follows that P is Hurwitz.

Case 3: We assume (iii) and that n is odd. In this case we let

$$\boldsymbol{A}_2(P) = \begin{bmatrix} -\frac{p_{n-2}}{p_n} & -\frac{p_{n-4}}{p_n} & \cdots & -\frac{p_1}{p_n} & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

and note that one can check to see that

$$\boldsymbol{C}(P)\boldsymbol{A}_2(P) = -\begin{bmatrix} \boldsymbol{D}(P) & \boldsymbol{0} \\ \boldsymbol{0}^T & 0 \end{bmatrix}. \tag{10.12}$$

As in Case 1, we may define the square root, $\boldsymbol{C}(P)^{1/2}$, of $\boldsymbol{C}(P)$, and ascertain that

$$\boldsymbol{C}(P)^{1/2}\boldsymbol{A}_2(P)\boldsymbol{C}(P)^{-1/2} = -\boldsymbol{C}(P)^{-1/2}\begin{bmatrix} \boldsymbol{D}(P) & \boldsymbol{0} \\ \boldsymbol{0}^T & 0 \end{bmatrix}\boldsymbol{C}(P)^{-1/2}.$$

Again, the conclusion is that $\boldsymbol{A}_2(P)$ is similar to a symmetric matrix, and so must have real eigenvalues. These eigenvalues are the roots of the characteristic polynomial

$$X^{(n+1)/2} + \frac{p_{n-2}}{p_n}X^{(n+1)/2-1} + \cdots + \frac{p_1}{p_n}X.$$

This polynomial clearly has a zero root. However, since (iii) holds, for positive real values of X, the characteristic polynomial takes on positive values, so the nonzero eigenvalues of $\boldsymbol{A}_2(P)$ must be negative, and there are $\frac{n+1}{2} - 1$ of these. From this and (10.12) it follows that the matrix

$$\begin{bmatrix} \boldsymbol{D}(P) & \boldsymbol{0} \\ \boldsymbol{0}^T & 0 \end{bmatrix}$$

has one zero eigenvalue and $\frac{n+1}{2} - 1$ positive real eigenvalues. Thus $\boldsymbol{D}(P)$ must be positive-definite, and P is then Hurwitz by Lemma 10.4.7 and Theorem 10.4.5.

Case 4: We assume (i) and that n is odd. As in Case 2, define $P^{-1}(X) = X^n P(\frac{1}{X})$. In this case one can ascertain that $\boldsymbol{C}(P^{-1})$ is obtained from $\boldsymbol{C}(P)$ by reversing rows and columns, and that $\boldsymbol{D}(P^{-1})$ is obtained from $\boldsymbol{D}(P)$ by reversing rows and columns. The difference from the situation in Case 2 arises because here we are taking n odd, while in Case 2 it was even. In any event, one may now apply Case 3 to P^{-1} to show that P^{-1} is Hurwitz. Then P is itself Hurwitz by Exercise 10.4.2.

Case 5: We assume (ii) and that n is odd. For $\epsilon > 0$ define $P_\epsilon \in \mathbb{R}[X]$ by $P_\epsilon(X) = (X + \epsilon)P(X)$. Thus the degree of P_ϵ is now even. Indeed,

$$P_\epsilon(X) = p_n X^{n+1} + (p_{n-1} + \epsilon p_n)X^n + \cdots + (p_0 + \epsilon p_1)X + \epsilon p_0.$$

One may readily determine that

$$\boldsymbol{C}(P_\epsilon) = \boldsymbol{C}(P) + \epsilon \boldsymbol{C}$$

for some matrix $\boldsymbol{C}$ which is independent of ϵ. In like manner, one may show that

$$\boldsymbol{D}(P_\epsilon) = \begin{bmatrix} \boldsymbol{D}(P) + \epsilon \boldsymbol{D}_{11} & \epsilon \boldsymbol{D}_{12} \\ \epsilon \boldsymbol{D}_{12} & \epsilon p_0^2 \end{bmatrix},$$

where $\boldsymbol{D}_{11}$ and $\boldsymbol{D}_{12}$ are independent of ϵ. Since $\boldsymbol{D}(P)$ is positive-definite and $a_0 > 0$, for ϵ sufficiently small we must have that $\boldsymbol{D}(P_\epsilon)$ is positive-definite. From the argument of Case 2, we may infer that P_ϵ is Hurwitz, from which it is obvious that P is also Hurwitz.

Case 6: We assume (iv) and that n is odd. We define $P^{-1}(X) = X^n P(\frac{1}{X})$ so that $\boldsymbol{C}(P^{-1})$ is obtained from $\boldsymbol{C}(P)$ by reversing rows and columns, and that $\boldsymbol{D}(P^{-1})$ is obtained from $\boldsymbol{D}(P)$ by reversing rows and columns. One can now use Case 5 to show that P^{-1} is Hurwitz, and so P is also Hurwitz by Exercise 10.4.2.

Case 7: We assume (iii) and that n is even. As with Case 5, we define $P_\epsilon(X) = (X + \epsilon)P(X)$ and in this case we compute

$$\boldsymbol{C}(P_\epsilon) = \begin{bmatrix} \boldsymbol{C}(P) + \epsilon \boldsymbol{C}_{11} & \epsilon \boldsymbol{C}_{12} \\ \epsilon \boldsymbol{C}_{12} & \epsilon p_0^2 \end{bmatrix}$$

and

$$\boldsymbol{D}(P_\epsilon) = \boldsymbol{D}(P) + \epsilon \boldsymbol{D},$$

where $\boldsymbol{C}_{11}$, $\boldsymbol{C}_{12}$, and $\boldsymbol{D}$ are independent of ϵ. By our assumption (iii), for $\epsilon > 0$ sufficiently small we have $\boldsymbol{C}(P_\epsilon)$ positive-definite. Thus, invoking the argument of Case 1, we may deduce that $\boldsymbol{D}(P_\epsilon)$ is also positive-definite. Therefore P_ϵ is Hurwitz by Lemma 10.4.7 and Theorem 10.4.3. Thus P is itself also Hurwitz.

Case 8: We assume (iv) and that n is even. Taking $P^{-1}(X) = X^n P(\frac{1}{X})$ we see that $\boldsymbol{C}(P^{-1})$ is obtained from $\boldsymbol{D}(P)$ by reversing the rows and columns, and that $\boldsymbol{D}(P^{-1})$ is obtained from $\boldsymbol{C}(P)$ by reversing the rows and columns. Now one may apply Case 7 to deduce that P^{-1}, and therefore P, is Hurwitz. ■

10.4.4 The Liénard–Chipart criterion

Although less well-known than the criterion of Routh and Hurwitz, the test we give next has the advantage of delivering fewer determinantal inequalities to test. This results from their being a dependence on some of the Hurwitz determinants.

10.4.9 Theorem (Liénard–Chipart criterion) *A polynomial*

$$P = X^n + p_{n-1}X^{n-1} + \cdots + p_1X + p_0 \in \mathbb{R}[X]$$

is Hurwitz if and only if any one of the following conditions holds:

(i) $p_{2k} > 0$, $k \in \{0, 1, \ldots, \lfloor \frac{n-1}{2} \rfloor\}$ *and* $\Delta_{2k+1} > 0$, $k \in \{0, 1, \ldots, \lfloor \frac{n-1}{2} \rfloor\}$;

(ii) $p_{2k} > 0$, $k \in \{0, 1, \ldots, \lfloor \frac{n-1}{2} \rfloor\}$ *and* $\Delta_{2k} > 0$, $k \in \{1, \ldots, \lfloor \frac{n}{2} \rfloor\}$;

(iii) $p_0 > 0$, $p_{2k+1} > 0$, $k \in \{0, 1, \ldots, \lfloor \frac{n-2}{2} \rfloor\}$ *and* $\Delta_{2k+1} > 0$, $k \in \{0, 1, \ldots, \lfloor \frac{n-1}{2} \rfloor\}$;

(iv) $p_0 > 0$, $p_{2k+1} > 0$, $k \in \{0, 1, \ldots, \lfloor \frac{n-2}{2} \rfloor\}$ *and* $\Delta_{2k} > 0$, $k \in \{1, \ldots, \lfloor \frac{n}{2} \rfloor\}$.

Here $\Delta_1, \ldots, \Delta_n$ *are the Hurwitz determinants.*

Proof The theorem follows immediately from and Theorem 10.4.8, after one checks lem:minors
that the principal minors of $\boldsymbol{C}(P)$ are exactly the odd Hurwitz determinants $\Delta_1, \Delta_3, \ldots$, and that the principal minors of $\boldsymbol{D}(P)$ are exactly the even Hurwitz determinants $\Delta_2, \Delta_4, \ldots$. ■

The advantage of the Liénard–Chipart test over the Hurwitz test is that one will generally have fewer determinants to compute. Let us illustrate the criterion in the simplest case, when $n = 2$.

10.4.10 Example (Liénard–Chipart criterion) We consider the polynomial $P = X^2 + aX + b$. Recall that the Hurwitz determinants were computed in Example 10.4.4:

$$\Delta_1 = a, \quad \Delta_2 = ab.$$

Let us write down the four conditions of Theorem 10.4.9:

1. $p_0 = b > 0$, $\Delta_1 = a > 0$;
2. $p_0 = b > 0$, $\Delta_2 = ab > 0$;
3. $p_0 = b > 0$, $p_1 = a > 0$, $\Delta_1 = a > 0$;
4. $p_0 = b > 0$, $p_1 = a > 0$, $\Delta_2 = ab > 0$.

We see that all of these conditions are equivalent in this case, and imply that P is Hurwitz if and only if $a, b > 0$, as expected. This example is really too simple to illustrate the potential advantages of the Liénard-Chipart criterion, but we refer the reader to Exercise 10.4.3 to see how the test can be put to good use. •

10.4.5 Kharitonov's test

It is sometimes the case that one does not know exactly the coefficients for a given polynomial. In such instances, one may know bounds on the coefficients. That is, for a polynomial

$$P(s) = p_n s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0, \tag{10.13}$$

one may know that the coefficients satisfy inequalities of the form

$$p_i^{\min} \leq p_i \leq p_i^{\max}, \quad i = 0, 1, \ldots, n. \tag{10.14}$$

In this case, the following remarkable theorem gives a simple test for the stability of the polynomial for all possible values for the coefficients.

10.4.11 Theorem (Kharitonov's criterion) *Given a polynomial of the form (10.13) with the coefficients satisfying the inequalities (10.14), define four polynomials*

$$\begin{aligned}
Q_1(s) &= p_0^{\min} + p_1^{\min}s + p_2^{\max}s^2 + p_3^{\max}s^3 + \cdots\\
Q_2(s) &= p_0^{\min} + p_1^{\max}s + p_2^{\max}s^2 + p_3^{\min}s^3 + \cdots\\
Q_3(s) &= p_0^{\max} + p_1^{\max}s + p_2^{\min}s^2 + p_3^{\min}s^3 + \cdots\\
Q_4(s) &= p_0^{\max} + p_1^{\min}s + p_2^{\min}s^2 + p_3^{\max}s^3 + \cdots
\end{aligned}$$

Then P *is Hurwitz for all*

$$(p_0, p_1, \ldots, p_n) \in [p_0^{\min}, p_0^{\max}] \times [p_1^{\min}, p_1^{\max}] \times \cdots \times [p_n^{\min}, p_n^{\max}]$$

if and only if the polynomials Q_1, Q_2, Q_3, *and* Q_4 *are Hurwitz.*

Proof Let us first assume without loss of generality that $p_j^{\min} > 0$, $j = 0, \ldots, n$. Indeed, by Exercise 10.4.1, for a polynomial to be Hurwitz, its coefficients must have the same sign, and we may as well suppose this sign to be positive. If

$$\boldsymbol{p} = (p_0, p_1, \ldots, p_n) \in [p_0^{\min}, p_0^{\min}] \times [p_1^{\min}, p_1^{\min}] \times \cdots \times [p_n^{\min}, p_n^{\min}],$$

then let us say, for convenience, that $\boldsymbol{p}$ is ***allowable***. For $\boldsymbol{p}$ allowable denote

$$P_{\boldsymbol{p}}(s) = p_n s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0.$$

It is clear that if all polynomials $P_{\boldsymbol{p}}$ are allowable then the polynomials Q_1, Q_2, Q_3, and Q_4 are Hurwitz. Thus suppose for the remainder of the proof that Q_1, Q_2, Q_3, and Q_4 are Hurwitz, and we shall deduce that $P_{\boldsymbol{p}}$ is also Hurwitz for every allowable $\boldsymbol{p}$.

For $\omega \in \mathbb{R}$ define

$$R(\omega) = \{P_{\boldsymbol{p}}(i\omega) \mid \boldsymbol{p} \text{ allowable}\}.$$

The following property of $R(\omega)$ lies at the heart of our proof. It is first noticed by Dasgupta [1988].

1 Lemma *For each* $\omega \in \mathbb{R}$, $R(\omega)$ *is a rectangle in* $\mathbb{C}$ *whose sides are parallel to the real and imaginary axes, and whose corners are* $Q_1(i\omega)$, $Q_2(i\omega)$, $Q_3(i\omega)$, *and* $Q_4(i\omega)$.

Proof We note that for $\omega \in \mathbb{R}$ we have

$$\begin{aligned}
\operatorname{Re}(Q_1(i\omega)) &= \operatorname{Re}(Q_2(i\omega)) = p_0^{\min} - p^{\max}\omega^2 + p_4^{\min}\omega^4 + \cdots\\
\operatorname{Re}(Q_3(i\omega)) &= \operatorname{Re}(Q_4(i\omega)) = p_0^{\max} - p^{\min}\omega^2 + p_4^{\max}\omega^4 + \cdots\\
\operatorname{Im}(Q_1(i\omega)) &= \operatorname{Im}(Q_4(i\omega)) = \omega\big(p^{\min} - p^{\max}\omega^2 + p_4^{\min}\omega^4 + \cdots\big)\\
\operatorname{Im}(Q_2(i\omega)) &= \operatorname{Im}(Q_3(i\omega)) = \omega\big(p^{\max} - p^{\min}\omega^2 + p_4^{\max}\omega^4 + \cdots\big).
\end{aligned}$$

From this we deduce that for any allowable $\boldsymbol{p}$ we have

$$\begin{aligned}
\operatorname{Re}(Q_1(i\omega)) = \operatorname{Re}(Q_2(i\omega)) \le \operatorname{Re}(P_{\boldsymbol{p}}(i\omega)) \le \operatorname{Re}(Q_3(i\omega)) = \operatorname{Re}(Q_4(i\omega))\\
\operatorname{Im}(Q_1(i\omega)) = \operatorname{Im}(Q_4(i\omega)) \le \operatorname{Im}(P_{\boldsymbol{p}}(i\omega)) \le \operatorname{Im}(Q_2(i\omega)) = \operatorname{Im}(Q_3(i\omega)).
\end{aligned}$$

This leads to the picture shown in Figure 10.8 for $R(\omega)$. The lemma follows immediately from this. ▼

Using the lemma, we now claim that if $\boldsymbol{p}$ is allowable, then $P_{\boldsymbol{p}}$ has no imaginary axis roots. To do this, we record the following useful property of Hurwitz polynomials.

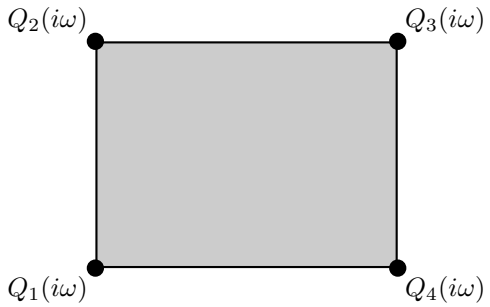


Figure 10.8 $R(\omega)$

2 Lemma *If* $\mathrm{P} \in \mathbb{R}[\mathrm{s}]$ *is monic and Hurwitz with* $\deg(\mathrm{P}) \geq 1$, *then* $\arg(\mathrm{P}(\mathrm{i}\omega))$ *is a continuous and strictly increasing function of* ω.

Proof Write

$$P(s) = \prod_{j=1}^{n}(s - z_j)$$

where $z_j = \sigma_j + i\omega_j$ with $\sigma_j < 0$. Thus

$$\arg(P(i\omega)) = \sum_{j=1}^{n} \arg((i\omega + |\sigma_j| - i\omega_j)) = \sum_{j=1}^{n} \arctan\Big(\frac{\omega - \omega_j}{|\sigma_j|}\Big).$$

Since $|\sigma_j| > 0$, each term in the sum is continuous and strictly increasing, and thus so too is $\arg(P(i\omega))$. ▼

To show that $0 \notin R(\omega)$ for $\omega \in \mathbb{R}$, first note that $0 \notin R(0)$. Now, since the corners of $R(\omega)$ are continuous functions of ω, if $0 \in R(\omega)$ for some $\omega > 0$, then it must be the case that for some $\omega_0 \in [0, \omega]$ the point $0 \in \mathbb{C}$ lies on the boundary of $R(\omega_0)$. Suppose that 0 lies on the lower boundary of the rectangle $R(\omega_0)$. This means that $Q_1(i\omega_0) < 0$ and $Q_4(i\omega_0) > 0$ since the corners of $R(\omega)$ cannot pass through 0. Since Q_1 is Hurwitz, by Lemma 2 we must have $Q_1(i(\omega_0 + \delta))$ in the $(-,-)$ quadrant in $\mathbb{C}$ and $Q_4(i(\omega_0 + \delta))$ in the $(+,+)$ quadrant in $\mathbb{C}$ for $\delta > 0$ sufficiently small. However, since $\mathrm{Im}(Q_1(i\omega)) = \mathrm{Im}(Q_4(i\omega))$ for all $\omega \in \mathbb{R}$, this cannot be. Therefore 0 cannot lie on the lower boundary of $R(\omega_0)$ for any $\omega_0 > 0$. Similar arguments establish that 0 cannot lie on either of the other three boundaries either. This then prohibits 0 from lying in $R(\omega)$ for any $\omega > 0$.

Now suppose that $P_{\boldsymbol{p}_0}$ is not Hurwitz for some allowable $\boldsymbol{p}_0$. For $\lambda \in [0, 1]$ each of the polynomials

$$\lambda Q_1 + (1 - \lambda)P_{\boldsymbol{p}_0} \tag{10.15}$$

is of the form $P_{\boldsymbol{p}_\lambda}$ for some allowable $\boldsymbol{p}_\lambda$. Indeed, the equation (10.15) defines a straight line from Q_1 to $P_{\boldsymbol{p}_0}$, and since the set of allowable $\boldsymbol{p}$'s is convex (it is a cube), this line remains in the set of allowable polynomial coefficients. Now, since Q_1 is Hurwitz and $P_{\boldsymbol{p}_0}$ is not, by continuity of the roots of a polynomial with respect to the coefficients, we deduce that for some $\lambda \in [0, 1)$, the polynomial $P_{\boldsymbol{p}_\lambda}$ must have an imaginary axis

root. However, we showed above that $0 \notin R(\omega)$ for all $\omega \in \mathbb{R}$, denying the possibility of such imaginary axis roots. Thus all polynomials P_p are Hurwitz for allowable p. ■

10.4.12 Remarks

1. Note the pattern of the coefficients in the polynomials Q_1, Q_2, Q_3, and Q_4 has the form $(\dots, \max, \max, \min, \min, \dots)$ This is charmingly referred to as the ***Kharitonov melody***.
2. One would anticipate that to check the stability for P one should look at all possible extremes for the coefficients, giving 2^n polynomials to check. That this can be reduced to four polynomial checks is an unobvious simplification. •

Let us apply the Kharitonov test in the simplest case when $n = 2$.

10.4.13 Example We consider

$$P(s) = s^2 + as + b$$

with the coefficients satisfying

$$(a, b) \in [a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}].$$

The polynomials required by Theorem 10.4.11 are

$$\begin{aligned} Q_1(s) &= s^2 + a_{\min}s + b_{\min} \\ Q_2(s) &= s^2 + a_{\max}s + b_{\min} \\ Q_3(s) &= s^2 + a_{\max}s + b_{\max} \\ Q_4(s) &= s^2 + a_{\min}s + b_{\max}. \end{aligned}$$

We now apply the Routh/Hurwitz criterion to each of these polynomials. This indicates that all coefficients of the four polynomials Q_1, Q_2, Q_3, and Q_4 should be positive. This reduces to requiring that

$$a_{\min}, a_{\max}, b_{\min}, b_{\max} > 0.$$

That is, $a_{\min}, b_{\min} > 0$. In this simple case, we could have guessed the result ourselves since the Routh/Hurwitz criterion are so simple to apply for degree two polynomials. Nonetheless, the simple example illustrates how to apply Theorem 10.4.11. •

10.4.6 Notes

It is interesting to note that the method of Edward John Routh (1831–1907) was developed in response to a famous paper of James Clerk Maxwell[6] (1831–1879) on the use of governors to control a steam engine. This paper of Maxwell [1868] can be regarded as the first paper in mathematical control theory.

[6]Maxwell, of course, is better known for his famous equations of electromagnetism.

Theorem 10.4.1 is due to Routh [1877].

Theorem 10.4.3 is due to Hurwitz [1895].

Theorem 10.4.5 is due to Charles Hermite (1822–1901) [see Hermite 1854]. The slick proof using Lyapunov methods comes from the paper of Parks [1962].

Our proof of Theorem 10.4.8 follows that of Anderson [1972].

Theorem 10.4.9 is from Liénard and Chipart [1914][7] This is given thorough discussion by Gantmacher [1959]. Here we state the result, and give a proof due to Anderson [1972] that is more elementary than that of Gantmacher. The observation in the proof of Theorem 10.4.9 is made by a computation which we omit, and appears to be first been noticed by Fujiwara [1915].

Theorem 10.4.11 is due to Kharitonov [1978]. Since the publication of Kharitonov's result, or more properly its discovery by the non-Russian speaking world, there have been many simplifications of the proof [e.g., Chapellat and Bhattacharyya 1989, Dasgupta 1988, Mansour and Anderson 1993]. The proof we give essentially follows Minnichelli, Anagnost, and Desoer [1989]. Anderson, Jury, and Mansour [1987] observe that for polynomials of degree 3, 4, or 5, it suffices to check not four, but one, two, or three polynomials, respectively, as being Hurwitz. A proof of Kharitonov's theorem, using Lyapunov methods (see Section 10.7.4), is given by Mansour and Anderson [1993].

Exercises

10.4.1 A useful necessary condition for a polynomial to have all roots in $\mathbb{C}_-$ is given by the following theorem.

Theorem *If the polynomial*

$$P = X^n + p_{n-1}X^{n-1} + \cdots + p_1X + p_0 \in \mathbb{R}[X]$$

is Hurwitz, then the coefficients $p_0, p_1, \ldots, p_{n-1}$ *are all positive.*

(a) Prove this theorem.

(b) Is the converse of the theorem true? If so, prove it, if not, give a counterexample.

10.4.2 Consider a polynomial

$$P = p_nX^n + p_{n-1}X^{n-1} + \cdots + p_1X + p_0 \in \mathbb{R}[X]$$

with $p_0, p_n \neq 0$, and define $P^{-1} \in \mathbb{R}[X]$ by $P^{-1}(X) = X^nP(\frac{1}{X})$.

(a) Show that the roots for P^{-1} are the reciprocals of the roots for P.

(b) Show that P is Hurwitz if and only if P^{-1} is Hurwitz.

[7] Perhaps the relative obscurity of the test reflects that of its authors; I was unable to find a biographical reference for either Liénard or Chipart. I do know that Liénard did work in differential equations, with the ***Liénard equation*** being a well-studied second-order linear differential equation.

10.4.3 For the following two polynomials,

(a) $P = X^3 + aX^2 + bX + c$,

(b) $P = X^4 + aX^3 + bX^2 + cX + d$,

write down the four conditions of the Liénard–Chipart criterion, Theorem 10.4.9, and determine which is the least restrictive.

Section 10.5

Lyapunov's First (or Indirect) Method

The First Method of Lyapunov relates the stability of an equilibrium point to the stability of the linearisation about this equilibrium point. Therefore, in this section we provide a concrete impetus for the process of linearisation developed in Section 5.1. We shall discuss separately the First Method of Lyapunov in the nonautonomous and autonomous situation, since the autonomous case is much easier.

Let us briefly recall here the process of the linearisation of an ordinary differential equation $\boldsymbol{F}$ about an equilibrium state $\boldsymbol{x}_0$. We suppose that the right-hand side $\widehat{\boldsymbol{F}}$ is differentiable with respect to $\boldsymbol{x}$. Then the linearisation is the linear ordinary differential equation $\boldsymbol{F}_{L,\boldsymbol{x}_0}$ on $\mathbb{R}^n$ whose right-hand side is

$$\begin{aligned}\widehat{\boldsymbol{F}}_{L,\boldsymbol{x}_0} \colon \mathbb{T} \times \mathbb{R}^n &\to \mathbb{R}^n \\ (t, \boldsymbol{v}) &\mapsto \boldsymbol{D}\widehat{\boldsymbol{F}}_t(\boldsymbol{x}_0) \cdot \boldsymbol{v}.\end{aligned}$$

10.5.1 The First Method for nonautonomous equations

We shall work with a system of first-order ordinary differential equations $\boldsymbol{F}$ with right-hand side

$$\widehat{\boldsymbol{F}} \colon \mathbb{T} \times U \to \mathbb{R}^n,$$

where $U \subseteq \mathbb{R}^n$ is the state space, i.e., an open subset of $\mathbb{R}^n$. We shall consider an equilibrium point $\boldsymbol{x}_0 \in U$; thus, by Proposition 5.1.5, $\widehat{\boldsymbol{F}}(t, \boldsymbol{x}_0) = \boldsymbol{0}$ for all $t \in \mathbb{T}$.

The main theorem for this setting is then the following.

10.5.1 Theorem (Uniform asymptotic stability for linearisation implies uniform exponential stability for equilibria I) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}^{\mathrm{n}}$$

and let $\mathbf{x}_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$. *Assume that* $\sup \mathbb{T} = \infty$, *that* $\widehat{\mathbf{F}}$ *is continuously differentiable, and that there exist* $\mathrm{r}, \mathrm{L}, \mathrm{M} \in \mathbb{R}_{>0}$ *such that*

$$\left| \frac{\partial \widehat{\mathrm{F}}_{\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{t}, \mathbf{x}) \right| \leq \mathrm{M}, \qquad (\mathrm{t}, \mathbf{x}) \in \mathbb{T} \times \overline{\mathsf{B}}(\mathrm{r}, \mathbf{x}_0),\ \mathrm{j}, \mathrm{k} \in \{1, \dots, \mathrm{n}\}, \tag{10.16}$$

and

$$\left| \frac{\partial \widehat{\mathrm{F}}_{\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{t}, \mathbf{x}_1) - \frac{\partial \widehat{\mathrm{F}}_{\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{t}, \mathbf{x}_2) \right| \leq \mathrm{L} \|\mathbf{x}_1 - \mathbf{x}_2\|, \qquad \mathrm{t} \in \mathbb{T},\ \mathbf{x}_1, \mathbf{x}_2 \in \overline{\mathsf{B}}(\mathrm{r}, \mathbf{x}_0),\ \mathrm{j}, \mathrm{k} \in \{1, \dots, \mathrm{n}\}. \tag{10.17}$$

Then $\mathbf{x}_0$ is uniformly exponentially stable if its linearisation is uniformly asymptotically stable.

Proof First let us deduce some consequences of $\boldsymbol{F}$ satisfying the hypotheses of the theorem statement.

1 Lemma *If* $\mathbf{F}$ *is an ordinary differential equation whose right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathsf{U} \to \mathbb{R}^{\mathrm{n}}$$

satisfies:

(i) $\widehat{\mathbf{F}}$ *is continuously differentiable;*

(ii) there exist $\mathrm{r}, \mathrm{L}, \mathrm{M} \in \mathbb{R}_{>0}$ *such that (10.16) and (10.17) hold.*

Then there exists $\widehat{\mathbf{G}}\colon \mathbb{T}\times \mathsf{B}(\mathrm{r}, \mathbf{x}_0) \to \mathbb{R}^{\mathrm{n}}$ *and* $\mathrm{C} \in \mathbb{R}_{>0}$ *such that*

$$\widehat{\mathrm{F}}_{\mathrm{j}}(\mathrm{t}, \mathbf{x}) = \sum_{\mathrm{k}=1}^{\mathrm{n}} \frac{\partial \widehat{\mathrm{F}}_{\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{t}, \mathbf{x}_0)(\mathrm{x}_{\mathrm{k}} - \mathrm{x}_{0,\mathrm{k}}) + \widehat{\mathrm{G}}_{\mathrm{j}}(\mathrm{t}, \mathbf{x}), \qquad (\mathrm{t}, \mathbf{x}) \in \mathbb{T}\times \mathsf{B}(\mathrm{r}, \mathbf{x}_0),$$

where

$$\|\widehat{\mathbf{G}}(\mathrm{t}, \mathbf{x})\| \leq \mathrm{C}\|\mathbf{x} - \mathbf{x}_0\|^2, \qquad (\mathrm{t}, \mathbf{x}) \in \mathbb{T}\times \mathsf{B}(\mathrm{r}, \mathbf{x}_0) \tag{10.18}$$

ref *Proof* By the Mean Value Theorem, , we can write

$$\widehat{F}_j(t, \boldsymbol{x}) = \widehat{F}_j(t, \boldsymbol{x}_0) + \sum_{k=1}^{n} \frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{y})(x_k - x_{0,k})$$

for some $\boldsymbol{y} = s\boldsymbol{x}_0 + (1-s)\boldsymbol{x}$, $s \in [0,1]$. Since $\boldsymbol{x}_0$ is an equilibrium point, we rewrite this as

$$\widehat{F}_j(t, \boldsymbol{x}) = \sum_{k=1}^{n} \frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{x}_0)(x_k - x_{0,k}) + \sum_{k=1}^{n} \left(\frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{y}) - \frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{x}_0) \right)(x_k - x_{0,k}).$$

If we define

$$\widehat{G}_j = \sum_{k=1}^{n} \left(\frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{y}) - \frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{x}_0) \right)(x_k - x_{0,k}),$$

it only remains to verify the estimate (10.18) for a suitable $C \in \mathbb{R}_{>0}$. By the Cauchy–Bunyakovsky–Schwarz inequality, we have

$$\begin{aligned}
\|\widehat{G}(t, \boldsymbol{x})\| &= \left(\sum_{j=1}^{n} \left(\sum_{k=1}^{n} \left(\frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{y}) - \frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{x}_0) \right)(x_k - x_{0,k}) \right)^2 \right)^{1/2} \\
&\leq \left(\sum_{j=1}^{n} \left(\sum_{k=1}^{n} \left(\frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{y}) - \frac{\partial \widehat{F}_j}{\partial x_k}(t, \boldsymbol{x}_0) \right)^2 \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 \right) \right)^{1/2} \\
&\leq \left(\sum_{j=1}^{n} L^2 \|\boldsymbol{y} - \boldsymbol{x}_0\|^2 \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 \right)^{1/2} \\
&= \sqrt{n} L (1-s) \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 \leq \sqrt{n} L \|\boldsymbol{x} - \boldsymbol{x}_0\|^2,
\end{aligned}$$

and the lemma follows taking $C = \sqrt{n}L$. ▼

For brevity, let us denote $\boldsymbol{A}(t) = \widehat{\boldsymbol{DF}}(t, x_0)$. The assumptions of the theorem ensure that $\boldsymbol{A}$ satisfies the hypotheses of Theorem 10.10.2. Thus, since the linearisation is uniformly asymptotically stable, there exists $\boldsymbol{P}\colon \mathbb{T} \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ such that $(\boldsymbol{P}, \boldsymbol{I}_n)$ is a Lyapunov pair for $\boldsymbol{F}_{L,x_0}$. We define

$$\begin{aligned} V\colon \mathbb{T} \times U &\to \mathbb{R} \\ (t, x) &\mapsto f_{\boldsymbol{P}}(t, x - x_0). \end{aligned}$$

Let $(t_0, x) \in \mathbb{T} \times \mathsf{B}(r, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = x.$$

Then calculate, using the lemma above,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} V(t, \xi(t)) &= \frac{\mathrm{d}}{\mathrm{d}t} \langle \boldsymbol{P}(t)(\xi(t) - x_0), \xi(t) - x_0 \rangle_{\mathbb{R}^n} \\ &= \langle \dot{\boldsymbol{P}}(t)(\xi(t)), \xi(t) - x_0 \rangle_{\mathbb{R}^n} + \langle \boldsymbol{P}(t)(\widehat{\boldsymbol{F}}(t, \xi(t))), \xi(t) - x_0 \rangle_{\mathbb{R}^n} \\ &\quad + \langle \boldsymbol{P}(t)(\xi(t) - x_0), \widehat{\boldsymbol{F}}(t, \xi(t)) \rangle_{\mathbb{R}^n} \\ &= \langle (\dot{\boldsymbol{P}}(t) + \boldsymbol{P}(t)\boldsymbol{A}(t) + \boldsymbol{A}^T(t)\boldsymbol{P}(t))(\xi(t) - x_0), \xi(t) - x_0 \rangle_{\mathbb{R}^n} \\ &\quad + 2\langle \boldsymbol{P}(t)(\xi(t) - x_0), \widehat{\boldsymbol{G}}(t, \xi(t)) \rangle_{\mathbb{R}^n} \\ &= -\|\xi(t) - x_0\|^2 + 2\langle \boldsymbol{P}(t)(\xi(t) - x_0), \widehat{\boldsymbol{G}}(t, \xi(t)) \rangle_{\mathbb{R}^n}. \end{aligned}$$

Evaluating at $t = t_0$ and using Lemma 10.7.3, this shows that

$$\mathscr{L}_{\boldsymbol{F}} V(t_0, x) = -\|x - x_0\|^2 + 2\langle \boldsymbol{P}(t_0)(x - x_0), \widehat{\boldsymbol{G}}(t_0, x) \rangle_{\mathbb{R}^n}$$

for $(t_0, x) \in \mathbb{T} \times \mathsf{B}(r, x_0)$. By Lemma 10.6.17, let $B \in \mathbb{R}_{>0}$ be such that

$$B\|v\|^2 \leq \|\boldsymbol{P}(t)(v)\|^2 \leq B^{-1}\|v\|^2, \qquad (t, v) \in \mathbb{T} \times \mathbb{R}^n.$$

We have

$$\begin{aligned} |\langle \boldsymbol{P}(t)(x - x_0), \widehat{\boldsymbol{G}}(t, x) \rangle_{\mathbb{R}^n}| &\leq \|\boldsymbol{P}(t)(x - x_0)\| \|\widehat{\boldsymbol{G}}(t, x)\| \\ &\leq C\sqrt{B^{-1}}\|x - x_0\|^3 \leq C\sqrt{B^{-1}} r \|x - x_0\|^2, \end{aligned}$$

where C is as in the lemma. Therefore, if we shrink r sufficiently that $1 - 2C\sqrt{B^{-1}}r > \frac{1}{2}$, then

$$\mathscr{L}_{\boldsymbol{F}} V(t, x) \leq -\frac{1}{2}\|x - x_0\|^2, \qquad (t, x) \in \mathbb{T} \times \mathsf{B}(r, x_0).$$

Since we also have

$$B\|x - x_0\|^2 \leq V(t, x) \leq B^{-1}\|x - x_0\|^2, \qquad (t, x) \in \mathbb{T} \times \mathsf{B}(r, x_0),$$

by Theorem 10.10.2, the theorem follows from Theorem 10.7.6. ∎

10.5.2 The First Method for autonomous equations

Next we turn to Lyapunov's First Method for determining the stability of equilibria for nonautonomous ordinary differential equations.

10.5.2 Theorem (Asymptotic stability for linearisation implies exponential stability for equilibria II) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U} \to \mathbb{R}^{\mathrm{n}}$$
$$(\mathrm{t}, \mathbf{x}) \mapsto \widehat{\mathbf{F}}_0(\mathbf{x})$$

and let $\mathbf{x}_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}_0$. *Assume that* $\sup \mathbb{T} = \infty$, *that* $\widehat{\mathbf{F}}$ *is continuously differentiable, and that there exist* $\mathrm{r}, \mathrm{L} \in \mathbb{R}_{>0}$ *such that*

$$\left|\frac{\partial \widehat{\mathrm{F}}_{0,\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathbf{x}_1) - \frac{\partial \widehat{\mathrm{F}}_{0,\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathbf{x}_2)\right| \leq \mathrm{L}\|\mathbf{x}_1 - \mathbf{x}_2\|, \qquad \mathbf{x}_1, \mathbf{x}_2 \in \overline{\mathrm{B}}(\mathrm{r}, \mathbf{x}_0),\ \mathrm{j}, \mathrm{k} \in \{1, \dots, \mathrm{n}\}. \tag{10.19}$$

Then $\mathbf{x}_0$ *is exponentially stable if its linearisation is asymptotically stable.*

We offer two proofs of this theorem, one assuming Theorem 10.5.1 and the other an independent proof.

Proof of Theorem 10.5.2, assuming Theorem 10.5.1 The hypotheses of Theorem 10.5.2 clearly imply those of Theorem 10.5.1 since, in Theorem 10.5.2, $\widehat{F}$ is independent of t. Therefore, the hypotheses of Theorem 10.5.2 imply the conclusions of Theorem 10.5.1, i.e., that uniform asymptotic stability of the linearisation implies uniform exponential stability of the equilibrium. The proof in this case is concluded by recalling from Proposition 10.2.5 that the various flavours of uniform stability are equivalent to the corresponding flavours of stability for autonomous equations. ■

Independent proof of Theorem 10.5.2 First let us deduce some consequences of $\boldsymbol{F}$ satisfying the hypotheses of the theorem statement.

1 Lemma *If* $\mathbf{F}$ *is an autonomous ordinary differential equation whose right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U} \to \mathbb{R}^{\mathrm{n}}$$
$$(\mathrm{t}, \mathbf{x}) \mapsto \widehat{\mathbf{F}}_0(\mathbf{x})$$

satisfies:

(i) $\widehat{\mathbf{F}}_0$ *is continuously differentiable;*

(ii) there exist $\mathrm{r}, \mathrm{L} \in \mathbb{R}_{>0}$ *such that (10.19) holds.*

Then there exists $\widehat{\mathbf{G}}_0\colon \mathrm{B}(\mathrm{r}, \mathbf{x}_0) \to \mathbb{R}^{\mathrm{n}}$ *and* $\mathrm{C} \in \mathbb{R}_{>0}$ *such that*

$$\widehat{\mathrm{F}}_{0,\mathrm{j}}(\mathrm{t}, \mathbf{x}) = \sum_{\mathrm{k}=1}^{\mathrm{n}} \frac{\partial \widehat{\mathrm{F}}_{0,\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathbf{x}_0)(\mathrm{x}_{\mathrm{k}} - \mathrm{x}_{0,\mathrm{k}}) + \widehat{\mathrm{G}}_{0,\mathrm{j}}(\mathbf{x}), \qquad \mathbf{x} \in \mathrm{B}(\mathrm{r}, \mathbf{x}_0),$$

where

$$\|\widehat{\mathbf{G}}_0(\mathbf{x})\| \leq \mathrm{C}\|\mathbf{x} - \mathbf{x}_0\|^2, \qquad \mathbf{x} \in \mathrm{B}(\mathrm{r}, \mathbf{x}_0) \tag{10.20}$$

ref *Proof* By the Mean Value Theorem, , we can write

$$\widehat{F}_{0,j}(\boldsymbol{x}) = \widehat{F}_{0,j}(\boldsymbol{x}_0) + \sum_{k=1}^{n} \frac{\partial \widehat{F}_{0,j}}{\partial x_k}(\boldsymbol{y})(x_k - x_{0,k})$$

for some $y = sx_0 + (1-s)x$, $s \in [0,1]$. Since x_0 is an equilibrium point, we rewrite this as

$$\widehat{F}_{0,j}(x) = \sum_{k=1}^{n} \frac{\partial \widehat{F}_{0,j}}{\partial x_k}(x_0)(x_k - x_{0,k}) + \sum_{k=1}^{n} \left(\frac{\partial \widehat{F}_{0,j}}{\partial x_k}(y) - \frac{\partial \widehat{F}_{0,j}}{\partial x_k}(x_0) \right) (x_k - x_{0,k}).$$

If we define

$$\widehat{G}_{0,j} = \sum_{k=1}^{n} \left(\frac{\partial \widehat{F}_{0,j}}{\partial x_k}(y) - \frac{\partial \widehat{F}_{0,j}}{\partial x_k}(x_0) \right) (x_k - x_{0,k}),$$

it only remains to verify the estimate (10.20) for a suitable $C \in \mathbb{R}_{>0}$. By the Cauchy–Bunyakovsky–Schwarz inequality, we have

$$\begin{aligned}
\|\widehat{G}_0(x)\| &= \left(\sum_{j=1}^{n} \left(\sum_{k=1}^{n} \left(\frac{\partial \widehat{F}_{0,j}}{\partial x_k}(y) - \frac{\partial \widehat{F}_{0,j}}{\partial x_k}(x_0) \right) (x_k - x_{0,k}) \right)^2 \right)^{1/2} \\
&\leq \left(\sum_{j=1}^{n} \left(\sum_{k=1}^{n} \left(\frac{\partial \widehat{F}_{0,j}}{\partial x_k}(y) - \frac{\partial \widehat{F}_{0,j}}{\partial x_k}(x_0) \right)^2 \|x - x_0\|^2 \right) \right)^{1/2} \\
&\leq \left(\sum_{j=1}^{n} L^2 \|y - x_0\|^2 \|x - x_0\|^2 \right)^{1/2} \\
&= \sqrt{n} L (1-s) \|x - x_0\|^2 \leq \sqrt{n} L \|x - x_0\|^2,
\end{aligned}$$

and the lemma follows taking $C = \sqrt{n}L$. ▼

For brevity, let us denote $A = D\widehat{F}(x_0)$. Since the linearisation is asymptotically stable, by Theorem 10.10.3 there exists $P \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ such that (P, I_n) is a Lyapunov pair for F_{L,x_0}. We define

$$\begin{aligned}
V \colon U &\to \mathbb{R} \\
x &\mapsto f_P(x - x_0).
\end{aligned}$$

Let $x \in \mathsf{B}(r, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}_0(\xi(t)), \quad \xi(0) = x.$$

Then calculate, using the lemma above,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} V(\xi(t)) &= \frac{\mathrm{d}}{\mathrm{d}t} \langle P(\xi(t) - x_0), \xi(t) - x_0 \rangle_{\mathbb{R}^n} \\
&= \langle P(\widehat{F}_0(\xi(t))), \xi(t) - x_0 \rangle_{\mathbb{R}^n} + \langle P(\xi(t) - x_0), \widehat{F}_0(\xi(t)) \rangle_{\mathbb{R}^n} \\
&= \langle (PA + A^T P)(\xi(t) - x_0), \xi(t) - x_0 \rangle_{\mathbb{R}^n} \\
&\quad + 2\langle P(\xi(t) - x_0), \widehat{G}_0(\xi(t)) \rangle_{\mathbb{R}^n} \\
&= -\|\xi(t) - x_0\|^2 + 2\langle P(\xi(t) - x_0), \widehat{G}_0(\xi(t)) \rangle_{\mathbb{R}^n}.
\end{aligned}$$

Evaluating at $t = 0$ and using Lemma 10.7.3, this shows that

$$\mathscr{L}_F V(x) = -\|x - x_0\|^2 + 2\langle P(x - x_0), \widehat{G}_0(x) \rangle_{\mathbb{R}^n}$$

for $x \in \mathsf{B}(r, x_0)$. By Lemma 10.6.13, let $B \in \mathbb{R}_{>0}$ be such that

$$B\|v\|^2 \leq \|\boldsymbol{P}(v)\|^2 \leq B^{-1}\|v\|^2, \qquad v \in \mathbb{R}^n.$$

We have

$$\begin{aligned}|\langle \boldsymbol{P}(x - x_0), \widehat{G}_0(x)\rangle_{\mathbb{R}^n}| &\leq \|\boldsymbol{P}(x - x_0)\|\|\widehat{G}_0(x)\| \\ &\leq C\sqrt{B^{-1}}\|x - x_0\|^3 \leq C\sqrt{B^{-1}}r\|x - x_0\|^2,\end{aligned}$$

where C is as in the lemma. Therefore, if we shrink r sufficiently that $1 - 2C\sqrt{B^{-1}}r > \frac{1}{2}$, then

$$\mathscr{L}_F V(x) \leq -\frac{1}{2}\|x - x_0\|^2, \qquad x \in \mathsf{B}(r, x_0).$$

Since we also have

$$B\|x - x_0\|^2 \leq V(x) \leq B^{-1}\|x - x_0\|^2, \qquad x \in \mathsf{B}(r, x_0),$$

by Theorem 10.10.3, the theorem follows from Theorem 10.7.12. ■

10.5.3 An instability theorem

In this section we give a result that allows one to determine *instability* of an equilibrium from the linearisation. We shall work here only with autonomous ordinary differential equations.

10.5.3 Theorem (Spectral instability of linearisation implies instability for equilibria) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} &\to \mathbb{R}^{\mathrm{n}} \\ (\mathrm{t}, \mathbf{x}) &\mapsto \widehat{\mathbf{F}}_0(\mathbf{x})\end{aligned}$$

and let $\mathbf{x}_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$*. Assume that* $\sup \mathbb{T} = \infty$*, that* $\widehat{\mathbf{F}}_0$ *is continuously differentiable. Then* $\mathbf{x}_0$ *is unstable if* $\operatorname{spec}(\widehat{\mathbf{F}}_{\mathrm{L},\mathbf{x}_0}) \cap \mathbb{C}_+ \neq \varnothing$*.*

Proof For brevity, let us denote $A = \widehat{F}_{L,x_0}$. First let us suppose that $\operatorname{spec}(A) \cap \mathrm{i}\mathbb{R} = \varnothing$.

some decomposition result or other

Then, according to ■

10.5.4 A converse theorem

In this section we consider the extent to which stability of the linearisation exactly characterises stability of an equilibrium point. As we know from the results and examples above, it is definitely *not* the case that stability of an equilibrium point necessitates stability of the linearisation. The following result shows that this necessity holds when the type of stability we are discussing is exponential stability.

10.5.4 Theorem (Exponential stability of an equilibrium implies exponential stability of linearisation) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}^{\mathrm{n}}$$

and let $\mathbf{x}_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$*. Assume that* $\sup \mathbb{T} = \infty$*, that* $\widehat{\mathbf{F}}$ *is continuously differentiable, and that there exist* $\mathrm{r}, \mathrm{L}, \mathrm{M} \in \mathbb{R}_{>0}$ *such that*

$$\left| \frac{\partial \widehat{\mathrm{F}}_{\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{t}, \mathbf{x}) \right| \leq \mathrm{M}, \qquad (\mathrm{t}, \mathbf{x}) \in \mathbb{T} \times \overline{\mathsf{B}}(\mathrm{r}, \mathbf{x}_0),\ \mathrm{j}, \mathrm{k} \in \{1, \dots, \mathrm{n}\}, \tag{10.21}$$

and

$$\left| \frac{\partial \widehat{\mathrm{F}}_{\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{t}, \mathbf{x}_1) - \frac{\partial \widehat{\mathrm{F}}_{\mathrm{j}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{t}, \mathbf{x}_2) \right| \leq \mathrm{L} \|\mathbf{x}_1 - \mathbf{x}_2\|, \qquad \mathrm{t} \in \mathbb{T},\ \mathbf{x}_1, \mathbf{x}_2 \in \overline{\mathsf{B}}(\mathrm{r}, \mathbf{x}_0),\ \mathrm{j}, \mathrm{k} \in \{1, \dots, \mathrm{n}\}. \tag{10.22}$$

Then $\widehat{\mathbf{F}}_{\mathrm{L}, \mathbf{x}_0}$ *is globally exponentially stable if* $\mathbf{x}_0$ *is exponentially stable.*

Proof Let us abbreviate $A(t) = \widehat{F}_{L, x_0}(t)$. Let us write

$$A(t)x = \widehat{F}(t, x) - \underbrace{(\widehat{F}(t, x) - A(t)x)}_{\widehat{G}(t, x)}.$$

According to Lemma 1 from the proof of Theorem 10.5.1, there exists $C, r \in \mathbb{R}_{>0}$ such that

$$\|\widehat{G}(t, x)\| \leq C \|x - x_0\|^2, \qquad (t, x) \in \mathbb{T} \times \mathsf{B}(r, x_0).$$

Since ∎

Section 10.6

Lyapunov functions

We will be considering functions that, intuitively, have the equilibrium point x_0 as a maximum and whose derivative along solutions is nonincreasing. It is these notions of "maximum" and "nonincreasing" that we are concerned with here. It turns out that there is a great deal to say about these seemingly simple subjects.

10.6.1 Class $\mathcal{K}$-, class $\mathcal{L}$-, and class $\mathcal{KL}$-functions

It is convenient for many of our characterisations and for many of our proofs concerning Lyapunov's Second Method to have at hand two classes of scalar functions of a real variable, which leads to another class of scalar functions of two real variables.

10.6.1 Definition (Class $\mathcal{K}$, class $\mathcal{L}$, and class $\mathcal{KL}$) Let $a \in \mathbb{R}$ and $b, b' \in \mathbb{R}_{>0} \cup \{\infty\}$.

(i) A function $\phi\colon [0, b) \to \mathbb{R}_{\geq 0}$ is of ***class*** $\mathcal{K}$ if
   (a) ϕ is continuous,
   (b) ϕ is strictly increasing, i.e., $\phi(x) < \phi(y)$ if $x < y$, and
   (c) $\phi(0) = 0$.

   By $\mathcal{K}([0, b); [0, b'))$ we denote the set of functions of class $\mathcal{K}$ with domain $[0, b)$ and codomain $[0, b')$.

(ii) A function $\psi\colon [a, \infty) \to \mathbb{R}_{\geq 0}$ is of ***class*** $\mathcal{L}$ if
   (a) ψ is continuous,
   (b) ψ is strictly decreasing, i.e., $\psi(x) > \psi(y)$ is $x < y$, and
   (c) $\lim_{x \to \infty} \psi(x) = \infty$.

   By $\mathcal{L}([a, \infty); [0, b'))$ we denote the set of functions of class $\mathcal{L}$ with domain $[a, \infty)$ and codomain $[0, b')$.

(iii) A function $\psi\colon [0, b) \times [a, \infty) \to \mathbb{R}_{\geq 0}$ is of ***class*** $\mathcal{KL}$ if
   (a) $x \mapsto \psi(x, y)$ is of class $\mathcal{K}$ for each $y \in [a, \infty)$ and
   (b) $y \mapsto \psi(x, y)$ is of class $\mathcal{L}$ for each $x \in [0, b)$.

   By $\mathcal{KL}([0, b) \times [a, \infty); [0, b'))$ we denote the set of functions of class $\mathcal{KL}$ with domain $[0, b) \times [a, \infty)$ and codomain $[0, b')$. •

These sorts of functions are often collectively referred to as "comparison functions."

Let $\phi \in \mathcal{K}([0, b); \mathbb{R}_{\geq 0})$. Since ϕ is strictly increasing, the limit $\lim_{x \to b} \phi(b)$ exists, allowing that the limit mat be ∞. For this reason, we can unambiguously write $\phi(b)$, although b is not in the domain of ϕ.

In Exercises 10.7.1, 10.7.3, and 10.7.4 the reader can sort through some examples of functions that are or are not in these classes. Here we shall enumerate a few useful properties of such functions.

10.6.2 Lemma (Properties of class $\mathcal{K}$-, class $\mathcal{L}$-, and class $\mathcal{KL}$-functions) *Let* $b, b' \in \mathbb{R}_{>0} \cup \{\infty\}$ *and* $a \in \mathbb{R}$. *Then the following statements hold:*

(i) *if* $\phi \in \mathcal{K}([0,b);\mathbb{R}_{\geq 0})$, *then* $\phi^{-1} \in \mathcal{K}([0,\phi(b));\mathbb{R}_{\geq 0})$ *is well-defined and is of class* $\mathcal{K}$;

(ii) *if* $\phi_1 \in \mathcal{K}([0,b);[0,b'))$ *and* $\phi_2 \in \mathcal{K}([0,b');\mathbb{R}_{\geq 0})$, *then* $\phi_2 \circ \phi_1$ *is of class* $\mathcal{K}$;

(iii) *if* $\phi_1 \colon [0,b) \to [0,b')$ *and* $\phi_2 \colon [0,b') \to \mathbb{R}_{\geq 0}$ *are of class* $\mathcal{K}$, *and if* $\psi \colon [0,b) \times [a,\infty) \to [0,b')$ *is of class* $\mathcal{KL}$, *then the function*

$$[0,b) \times [a,\infty) \ni (x,y) \mapsto \phi_2(\psi(\phi_1(x),y)) \in \mathbb{R}_{\geq 0}$$

is of class $\mathcal{KL}$.

Proof These are all just a matter of working through definitions, and we leave this to the reader as Exercise 10.7.2. ■

One often encounters functions that are "almost" of class $\mathcal{K}$, and in this case it is sometimes possible to bound them from below by a class $\mathcal{K}$-function.

10.6.3 Lemma (Bounding nondecreasing functions by strictly increasing functions) *Let* $b \in \mathbb{R}_{>0} \cup \{\infty\}$ *and let* $f \colon [0,b) \to \mathbb{R}_{\geq 0}$ *have the following properties:*

(i) f *is continuous;*

(ii) f *is nondecreasing, i.e.,* $f(x_1) \leq f(x_2)$ *for* $x_1 < x_2$;

(iii) $f(x) \in \mathbb{R}_{>0}$ *for* $x \in (0,b)$;

(iv) $f(0) = 0$.

Then there exist $\phi_1, \phi_2 \in \mathcal{K}([0,b);\mathbb{R}_{\geq 0})$ *such that* $\phi_1(x) \leq f(x) \leq \phi_2(x)$ *for* $x \in [0,b)$. *Moreover,* ϕ_1 *can be chosen to be locally Lipschitz.*

Proof Let $(x_j)_{j\in\mathbb{Z}}$ be the strictly increasing doubly infinite sequence in $(0,b)$ given by

$$x_j = \begin{cases} \frac{b}{2}2^j, & j \leq 0, \\ b(1 - 2^{-j-1}), & j > 0, \end{cases}$$

noting that $\lim_{j\to-\infty} x_j = 0$ and $\lim_{j\to\infty} x_j = b$. Define a doubly infinite sequence $(\alpha_j)_{j\in\mathbb{Z}}$ by

$$\alpha_j = \begin{cases} 2^{j-1}, & j \leq 0, \\ 1 - 2^{-j-1}, & j > 0. \end{cases}$$

Note that both sequences are strictly increasing and that

$$\lim_{j\to-\infty} \alpha_j = 0, \quad \lim_{j\to\infty} \alpha_j = 1.$$

Let $N_1 \in \mathbb{Z}_{>0}$ be sufficiently large that $x_{j+1} - x_j < 1$ for $j \leq -N_1$. This is possible since $(x_{-j})_{j\in\mathbb{Z}_{>0}}$ converges to 0, and so is Cauchy. Let $N_2 \in \mathbb{Z}_{\geq 0}$ be the smallest positive integer such that

$$f(x_j) - f(x_{j-1}) < 1, \qquad j \leq -N_2.$$

This is possible since $(f(x_{-j}))_{j\in\mathbb{Z}_{>0}}$ converges to 0 (by continuity of f) and so is Cauchy. Let $N = \max\{N_1, N_2\}$. Now define

$$\phi_{1,j} = \begin{cases} (x_{j+1} - x_j)\alpha_j f(x_j), & |j| \geq N, \\ \alpha_j f(x_j), & |j| < N, \end{cases}$$

and

$$\phi_{2,j} = (1 + \alpha_j) f(x_j), \qquad j \in \mathbb{Z}.$$

Here are the key observations about the doubly infinite sequences $(\phi_{1,j})_{j\in\mathbb{Z}}$ and $(\phi_{2,j})_{j\in\mathbb{Z}}$.

1. We have $\phi_{1,j-1} < \phi_{1,j}$ for $j \in \mathbb{Z}$. This follows because
   (a) $x_j - x_{j-1} < x_{j+1} - x_j < 1$ for $j \leq -N$,
   (b) $\alpha_j < \alpha_{j-1}$ for $j \in \mathbb{Z}$, and
   (c) $f(x_{j-1}) \leq f(x_j)$ for $j \leq -N$.
2. $f(x) \geq \phi_{j,1}$ for $x \in [x_j, x_{j+1})$ and $j \in \mathbb{Z}$. This follows because
   (a) $x_{j+1} - x_j < 1$ for $j \leq -N$,
   (b) $\alpha_j < 1$ for $j \in \mathbb{Z}$, and
   (c) $f(x) \geq f(x_j)$ for $x \in [x_j, x_{j+1})$.
3. $\phi_{2,j} < \phi_{2,j}$ for $j \in \mathbb{Z}$. This follows because
   (a) $1 + \alpha_j < 1 + \alpha_{j+1}$ for $j \in \mathbb{Z}$ and
   (b) $f(x_j) \leq f(x_{j+1})$ for $j \in \mathbb{Z}$.
4. $f(x) \leq \phi_{2,j}$ for $x \in [x_{j-1}, x_j)$ and $j \in \mathbb{Z}$. This follows because
   (a) $1 + \alpha_j > 1$ for $j \in \mathbb{Z}$ and
   (b) $f(x) \leq f_j(x)$ for $x \in [x_{j-1}, x_j)$ and $j \in \mathbb{Z}$.

Now define

$$\phi_1(x) = \begin{cases} 0, & x = 0, \\ \phi_{1,j-1} + \frac{x - x_j}{x_{j+1} - x_j}(\phi_{1,j} - \phi_{1,j-1}), & x \in [x_j, x_{j+1}), \end{cases}$$

and

$$\phi_2(x) = \begin{cases} 0, & x = 0, \\ \phi_{2,j} + \frac{\phi_{2,j+1} - \phi_{2,j}}{x_j - x_{j-1}}(x - x_{j-1}), & x \in [x_{j-1}, x_j). \end{cases}$$

One can then directly verify that $\phi_1, \phi_2 \in \mathcal{K}([0, b); \mathbb{R}_{\geq 0})$ and that

$$\phi_2(x) \leq f(x) \leq \phi_2(x)$$

for all $x \in [0, b)$.

Let us now show that ϕ_1 is locally Lipschitz. Note that both ϕ_1 and ϕ_2 are piecewise linear on $(0, b)$, which means they are locally Lipschitz on $(0, b)$. In order to show that ϕ_1 can be chosen to be locally Lipschitz on $[0, b)$, we show that the slopes of the linear segments comprising ϕ_1 are bounded as we approach 0. The set of such slopes is

$$\left\{ \frac{\phi_{1,j} - \phi_{1,j-1}}{x_{j+1} - x_j} \;\middle|\; j \in \mathbb{Z} \right\},$$

and we will verify that

$$\limsup_{j\to-\infty} \frac{\phi_{1,j}-\phi_{1,j-1}}{x_{j+1}-x_j} < \infty.$$

We first note that all of these slopes are positive, as can be seen from the properties of $\phi_{1,j}$, $j \in \mathbb{Z}$. By definition of N, if $j \leq -N$,

$$\begin{aligned}\frac{\phi_{1,j}-\phi_{1,j-1}}{x_{j+1}-x_j} &= (1-\alpha_j)f(x_j), 1-(1-\alpha_{j-1})f(x_{j-1})\\ &\leq (1-\alpha_j)f(x_j) \leq (1-\alpha_N)f(x_N).\end{aligned}$$

Therefore,

$$\limsup_{j\to-\infty} \frac{\phi_{1,j}-\phi_{1,j-1}}{x_{j+1}-x_j} < \infty,$$

as claimed. Now let $x', x'' \in [0, b)$ satisfy $x' < x''$ and let $N \in \mathbb{Z}$ be such that $[x', x''] \subseteq [0, x_N)$. Letting

$$M = \sup\left\{ \frac{\phi_{1,j}-\phi_{1,j-1}}{x_{j+1}-x_j} \;\middle|\; j \leq N \right\},$$

we have

$$|\phi_1(x_1) - \phi_1(x_2)| \leq M|x_1 - x_2|,$$

which gives the desired conclusion. ■

A useful relationship between functions of class $\mathcal{K}$ and class $\mathcal{KL}$ is given by the following lemma.

10.6.4 Lemma (Solutions of differential equations with class $\mathcal{K}$ right-hand side) *Let $\phi \in \mathcal{K}([0, b); \mathbb{R}_{\geq 0})$ be locally Lipschitz. Then there exists $\psi \in \mathcal{KL}([0, b) \times \mathbb{R}_{\geq 0}; \mathbb{R}_{\geq 0})$ such that, if $x \in [0, b)$ and $t_0 \in \mathbb{R}$, then the solution to the initial value problem*

$$\dot{\xi}(t) = -\phi(\xi(t)), \quad \xi(t_0) = x,$$

is $\psi(x, t - t_0)$ for $t \geq t_0$.

Proof Using the method of Section 4.1.1, for $x \in (0, b)$ and for $t_0 \in \mathbb{R}$, the solution to the initial value problem

$$\dot{\xi}(t) = \phi(\xi(t)), \quad \xi(t_0) = x,$$

satisfies

$$\int_{t_0}^{t} d\tau = -\int_{x}^{\xi(t)} \frac{dy}{\phi(y)}.$$

To encode the dependence of this solution on the initial data, we shall denote it by $\xi_{t_0,x}$. Let us fix $x_0 \in (0, b)$ and define

$$\begin{aligned}\alpha &: [0, b) \to \mathbb{R}\\ x &\mapsto -\int_{x_0}^{x} \frac{dy}{\phi(y),}\end{aligned}$$

and note that α has the following properties.

1. *α is continuously differentiable:* This is due to the Fundamental Theorem of Calculus.
2. *α is strictly decreasing:* This is because ϕ is positive on $(0, b)$.
3. $\lim_{x\to 0} \alpha(x) = \infty$*:* Here we note that $\alpha(\xi_{0,x_0}(t)) = t$. Because ϕ is positive on $(0, b)$, it follows that $\lim_{t\to\infty} \xi_{0,x_0}(t) = 0$. Moreover, again since ϕ is positive on $(0, b)$, we cannot have $\xi_{0,x_0}(t) = 0$ for any finite t. Thus we have

$$\lim_{x\to 0} \alpha(x) = \lim_{t\to\infty} \alpha(\xi_{0,x_0}(t)) = \lim_{t\to\infty} t = \infty,$$

as asserted.

Now let $c = -\lim_{x\to b} \alpha(x)$, allowing that $c = \infty$. Thus image$(\alpha) = (-c, \infty)$ and, since α is strictly decreasing, we have a well-defined map $\alpha^{-1} \colon (-c, \infty) \to (0, b)$. Since

$$\alpha(\xi_{t_0,x}(t)) - \alpha(x) = t - t_0,$$

we have

$$\xi_{t_0,x}(t) = \alpha^{-1}(\alpha(x) + t - t_0).$$

Then define

$$\psi(x, s) = \begin{cases} \alpha^{-1}(\alpha(x) + s), & x \in (0, b), \\ 0, & x = 0. \end{cases}$$

It is clear that ψ is continuous on $(0, b) \times \mathbb{R}_{>0}$. Moreover, since

$$\lim_{(x,s)\to(0,0)} \psi(x, s) = \lim_{(x,s)\to(0,0)} \alpha^{-1}(\alpha(x) + s) = \lim_{(x,s)\to(0,0)} \xi_{0,x}(s) = 0,$$

we conclude continuity of ψ on its domain, and so we have continuity is each argument. Because $\xi_{0,x}(s) = \psi(x, s)$, we have

$$\frac{\partial \psi}{\partial s}(x, s) = \dot{\xi}_{0,x}(s) = -\phi(\xi_{0,x}(s)) = -\phi(\psi(x, s)) < 0$$

for $s \in \mathbb{R}_{>0}$, and so ψ is strictly decreasing in its second argument. It is also strictly increasing in its first argument since

$$\begin{aligned} \frac{\partial \psi}{\partial x}(x, s) &= \frac{\partial \alpha^{-1}}{\partial y}(\alpha(x) + s)\frac{\partial \alpha}{\partial y}(x) \\ &= \left(\frac{\partial \alpha}{\partial y}(\alpha^{-1}(\alpha(x) + s))\right)^{-1} \frac{\partial \alpha}{\partial y}(x) \\ &= \frac{\phi(\psi(x, s))}{\phi(x)} > 0, \end{aligned}$$

ref using the Inverse Function Theorem (). Finally,

$$\lim_{s\to\infty} \psi(x, s) = \lim_{t\to\infty} \xi(t) = 0,$$

and we have verified that $\psi \in \mathcal{KL}([0, b) \times \mathbb{R}_{\geq 0}; \mathbb{R}_{\geq 0})$ ■

10.6.2 General time-invariant functions

Now we give some definitions that, while simple, are not as simple as they seem.

10.6.5 Definition (Locally definite, locally semidefinite, decrescent I) Let $U \subseteq \mathbb{R}^n$ be an open set and let $x_0 \in U$. A function $f\colon U \to \mathbb{R}$ is:

(i) ***locally positive-definite*** about x_0 if
  (a) it is continuous,
  (b) $f(x_0) = 0$,
  (c) there exists $r \in \mathbb{R}_{>0}$ such that $f(x) \in \mathbb{R}_{>0}$ for $x \in \mathsf{B}(r, x_0) \setminus \{x_0\}$;

(ii) ***locally positive-semi definite*** about x_0 if
  (a) it is continuous,
  (b) $f(x_0) = 0$,
  (c) there exists $r \in \mathbb{R}_{>0}$ such that $f(x) \in \mathbb{R}_{\geq 0}$ for $x \in \mathsf{B}(r, x_0) \setminus \{x_0\}$;

(iii) ***locally negative-definite about*** $\mathbf{x_0}$ if $-f$ is positive-definite about x_0;

(iv) ***locally negative-semidefinite about*** $\mathbf{x_0}$ if $-f$ is positive-semidefinite about x_0;

(v) ***locally decrescent about*** $\mathbf{x_0}$ if there exists a locally positive-definite function $g\colon U \to \mathbb{R}$ around x_0 and $r \in \mathbb{R}_{>0}$ such that $f(x) \leq g(x)$ for every $x \in \mathsf{B}(r, x_0)$. •

If $f\colon U \to \mathbb{R}$ is locally positive-definite (resp. locally positive-semidefinite) about x_0 and if $r \in \mathbb{R}_{>0}$ is such that $f(x) \in \mathbb{R}_{>0}$ for $x \in \mathsf{B}(r, x_0)$, we shall say that f is ***locally positive-semidefinite about*** $\mathbf{x_0}$ ***in*** $\mathbf{B(r, x_0)}$ (resp. ***locally positive-semidefinite about*** $\mathbf{x_0}$ ***in*** $\mathbf{B(r, x_0)}$). Similar terminology applies, of course, for functions that are locally negative-definite or locally negative-semidefinite. In like manner, if f is locally decrescent about x_0, and if $r \in \mathbb{R}_{>0}$ and g, locally positive-definite about x_0 in $\mathsf{B}(r, x_0)$, are such that $f(x) \leq g(x)$ for $x \in \mathsf{B}(r, x_0)$, then we say that f is ***locally decrescent about*** $\mathbf{x_0}$ ***in*** $\mathbf{B(r, x_0)}$.

We introduce the following notation:

$\mathrm{LPD}_r(x_0)$ set of locally positive-definite functions about x_0 in $\mathsf{B}(r, x)$;
$\mathrm{LPSD}_r(x_0)$ set of locally positive-semidefinite functions about x_0 in $\mathsf{B}(r, x_0)$;
$\mathrm{LD}_r(x_0)$ set of locally decrescent functions about x_0 in $\mathsf{B}(r, x_0)$

and we also denote

$$\mathrm{LPD}(x_0) = \cup_{r \in \mathbb{R}_{>0}} \mathrm{LPD}_r(x_0), \quad \mathrm{LPSD}(x_0) = \cup_{r \in \mathbb{R}_{>0}} \mathrm{LPSD}_r(x_0), \quad \mathrm{LD}(x_0) = \cup_{r \in \mathbb{R}_{>0}} \mathrm{LD}_r(x_0).$$

The following lemma characterises some of the preceding types of functions by class $\mathcal{K}$-functions.

10.6.6 Lemma (Positive-definite and decrescent in terms of class $\mathcal{K}$ II) *For* $\mathrm{U} \subseteq \mathbb{R}^n$ *open, a continuous function* $\mathrm{f}\colon \mathrm{U} \to \mathbb{R}$, *and* $\mathrm{r} \in \mathbb{R}_{>0}$, *the following statements hold:*

(i) $\mathrm{f} \in \mathrm{LPD}_r(\mathbf{x}_0)$ *if and only if there exist* $\phi_1, \phi_2 \in \mathcal{K}([0, \mathrm{r}); \mathbb{R}_{\geq 0})$ *such that*

$$\phi_1(\|\mathbf{x} - \mathbf{x}_0\|) \leq \mathrm{f}(\mathbf{x}) \leq \phi_2(\|\mathbf{x} - \mathbf{x}_0\|)$$

for all $\mathbf{x} \in \mathsf{B}(\mathrm{r}, \mathbf{x}_0)$;

(ii) $\mathrm{f} \in \mathrm{LD}_{\mathrm{r}}(\mathbf{x}_0)$ *if and only if there exists* $\phi \in \mathcal{K}([0,\mathrm{r});\mathbb{R}_{\geq 0})$ *such that*

$$\mathrm{f}(\mathbf{x}) \leq \phi(\|\mathbf{x}-\mathbf{x}_0\|)$$

for all $\mathbf{x} \in \mathsf{B}(\mathrm{r},\mathbf{x}_0)$.

Proof (i) Suppose that $f \in \mathrm{LPD}_r(x_0)$. We first define $\psi_1 \colon [0,r) \to \mathbb{R}_{\geq 0}$ by

$$\psi_1(s) = \inf\{f(x) \mid \|x-x_0\| \in [s,r)\}.$$

We claim that (1) ψ_1 is continuous, (2) $\psi_1(0) = 0$, (3) $\psi_1(s) \in \mathbb{R}_{>0}$ for $s \in (0,r)$, (4) ψ_1 is nonincreasing, and (5) $f(x) \geq \psi_1(\|x-x_0\|)$. The only one of these that is not rather obvious is the continuity of ψ_1.

This we prove as follows. Let $s_0 \in [0,r)$ and let $\epsilon \in \mathbb{R}_{>0}$. For $x \in \mathsf{B}(r,x_0)$, let $\delta_x \in \mathbb{R}_{>0}$ be such that, if $x' \in \mathsf{B}(r,x_0)$ satisfies $\|x'-x\| < \delta_x$, then $|f(x')-f(x)| < \epsilon$. Now, by compactness of

$$S(s_0,x_0) = \{x \in \mathsf{B}(r,x_0) \mid \|x-x_0\| = s_0\},$$

let $x_1,\dots,x_k \in S(x_0,x_0)$ be such that $S(s_0,x_0) \subseteq \cup_{j=1}^k \mathsf{B}(\delta_{x_j},x_j)$. Define

$$\begin{aligned} d_{s_0} \colon S(s_0,x_0) &\to \mathbb{R}_{>0} \\ x &\mapsto \min\{\|x-x_1\|,\dots,\|x-x_k\|\}. \end{aligned}$$

ref Being a min of continuous functions, d_{s_0} is continuous (by). Being a continuous function on a compact set, there exists $\delta \in \mathbb{R}_{>0}$ such that $d_{s_0}(x) \geq \delta$ for every $x \in S(s_0,x_0)$. Now, let $s \in [0,r)$ be such that $|s-s_0| < \delta$. First suppose that $s > s_0$. Since ψ_1 is nondecreasing, $\psi_1(s) - \psi_1(s_0) \geq 0$. Now, if $x \in S(s_0,x_0)$, there exists $x' \in S(s,x_0)$ such that $|f(x')-f(x)| < \epsilon$. Thus

$$-\epsilon < f(x') - f(x) < \epsilon.$$

Since

$$\psi_1(s) \leq f(x'), \quad -\psi(s_0) \geq -f(x),$$

we have

$$\psi_1(s) - \psi(s_0) \leq f(x') - f(x) < \epsilon.$$

In like manner, if $s < s_0$, we have

$$\psi(s_0) - \psi(s) < \epsilon,$$

which gives $|\psi(s)-\psi(s_0)| < \epsilon$. This gives the asserted continuity of ψ_1.

Now, by Lemma 10.6.3, there exists $\phi_1 \in \mathcal{K}([0,r);\mathbb{R}_{\geq 0})$ such that

$$\phi_1(\|x-x_0\|) \leq \psi_1(\|x-x_0\|) \leq f(x)$$

for $x \in \mathsf{B}(r,x_0)$.

Next define $\psi_2 \colon [0,r) \to \mathbb{R}_{\geq 0}$ by

$$\psi_2(s) = \sup\{f(x) \mid \|x-x_0\| \leq s\}.$$

We can see that (1) ψ_2 is continuous, (2) $\psi_2(0) = 0$, (3) $\psi_2(s) \in \mathbb{R}_{>0}$ for $s \in (0,r)$, (4) ψ_2 is nondecreasing, and (5) $f(x) \leq \psi_2(\|x-x_0\|)$. Again, continuity is the only not completely

trivial assertion, and an argument like that above for ψ_1 can be easily made to prove this continuity assertion. Now, by Lemma 10.6.3, there exists $\phi_2 \in \mathcal{K}([0,r);\mathbb{R}_{\geq 0})$ such that

$$\phi_2(\|x - x_0\|) \geq \psi_1(\|x - x_0\|) \geq f(x)$$

for $x \in \mathsf{B}(r, x_0)$.

Next suppose that there exist $\psi_1, \phi_2 \in \mathcal{K}([0,r);\mathbb{R}_{\geq 0})$ such that

$$\phi_1(\|x - x_0\|) \leq f(x) \leq \phi_2(\|x - x_0\|)$$

for all $x \in \mathsf{B}(r, x_0)$. The left inequality immediately gives $f \in \mathrm{LPD}_r(x_0)$.

(ii) Suppose that $f \in \mathrm{LD}_r(x_0)$. Let $g \in \mathrm{LPD}_r(x_0)$ be such that $f(x) \leq g(x)$ for $x \in \mathsf{B}(r, x_0)$. By part (i) let $\phi \in \mathcal{K}([0,r);\mathbb{R}_{\geq 0})$ be such that

$$\phi(\|x - x\|) \geq g(x) \geq f(x),$$

as desired.

Finally, suppose that there exists $\phi \in \mathcal{K}([0,r);\mathbb{R}_{\geq 0})$ such that $f(x) \leq \phi(\|x - x_0\|)$ for $x \in \mathsf{B}(r, x_0)$. Since the function g defined on $\mathsf{B}(r, x_0)$ by $g(x) = \phi(\|x - x_0\|)$ is locally positive-definite about x_0 in $\mathsf{B}(r, x_0)$, the proof of the lemma is concluded. ■

10.6.3 General time-varying functions

Next we generalise the constructions of the preceding section to allow functions that depend on time.

10.6.7 Definition (Locally definite, locally semidefinite, decrescent II) Let $U \subseteq \mathbb{R}^n$ be an open set, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $x_0 \in U$. A function $f\colon \mathbb{T} \times U \to \mathbb{R}$ is:

(i) ***locally positive-definite*** about x_0 if
  (a) it is continuous,
  (b) $f(t, x_0) = 0$ for all $t \in \mathbb{T}$, and
  (c) there exist $r \in \mathbb{R}_{>0}$ and $f_0 \in \mathrm{LPD}_r(x_0)$ such that $f(t,x) \geq f_0(x)$ for every $(t,x) \in \mathbb{T} \times \mathsf{B}(r, x)$.

(ii) ***locally positive-semi definite*** about x_0 if
  (a) it is continuous,
  (b) $f(t, x_0) = 0$ for all $t \in \mathbb{T}$, and
  (c) there exist $r \in \mathbb{R}_{>0}$ and $f_0 \in \mathrm{LPSD}_r(x_0)$ such that $f(t,x) \geq f_0(x)$ for every $(t,x) \in \mathbb{T} \times \mathsf{B}(r, x)$.

(iii) ***locally negative-definite about*** $\mathbf{x_0}$ if $-f$ is positive-definite about x_0;

(iv) ***locally negative-semidefinite about*** $\mathbf{x_0}$ if $-f$ is positive-semidefinite about x_0;

(v) ***locally decrescent about*** $\mathbf{x_0}$ if there exist $r \in \mathbb{R}_{>0}$ and $g \in \mathrm{LPD}_r(x_0)$ such that $f(t,x) \leq g(x)$ for every $(t,x) \in \mathbb{T} \times \mathsf{B}(r, x_0)$. •

Let us introduce some notation for these classes of functions. As for time–invariant functions, we have all of the preceding notions of definiteness about x_0 "in $\mathsf{B}(r, x_0)$," with the obvious meaning. Let us not use all of the words required to make this obvious terminology precise. We also have the following symbols, keeping in mind that functions now are defined on $\mathbb{T} \times U$:

$\mathrm{TVLPD}_r(x_0)$ set of locally positive-definite functions about x_0 in $\mathsf{B}(r, x)$;
$\mathrm{TVLPSD}_r(x_0)$ set of locally positive-semidefinite functions about x_0 in $\mathsf{B}(r, x_0)$;
$\mathrm{TVLD}_r(x_0)$ set of locally decrescent functions about x_0 in $\mathsf{B}(r, x_0)$

and we also denote

$$\mathrm{TVLPD}(x_0) = \cup_{r \in \mathbb{R}_{>0}} \mathrm{TVLPD}_r(x_0), \quad \mathrm{TVLPSD}(x_0) = \cup_{r \in \mathbb{R}_{>0}} \mathrm{TVLPSD}_r(x_0),$$
$$\mathrm{TVLD}(x_0) = \cup_{r \in \mathbb{R}_{>0}} \mathrm{TVLD}_r(x_0).$$

An application of the definitions and of Lemma 10.6.6 gives the following lemma.

10.6.8 Lemma (Positive-definite and decrescent in terms of class $\mathcal{K}$ II) *For $\mathrm{U} \subseteq \mathbb{R}^n$ open, an interval $\mathbb{T} \subseteq \mathbb{R}$, a continuous function $\mathrm{f} \colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}$, and $\mathrm{r} \in \mathbb{R}_{>0}$, the following statements hold:*

(i) $\mathrm{f} \in \mathrm{TVLPD}_\mathrm{r}(\mathbf{x}_0)$ *if and only if there exist $\phi_1, \phi_2 \in \mathcal{K}([0, \mathrm{r}); \mathbb{R}_{\geq 0})$ such that*

$$\phi_1(\|\mathbf{x} - \mathbf{x}_0\|) \leq \mathrm{f}(\mathrm{t}, \mathbf{x}) \leq \phi_2(\|\mathbf{x} - \mathbf{x}_0\|)$$

for all $\mathrm{t} \in \mathbb{T}$ and $\mathbf{x} \in \mathsf{B}(\mathrm{r}, \mathbf{x}_0)$;

(ii) $\mathrm{f} \in \mathrm{TVLD}_\mathrm{r}(\mathbf{x}_0)$ *if and only if there exists $\phi \in \mathcal{K}([0, \mathrm{r}); \mathbb{R}_{\geq 0})$ such that*

$$\mathrm{f}(\mathrm{t}, \mathbf{x}) \leq \phi(\|\mathbf{x} - \mathbf{x}_0\|)$$

for all $\mathrm{t} \in \mathbb{T}$ and $\mathbf{x} \in \mathsf{B}(\mathrm{r}, \mathbf{x}_0)$.

10.6.9 Remark (The uniformity in time of time-varying definitions) The reader will note that, in the definition of $\mathrm{TVLPD}(x_0)$, etc., the characterisations are in terms of *time-invariant* functions from $\mathrm{LPD}(x_0)$, etc., and are required to hold for every $t \in \mathbb{T}$. One says, in this case, that the bounds required for elements of $\mathrm{TVLPD}(x_0)$, etc., hold ***uniformly*** in t. One might imagine conditions that are *not* uniform in t, but just what is required of such a definition is rather complicated. Our lack of consideration of these cases reflected in Sections 10.7.1 and 10.7.3, where we only consider Lyapunov's Second Method for characterising *uniform* stability, since nonuniform counterparts are more complicated. •

10.6.4 Time-invariant quadratic functions

When we apply Lyapunov's Second Method to linear differential equations, we will use locally positive-definite functions as in the general case. However, because

of the extra structure of linear equations, it is natural to consider locally positive-definite functions of a very particular form. In this section we shall consider the time-invariant case.

As we do when talking about linear ordinary differential equations, we shall work with equations whose state space is a finite-dimensional $\mathbb{R}$-vector space V. In such a case, the definitions of locally positive-definite, etc., are modified to account for the fact that we are principally interested in what is happening with the zero vector when talking about linear systems. The appropriate definitions require having at hand an inner product that generalises the Euclidean inner product.[8] That is, we suppose that we assign to each pair of vectors $v_1, v_2 \in \mathsf{V}$ a number $\langle v_1, v_2 \rangle \in \mathbb{R}$, and this assignment has the following properties:

1. for fixed $v_2 \in \mathsf{V}$, the function $v_1 \mapsto \langle v_1, v_2 \rangle$ is linear;
2. for fixed $v_1 \in \mathsf{V}$, the function $v_2 \mapsto \langle v_1, v_2 \rangle$ is linear;
3. $\langle v_1, v_2 \rangle = \langle v_2, v_1 \rangle$ for all $v_1, v_2 \in \mathsf{V}$;
4. $\langle v, v \rangle \in \mathbb{R}_{\geq 0}$ for all $v \in \mathsf{V}$;
5. $\langle v, v \rangle = 0$ only if $v = 0$.

We think of $\langle v_1, v_2 \rangle$ as being the "angle" between v_1 and v_2. The following are terminology and facts we shall require about inner products.

1. The assignment $v \mapsto \sqrt{\langle v, v \rangle}$ defines a norm on V that we shall simply denote by $\|\cdot\|$.
2. Given $L \in \mathrm{L}(\mathsf{V}; \mathsf{V})$, the ***transpose*** of L is the linear map $L^T \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ defined by

$$\langle L^T(v_1), v_2 \rangle = \langle v_1, L(v_2) \rangle, \qquad v_1, v_2 \in \mathsf{V}.$$

   A linear map L is ***symmetric*** if $L^T = L$.
3. If V is n-dimensional and if $L \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ is symmetric, then
   (a) its eigenvalues are real and
   (b) there is an orthonormal basis $\{e_1, \ldots, e_n\}$ of eigenvectors, i.e., (i) each of the vectors e_j, $j \in \{1, \ldots, n\}$, is an eigenvector for some eigenvalue, (ii) $\langle e_j, e_k \rangle = 0$ for $j \neq k$, and (iii) $\|e_j\| = 1$, $j \in \{1, \ldots, n\}$.

The functions of interest to us are then those prescribed by the following definition.

[8] Children call the Euclidean inner product the "dot" product, and it is defined by

$$(\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto \sum_{j=1}^{n} x_{1,j} x_{2,j}.$$

The expression on the right is often denoted $\boldsymbol{x}_1 \cdot \boldsymbol{x}_2$. However, we eschew the "$\cdot$"-notation, which is for babies, and instead write it as $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_{\mathbb{R}^n}$.

10.6.10 Definition (Quadratic function) Let V be an n-dimensional $\mathbb{R}$-vector space, let $\langle \cdot, \cdot \rangle$ be an inner product on V, and let $Q \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ be a symmetric linear map. The ***quadratic function*** associated to Q is

$$\begin{aligned} f_Q \colon \mathsf{V} &\to \mathbb{R} \\ v &\mapsto \langle Q(v), v \rangle. \end{aligned}$$

•

Now we classify various sorts of quadratic functions.

10.6.11 Definition (Locally definite, locally semidefinite, decrescent III) Let V be an n-dimensional $\mathbb{R}$-vector space, let $\langle \cdot, \cdot \rangle$ be an inner product on V, and let $Q \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ be a symmetric linear map. The linear map Q is:

(i) ***positive-definite*** if $f_Q(v) \in \mathbb{R}_{>0}$ for $v \in \mathsf{V} \setminus \{0\}$;

(ii) ***positive-semi definite*** if $f_Q(v) \in \mathbb{R}_{\geq 0}$ for $v \in \mathsf{V}$;

(iii) ***negative-definite*** if $-Q$ is positive-definite;

(iv) ***negative-semidefinite*** if $-Q$ is positive-semidefinite;

(v) ***decrescent*** if there exists a positive-definite symmetric linear map $Q_0 \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ such that $f_Q(v) \leq f_{Q_0}(v)$ for $v \in \mathsf{V}$. •

Let us relate these notions to local definiteness notions for general functions, and also to the eigenvalues of Q.

10.6.12 Lemma (Characterisations of definite, semidefinite, and decrescent symmetric linear maps) *Let* V *be an* n*-dimensional* $\mathbb{R}$*-vector space, let* $\langle \cdot, \cdot \rangle$ *be an inner product on* V*, and let* $\mathrm{Q} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ *be a symmetric linear map. Then the following statements hold.*

(i) The following statements are equivalent:

(a) Q *is positive-definite;*

(b) $\mathrm{f_Q} \in \mathrm{LPD}(0)$*;*

(c) $\mathrm{spec}(\mathrm{Q}) \subseteq \mathbb{R}_{>0}$.

(ii) The following statements are equivalent:

(a) Q *is positive-semidefinite;*

(b) $\mathrm{f_Q} \in \mathrm{LPSD}(0)$*;*

(c) $\mathrm{spec}(\mathrm{Q}) \subseteq \mathbb{R}_{\geq 0}$.

(iii) Q *is decrescent.*

Proof First of all, because Q is symmetric, all eigenvalues of Q are real and there is an orthonormal basis $\{e_1, \ldots, e_n\}$ of eigenvectors. Thus there exist $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ such that

$$Q(e_j) = \lambda_j e_j, \qquad j \in \{1, \ldots, n\}.$$

Therefore, if $v = \sum_{j=1}^{n} v_j e_v$, then

$$f_Q(v) = \left\langle Q\left(\sum_{j=1}^{n} v_j e_j\right), \sum_{k=1}^{n} v_k e_k \right\rangle = \sum_{j,k=1}^{n} v_j v_k \langle Q(e_j), e_k \rangle$$
$$= \sum_{j,k=1}^{k} \lambda_j v_j v_k \langle e_j, e_k \rangle = \sum_{j=1}^{n} \lambda_j v_j^2.$$

With this formula in hand, we prove the lemma.

(i) If Q is positive-definite, then it is clear that f_Q is locally positive-definite, from the definition.

Now, we claim that, if $\operatorname{spec}(Q) \not\subset \mathbb{R}_{>0}$, then f_Q is not locally positive definite about 0. Indeed, suppose that $\lambda_j \leq 0$ for some $j \in \{1, \dots, n\}$. Then, for any $\epsilon \in \mathbb{R}_{>0}$,

$$f_Q(\epsilon e_j) = \lambda_j \epsilon^2 \leq 0.$$

Since, for any $r \in \mathbb{R}_{>0}$, we can choose $\epsilon = \frac{r}{2} \in \mathbb{R}_{>0}$ so that $\epsilon e_j \in \mathsf{B}(r, 0) \setminus \{0\}$, it cannot be the case that f_Q is locally positive-definite.

Finally, if $\operatorname{spec}(Q) \subseteq \mathbb{R}_{>0}$, then the formula

$$f_Q(v) = \sum_{j=1}^{n} \lambda_j v_j^2$$

ensures that Q is positive-definite.

(ii) The proof follows along the lines of the first part of the proof, *mutatis mutandis*.

(iii) As in the opening paragraph of the proof, we write

$$f_Q(v) = \lambda_j v_j^2,$$

where $\lambda_1, \dots, \lambda_j$ are the eigenvalues of Q. We then let

$$C = \max\{1, \lambda_1, \dots, \lambda_n\}$$

and define $Q_0 \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ so that

$$f_{Q_0}(v) = C \sum_{j=1}^{n} v_j^2,$$

and observe that Q_0 is positive-definite (by part (i)) and that $f_Q(v) \leq f_{Q_0}(v)$ for all $v \in \mathsf{V}$. ■

The vacuous nature of the nature of decrescent symmetric linear maps (every symmetric linear map is decrescent) arises simply because this notion is not really a valuable one for time-invariant quadratic functions. We state the definition simply for the sake of preserving symmetry of the definitions.

Along these lines, the following result will be helpful to us in the next section.

10.6.13 Lemma (Upper and lower bounds for positive-definite quadratic functions) *Let* V *be an* n*-dimensional* $\mathbb{R}$*-vector space, let* $\langle\cdot,\cdot\rangle$ *be an inner product on* V*, and let* $\mathrm{Q} \in \mathrm{L}(\mathsf{V};\mathsf{V})$ *be a positive-definite symmetric linear map. Then there exists* $\mathrm{C} \in \mathbb{R}_{>0}$ *such that, for every* $\mathrm{v} \in \mathsf{V}$*, we have*

$$\mathrm{C}\langle \mathrm{v}, \mathrm{v}\rangle \leq \mathrm{f}_{\mathrm{Q}}(\mathrm{v}) \leq \mathrm{C}^{-1}\langle \mathrm{v}, \mathrm{v}\rangle.$$

Proof As in the proof of Lemma 10.6.12, for an orthonormal basis of eigenvectors $\{e_1, \dots, e_n\}$, we have

$$f_Q(v) = \sum_{j=1}^{n} \lambda_j v_j^2$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues. The result follows by taking requiring that

$$C \leq \min\{\lambda_1, \dots, \lambda_n\}$$

and

$$C^{-1} \geq \max\{\lambda_1, \dots, \lambda_n\}.$$

■

10.6.5 Time-varying quadratic functions

The final collection of functions we consider are those that are quadratic, as in the preceding section, and vary with time. A reader who has been paying attention while reading the preceding sections will likely be able to write down the definitions and characterisations we give next, as these follow quite naturally from what we have done already.

10.6.14 Definition (Time-varying quadratic function) Let V be an n-dimensional $\mathbb{R}$-vector space, let $\langle\cdot,\cdot\rangle$ be an inner product on V, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $Q\colon \mathbb{T} \to \mathrm{L}(\mathsf{V};\mathsf{V})$ be such that $Q(t)$ is a symmetric linear map for every $t \in \mathbb{T}$. The ***time-varying quadratic function*** associated to Q is

$$\begin{aligned} f_Q\colon \mathbb{T} \times \mathsf{V} &\to \mathbb{R} \\ (t, v) &\mapsto \langle Q(t)(v), v\rangle. \end{aligned}$$

•

10.6.15 Definition (Locally definite, locally semidefinite, decrescent IV) Let V be an n-dimensional $\mathbb{R}$-vector space, let $\langle\cdot,\cdot\rangle$ be an inner product on V, let $\mathbb{T} \subseteq \mathbb{R}$ be an interval, and let $Q\colon \mathbb{T} \to \mathrm{L}(\mathsf{V};\mathsf{V})$ be such that $Q(t)$ is a symmetric linear map for every $t \in \mathbb{T}$. The function Q is:

(i) ***positive-definite*** if there exists a positive-definite symmetric linear map $Q_0 \in \mathrm{L}(\mathsf{V};\mathsf{V})$ such that $f_Q(t, v) \geq f_{Q_0}(v)$ for $(t, v) \in \mathbb{T} \times \mathsf{V}$;

(ii) ***positive-semi definite*** if there exists a positive-definite symmetric linear map $Q_0 \in \mathrm{L}(\mathsf{V};\mathsf{V})$ such that $f_Q(t, v) \geq f_{Q_0}(v)$ for $(t, v) \in \mathbb{T} \times \mathsf{V}$;

(iii) ***negative-definite*** if $-Q$ is positive-definite;

(iv) ***negative-semidefinite*** if $-Q$ is positive-semidefinite.

(v) ***decrescent*** if there exists a positive-definite symmetric linear map $Q_0 \in \mathrm{L}(\mathsf{V};\mathsf{V})$ such that $f_Q(t, v) \leq f_{Q_0}(v)$ for $(t, v) \in \mathbb{T} \times \mathsf{V}$. •

10.6.16 Lemma (Characterisations of definite, semidefinite, and decrescent time-varying symmetric linear maps) *Let* V *be an* n*-dimensional* $\mathbb{R}$*-vector space, let* $\langle \cdot, \cdot \rangle$ *be an inner product on* V*, let* $\mathbb{T} \subseteq \mathbb{R}$ *be an interval, and let* $Q \colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$ *be such that* $Q(t)$ *is a symmetric linear map for every* $t \in \mathbb{T}$*. Then the following statements hold.*

(i) The following statements are equivalent:

(a) Q *is positive-definite;*

(b) $f_Q \in \mathrm{TVLPD}(0)$*;*

(c) there exists $\ell \in \mathbb{R}_{>0}$ *such that*

$$\ell \leq \inf\{\lambda \in \mathbb{R} \mid \lambda \in \operatorname{spec}(Q(t)) \text{ for some } t \in \mathbb{T}\}.$$

(ii) The following statements are equivalent:

(a) Q *is positive-semidefinite;*

(b) $f_Q \in \mathrm{TVLPSD}(0)$*;*

(c) there exists $\ell \in \mathbb{R}_{\geq 0}$ *such that*

$$\ell \leq \inf\{\lambda \in \mathbb{R} \mid \lambda \in \operatorname{spec}(Q(t)) \text{ for some } t \in \mathbb{T}\}.$$

(iii) The following statements are equivalent:

(a) Q *is decrescent;*

(b) $f_Q \in \mathrm{TVLD}(0)$*;*

(c) there exists $\mu \in \mathbb{R}_{>0}$ *such that*

$$\mu \geq \sup\{\lambda \in \mathbb{R} \mid \lambda \in \operatorname{spec}(Q(t)) \text{ for some } t \in \mathbb{T}\}.$$

Proof (i) First suppose that Q is positive-definite. By definition, by Lemma 10.6.12(i), and by Definition 10.6.7(i), $f_Q \in \mathrm{TVLPD}(0)$.

Next, suppose that

$$\inf\{\lambda \in \mathbb{R} \mid \lambda \in \operatorname{spec}(Q(t)) \text{ for some } t \in \mathbb{T}\} \leq 0.$$

For $t \in \mathbb{T}$, let $\lambda_1(t), \dots, \lambda_n(t) \subseteq \mathbb{R}$ be the eigenvalues of $Q(t)$. Without loss of generality, suppose that

$$\lambda_1(t) = \min\{\lambda_1(t), \dots, \lambda_n(t)\}, \qquad t \in \mathbb{T}.$$

For $t \in \mathbb{T}$, let $v_1(t) \in \mathsf{V}$ be an eigenvector for the eigenvalue $\lambda_1(t)$, and suppose that $\|v_1(t)\| = 1$, and note that

$$f_Q(t, v_1(t)) = \langle Q(t)(v_1(t)), v_1(t) \rangle = \lambda_1(t) \langle v_1(t), v_1(t) \rangle = \lambda_1(t).$$

By assumption $\inf\{f_Q(t, v_1(t)) \mid t \in \mathbb{T}\} \leq 0$. This means that there exists a sequence $(t_j)_{j \in \mathbb{Z}_{>0}}$ such that

$$\lim_{j \to \infty} f_Q(t_j, v_1(t_j)) \leq 0.$$

Now let $r \in \mathbb{R}_{>0}$ and $g \in \mathrm{TVLPD}_r(0)$. By Lemma 10.6.8(i), let $\phi \in \mathcal{K}([0,r);\mathbb{R}_{\geq 0})$ be such that $\phi(|x|) \leq g(x)$ for all $x \in \mathsf{B}(r,0)$. For $\epsilon \in \mathbb{R}_{>0}$ such that $\epsilon^2 < r$, we have

$$\lim_{j\to\infty} f(t_j, \epsilon^2 v_1(t_j)) = \epsilon \lim_{j\to\infty} f(t_j, v_1(t_j)) \leq 0 < \phi(\epsilon^2) \leq g(\epsilon^2 v)$$

for every $v \in \mathsf{V}$ for which $\|v\| = 1$. This means that there exists $N \in \mathbb{Z}_{>0}$ such that

$$f(t_j, \epsilon^2 v_1(t_j)) < g(\epsilon^2 v_1(t_j)), \qquad j \geq N.$$

Since g and r were arbitrary, this prohibits f from being in TVLPD(0).

Finally, suppose that

$$\inf\{\lambda \in \mathbb{R} \mid \lambda \in \mathrm{spec}(Q(t)) \text{ for some } t \in \mathbb{T}\} > 0.$$

Let $\ell \in \mathbb{R}_{>0}$ be such that

$$\inf\{\lambda \in \mathbb{R} \mid \lambda \in \mathrm{spec}(Q(t)) \text{ for some } t \in \mathbb{T}\} \geq \ell$$

and define the symmetric positive-definite linear map Q_0 so that $f_{Q_0}(v) = \ell\langle v, v\rangle$ for all $v \in \mathsf{V}$. Then, for $t \in \mathbb{T}$, let $\lambda_1(t), \dots, \lambda_n(t)$ be the eigenvalues for $Q(t)$ and let $\{e_1(t), \dots, e_n(t)\}$ be an orthonormal basis of eigenvectors. If $v \in \mathsf{V}$, write

$$v = \sum_{j=1}^{n} v_j(t) e_j(t)$$

for uniquely defined $v_1(t), \dots, v_n(t) \in \mathbb{R}$. Then, recalling the calculations from the proof of Lemma 10.6.12,

$$f_Q(t, v) = \sum_{j=1}^{n} \lambda_j(t) v_j(t)^2 \geq \ell \sum_{j=1}^{n} v_j(t)^2 = \ell\langle v, v\rangle = f_{Q_0}(v),$$

and so Q is positive-definite.

(ii) This follows, *mutatis mutandis*, as does the preceding part of the lemma.

(iii) This also follows, *mutatis mutandis*, from the proof of part (i). ■

10.6.17 Lemma (Upper and lower bounds for time-varying positive-definite and decrescent quadratic functions) *Let* V *be an* n*-dimensional* $\mathbb{R}$*-vector space, let* $\langle\cdot,\cdot\rangle$ *be an inner product on* V*, let* $\mathbb{T} \subseteq \mathbb{R}$ *be an interval, and let* $\mathrm{Q}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V};\mathsf{V})$ *be such that* $\mathrm{Q}(\mathrm{t})$ *is symmetric for every* $\mathrm{t} \in \mathbb{T}$*. Then the following statements hold:*

(i) Q *is positive-definite if and only if there exists* $\mathrm{C} \in \mathbb{R}_{>0}$ *such that* $\mathrm{C}\langle \mathrm{v}, \mathrm{v}\rangle \leq \mathrm{f}_{\mathrm{Q}}(\mathrm{t}, \mathrm{v})$ *for every* $(\mathrm{t}, \mathrm{v}) \in \mathbb{T} \times \mathsf{V}$*;*

(ii) Q *is decrescent if and only if there exists* $\mathrm{C} \in \mathbb{R}_{>0}$ *such that* $\mathrm{f}_{\mathrm{Q}}(\mathrm{t}, \mathrm{v}) \leq \mathrm{C}\langle \mathrm{v}, \mathrm{v}\rangle$ *for every* $(\mathrm{t}, \mathrm{v}) \in \mathbb{T} \times \mathsf{V}$*.*

Proof As in the proof of Lemma 10.6.16, for $t \in \mathbb{T}$ we let $\lambda_1(t), \dots, \lambda_j(t)$ be the eigenvalues for $Q(t)$ and let $\{e_1(t), \dots, e_n(t)\}$ be an orthonormal basis of eigenvectors for $Q(t)$. If we write

$$v = \sum_{j=1}^{n} v_j(t) e_j(t),$$

we then have

$$f_Q(t, v) = \sum_{j=1}^{n} \lambda_j(t) v_j(t)^2.$$

Then Q is positive-definite if and only if there exists $C \in \mathbb{R}_{>0}$ such that

$$C \leq \lambda_j(t), \qquad j \in \{1, \dots, n\},\ t \in \mathbb{T},$$

and Q is decrescent if and only if there exists $C \in \mathbb{R}_{>0}$ such that

$$\lambda_j(t) \leq C, \qquad j \in \{1, \dots, n\},\ t \in \mathbb{T}.$$

The result then follows by a simple computation, mirroring many we have already done. ■

10.6.6 Stability in terms of class $\mathcal{K}$- and class $\mathcal{KL}$-functions

In this section, whose content consists of a single lemma with its lengthy proof, we characterise various notions of stability in terms of class $\mathcal{K}$- and class $\mathcal{KL}$-functions. While it is possible to prove some of our results relating to Lyapunov's Second Method, the characterisations we give in the lemma are useful in capturing the essence of some of the proofs, and of uniting their style.

Here is the lemma of which we speak.

10.6.18 Lemma (Stability of equilibria for nonautonomous equations in terms of class $\mathcal{K}$- and class $\mathcal{KL}$-functions) *Let* $\mathbf{F}$ *be a system of ordinary differential equations with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}^n$$

with $\sup \mathbb{T} = \infty$ *and satisfying Assumption 10.2.1. For an equilibrium point* $\mathbf{x}_0 \in \mathrm{U}$ *for* $\mathbf{F}$*, the following statements hold:*

(i) $\mathbf{x}_0$ *is stable if and only if, for each* $\mathrm{t}_0 \in \mathbb{T}$*, there exist* $\delta \in \mathbb{R}_{>0}$ *and* $\alpha \in \mathcal{K}([0, \delta); \mathbb{R}_{\geq 0})$ *such that, for every* $\mathbf{x} \in \mathrm{U}$ *satisfying* $\|\mathbf{x} - \mathbf{x}_0\| < \delta$*, the solution* ξ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \widehat{\mathbf{F}}(\mathrm{t}, \xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathbf{x},$$

satisfies $\|\xi(\mathrm{t})\| \leq \alpha(\|\mathbf{x} - \mathbf{x}_0\|)$ *for* $\mathrm{t} \geq \mathrm{t}_0$*;*

(ii) $\mathbf{x}_0$ *is uniformly stable if and only if there exist* $\delta \in \mathbb{R}_{>0}$ *and* $\alpha \in \mathcal{K}([0, \delta); \mathbb{R}_{\geq 0})$ *such that, for every* $(\mathrm{t}_0, \mathbf{x}) \in \mathbb{T} \times \mathrm{U}$ *satisfying* $\|\mathbf{x} - \mathbf{x}_0\| < \delta$*, the solution* ξ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \widehat{\mathbf{F}}(\mathrm{t}, \xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathbf{x},$$

satisfies $\|\xi(\mathrm{t}) - \mathbf{x}_0\| \leq \alpha(\|\mathbf{x} - \mathbf{x}_0\|)$ *for* $\mathrm{t} \geq \mathrm{t}_0$*;*

(iii) $\mathbf{x}_0$ *is asymptotically stable if and only if, for every* $\mathrm{t}_0 \in \mathbb{T}'$*, there exist* $\delta \in \mathbb{R}_{>0}$ *and* $\beta \in \mathcal{KL}([0,\delta) \times [\mathrm{t}_0, \infty); \mathbb{R}_{\geq 0})$ *such that, if* $\mathbf{x} \in \mathrm{U}$ *satisfies* $\|\mathbf{x} - \mathbf{x}_0\| < \delta$*, then the solution* $\boldsymbol{\xi}$ *to the initial value problem*

$$\dot{\boldsymbol{\xi}}(\mathrm{t}) = \widehat{\mathbf{F}}(\mathrm{t}, \boldsymbol{\xi}(\mathrm{t})), \quad \boldsymbol{\xi}(\mathrm{t}_0) = \mathbf{x},$$

satisfies $\|\boldsymbol{\xi}(\mathrm{t}) - \boldsymbol{\xi}_0(\mathrm{t})\| \leq \beta(\|\mathbf{x} - \mathbf{x}_0\|, \mathrm{t})$ *for* $\mathrm{t} \geq \mathrm{t}_0$*;*

(iv) $\mathbf{x}_0$ *is uniformly asymptotically stable if and only if there exist* $\delta \in \mathbb{R}_{>0}$ *and* $\beta \in \mathcal{KL}([0,\delta) \times [0, \infty); \mathbb{R}_{\geq 0})$ *such that, if* $(\mathrm{t}_0, \mathbf{x}) \in \mathbb{T} \times \mathrm{U}$ *satisfies* $\|\mathbf{x} - \mathbf{x}_0\| < \delta$*, then the solution* $\boldsymbol{\xi}$ *to the initial value problem*

$$\dot{\boldsymbol{\xi}}(\mathrm{t}) = \widehat{\mathbf{F}}(\mathrm{t}, \boldsymbol{\xi}(\mathrm{t})), \quad \boldsymbol{\xi}(\mathrm{t}_0) = \mathbf{x},$$

satisfies $\|\boldsymbol{\xi}(\mathrm{t}) - \boldsymbol{\xi}_0(\mathrm{t})\| \leq \beta(\|\mathbf{x} - \mathbf{x}_0\|, \mathrm{t} - \mathrm{t}_0)$ *for* $\mathrm{t} \geq \mathrm{t}_0$.

Proof (i) First suppose that, for each $t_0 \in \mathbb{T}$, there exist $\delta \in \mathbb{R}_{>0}$ and $\alpha \in \mathcal{K}([0,\delta); \mathbb{R}_{\geq 0})$ such that, for every $\boldsymbol{x} \in U$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta$, the solution $\boldsymbol{\xi}$ to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x},$$

satisfies $\|\boldsymbol{\xi}(t)\| \leq \alpha(\|\boldsymbol{x} - \boldsymbol{x}_0\|)$ for $t \geq t_0$. Let $t_0 \in \mathbb{T}$ and let δ and α be as above. Let $\epsilon \in \mathbb{R}_{>0}$ and let $\epsilon' = \min\{\epsilon, \alpha(\delta)\}$. Let $\delta' = \min\{\delta, \alpha^{-1}(\frac{\epsilon'}{2})\}$. Let $\boldsymbol{x} \in U$ satisfy $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta' \leq \delta$ and let $\boldsymbol{\xi}$ be the solution to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x}.$$

Since

$$\|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| \leq \alpha(\|\boldsymbol{x} - \boldsymbol{x}_0\|) \leq \alpha(\delta') \leq \alpha(\alpha^{-1}(\tfrac{\epsilon'}{2})) = \tfrac{\epsilon'}{2} < \epsilon,$$

we conclude stability of $\boldsymbol{x}_0$.

Next suppose that $\boldsymbol{x}_0$ is stable and let $t_0 \in \mathbb{T}$. For $\epsilon \in \mathbb{R}_{>0}$, let $A(\epsilon) \subseteq \mathbb{R}_{>0}$ be the set of positive numbers δ such that, for every $\boldsymbol{x} \in U$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta$, the solution $\boldsymbol{\xi}$ to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x},$$

satisfies $\|\boldsymbol{\xi}(t)\| < \epsilon$ for $t \geq t_0$. Then denote $\overline{\delta}(\epsilon) = \sup A(\epsilon)$. This then defines, for
what some $\epsilon_0 \in \mathbb{R}_{>0}$, a function $\overline{\delta}\colon [0, \epsilon_0) \to \mathbb{R}_{\geq 0}$ that is nondecreasing. By there exists $\overline{\alpha} \in \mathcal{K}([0,\epsilon_0); \mathbb{R}_{\geq 0})$ such that $\overline{\alpha}(\epsilon) \leq \overline{\delta}(\epsilon)$ for every $\epsilon \in [0, \epsilon_0)$. We can suppose that ϵ_0 is sufficiently small that image$(\overline{\alpha}) = [0, \delta_0)$ for $\delta_0 \in \mathbb{R}_{>0}$. Define $\alpha = \overline{\alpha}^{-1}$, which is of class $\mathcal{K}$ by Lemma 10.6.2(i). Now, let $\boldsymbol{x} \in U$ satisfies $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta_0$ and let $\boldsymbol{\xi}$ be the solution to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x}.$$

Then let $\epsilon = \alpha(\|\boldsymbol{x} - \boldsymbol{x}_0\|)$ note that

$$\|\boldsymbol{x} - \boldsymbol{x}_0\| = \overline{\alpha}(\epsilon) \leq \overline{\delta}(\epsilon).$$

Therefore,

$$\|\xi(t) - x_0\| < \epsilon = \alpha(\|x - x_0\|),$$

completing this part of the proof.

(ii) First suppose that there exist $\delta \in \mathbb{R}_{>0}$ and $\alpha \in \mathcal{K}([0,\delta);\mathbb{R}_{\geq 0})$ such that, for every $(t_0, x) \in \mathbb{T} \times U$ satisfying $\|x - x_0\| < \delta$, the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t) - x_0\| \leq \alpha(\|x - x_0\|)$ for $t \geq t_0$. Let δ and α be as above. Let $\epsilon \in \mathbb{R}_{>0}$ and let $\epsilon' = \min\{\epsilon, \alpha(\delta)\}$. Let $\delta' = \min\{\delta, \alpha^{-1}(\frac{\epsilon'}{2})\}$. Let $(t_0, x) \in \mathbb{T} \times U$ satisfy $\|x - x_0\| < \delta' \leq \delta$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x.$$

Since

$$\|\xi(t) - x_0\| \leq \alpha(\|x - x_0\|) \leq \alpha(\delta') \leq \alpha(\alpha^{-1}(\tfrac{\epsilon'}{2})) = \tfrac{\epsilon'}{2} < \epsilon,$$

we conclude uniform stability of x_0.

Next suppose that x_0 is uniformly stable. For $\epsilon \in \mathbb{R}_{>0}$, let $A(\epsilon) \subseteq \mathbb{R}_{>0}$ be the set of positive numbers δ such that, for every $(t_0, x) \in \mathbb{T} \times U$ satisfying $\|x - x_0\| < \delta$, the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t)\| < \epsilon$ for $t \geq t_0$. Then denote $\overline{\delta}(\epsilon) = \sup A(\epsilon)$. This then defines, for some $\epsilon_0 \in \mathbb{R}_{>0}$, a function $\overline{\delta}\colon [0,\epsilon_0) \to \mathbb{R}_{\geq 0}$ that is nondecreasing. By there exists what $\overline{\alpha} \in \mathcal{K}([0,\epsilon_0);\mathbb{R}_{\geq 0})$ such that $\overline{\alpha}(\epsilon) \leq \overline{\delta}(\epsilon)$ for every $\epsilon \in [0,\epsilon_0)$. We can suppose that ϵ_0 is sufficiently small that image$(\overline{\alpha}) = [0,\delta_0)$ for $\delta_0 \in \mathbb{R}_{>0}$. Define $\alpha = \overline{\alpha}^{-1}$, which is of class $\mathcal{K}$ by Lemma 10.6.2(i). Now, let $x \in U$ satisfy $\|x - x_0\| < \delta_0$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x.$$

Then let $\epsilon = \alpha(\|x - x_0\|)$ note that

$$\|x - x_0\| = \overline{\alpha}(\epsilon) \leq \overline{\delta}(\epsilon).$$

Therefore,

$$\|\xi(t) - x_0\| < \epsilon = \alpha(x - x_0),$$

completing this part of the proof.

(iii) First suppose that, for every $t_0 \in \mathbb{T}'$, there exist $\delta \in \mathbb{R}_{>0}$ and $\beta \mathcal{KL}([0,\delta) \times [t_0,\infty);\mathbb{R}_{\geq 0})$ such that, if $x \in U$ satisfies $\|x - x_0\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t) - \xi_0(t)\| \leq \beta(\|x - x_0\|, t)$ for $t \geq t_0$. Let $t_0 \in \mathbb{T}$ and let δ and β be as above. If $x \in U$ satisfies $\|x - x_0\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies

$$\|\xi(t) - \xi_0(t)\| \leq \beta(\|x - x_0\|, t) \leq \beta(\|x - x_0\|, t_0)$$

for $t \geq t_0$. By (ii) we conclude that x_0 is stable. Also, let $\epsilon \in \mathbb{R}_{>0}$ and let $T \in \mathbb{R}_{>0}$ be sufficiently large that $\beta(\frac{\delta}{2}, t_0 + T) < \epsilon$. Then, if $(t_0, x) \in \mathbb{T} \times U$ satisfy $\|x - x_0\| < \frac{\delta}{2}$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies

$$\|\xi(t) - x_0\| \leq \beta(\tfrac{\delta}{2}, t) \leq \beta(\tfrac{\delta}{2}, t_0 + T) < \epsilon$$

for $t \geq t_0 + T$. This gives asymptotic stability of x_0.

Next suppose that x_0 is asymptotically stable. Let $t_0 \in \mathbb{T}$. Since x_0 is stable (by definition), by part (i) there exists $\delta_0 \in \mathbb{R}_{>0}$ and $\alpha \in \mathcal{K}([0, \delta_0); \mathbb{R}_{\geq 0})$ such that, for $\delta \in [0, \delta_0]$, if $x \in U$ satisfies $\|x - x_0\| < \delta$, then the solution ξ of the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies

$$\|\xi(t) - x_0\| \leq \alpha(\|x - x_0\|) < \alpha(r).$$

Now, if $(\delta, \epsilon) \in [0, \delta_0] \times \mathbb{R}_{>0}$, then let $A(\delta, \epsilon) \subseteq \mathbb{R}_{>0}$ be the set of $T \in \mathbb{R}_{>0}$ such that, if $x \in U$ satisfies $\|x - x_0\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t) = x_0\| < \epsilon$ for $t \geq t_0 + T$, this being possible by asymptotic stability. Then define $\overline{T}(\delta, \epsilon) = \inf A(\delta, \epsilon)$.

Let us record some useful properties of $\overline{T}$.

1 Lemma

(i) $\overline{\mathrm{T}}(\delta, \epsilon) \in \mathbb{R}_{\geq 0}$ *for all* $(\delta, \epsilon) \in [0, \delta_0] \times \mathbb{R}_{>0}$*;*

(ii) $\delta \mapsto \overline{\mathrm{T}}(\delta, \epsilon)$ *is nondecreasing for every* $\epsilon \in \mathbb{R}_{>0}$*, i.e.,* $\overline{\mathrm{T}}(\delta_1, \epsilon) \leq \overline{\mathrm{T}}(\delta_2, \epsilon)$ *for* $\delta_1 < \delta_2$*;*

(iii) $\epsilon \mapsto \overline{\mathrm{T}}(\delta, \epsilon)$ *is nonincreasing for every* $\delta \in [0, \delta_0]$*, i.e.,* $\overline{\mathrm{T}}(\delta, \epsilon_1) \geq \overline{\mathrm{T}}(\delta, \epsilon_2)$ *for* $\epsilon_1 < \epsilon_2$*;*

(iv) $\overline{\mathrm{T}}(\delta, \epsilon) = 0$ *if* $\epsilon > \alpha(\delta)$*.*

Proof (i) This follows since, if $T \in A(\delta, \epsilon)$, then $T \in \mathbb{R}_{\geq 0}$.

(ii) Let $\delta_1 < \delta_2$. By definition, if $T \in A(\delta_2, \epsilon)$ then it is also the case that $T \in A(\delta_1, \epsilon)$. That is, $A(\delta_2, \epsilon) \subseteq A(\delta_1, \epsilon)$ and so $\inf A(\delta_1, \epsilon) \leq \inf A(\delta_2, \epsilon)$.

(iii) Let $\epsilon_1 < \epsilon_2$. Here, if $T \in A(\delta, \epsilon_1)$ then $T \in A(\delta, \epsilon_2)$, and this gives the result.

(iv) If $\epsilon > \alpha(\delta)$, then, if $(t_0, x) \in \mathbb{T} \times U$ satisfies $\|x - x_0\| < \delta$, the solution ξ of the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies

$$\|\xi(t) - x_0\| \leq \alpha(\|x - x_0\|) < \alpha(\delta) < \epsilon$$

for all $t \geq t_0$. Thus $0 \in A(\delta, \epsilon)$. ▼

For $(\delta, \epsilon) \in [0, \delta_0] \times \mathbb{R}_{>0}$, define

$$\tau(\delta, \epsilon) = \frac{2}{\epsilon} \int_{\epsilon/2}^{\epsilon} \overline{T}(\delta, x), \mathrm{d}x + \frac{\delta}{\epsilon}.$$[9]

Let us record some properties of τ.

2 Lemma

(i) $\tau(\delta, \epsilon) \in \mathbb{R}_{>0}$ *for every* $(\delta, \epsilon) \in [0, \delta_0] \times \mathbb{R}_{>0}$;

(ii) $\epsilon \mapsto \tau(\delta, \epsilon)$ *is continuous for every* $\delta \in [0, \delta_0]$;

(iii) $\lim_{\epsilon \to \infty} \tau(\delta, \epsilon) = 0$ *for every* $\delta \in [0, \delta_0]$;

(iv) $\delta \mapsto \tau(\delta, \epsilon)$ *is strictly increasing for every* $\epsilon \in \mathbb{R}_{>0}$;

(v) $\epsilon \mapsto \tau(\delta, \epsilon)$ *is strictly decreasing for every* $\delta \in [0, \delta_0]$;

(vi) $\tau(\delta, \epsilon) \geq \overline{\mathrm{T}}(\delta, \epsilon) + \frac{\delta}{\epsilon}$.

Proof (i) This follows since $\overline{T}$ is $\mathbb{R}_{\geq 0}$-valued by Lemma 1(i).

(ii) By the Fundamental Theorem of Calculus, the function

$$\epsilon \mapsto \int_{\epsilon/2}^{\epsilon} \overline{T}(\delta, x), \mathrm{d}x$$

is continuous, and from this the continuity of τ follows.

(iii) For fixed δ, we have $\overline{T}(\delta, \epsilon) = 0$ for $\epsilon > \alpha(\delta)$ by Lemma 1(iv), and so

$$\lim_{\epsilon \to \infty} \tau(\delta, \epsilon) = \lim_{\epsilon \to \infty} \frac{\delta}{\epsilon} = 0.$$

(iv) This follows since

$$\delta \mapsto \frac{2}{\epsilon} \int_{\epsilon/2}^{\epsilon} \overline{T}(\delta, x) \, \mathrm{d}x$$

is nondecreasing by Lemma 1(ii) and since $\delta \mapsto \frac{\delta}{\epsilon}$ is strictly increasing.

(v) This follows since $\epsilon \mapsto \frac{2}{\epsilon}$ is strictly decreasing, since

$$\int_{\epsilon/2}^{\epsilon} \overline{T}(\delta, x) \, \mathrm{d}x$$

is nonincreasing by Lemma 1(iii) and since $\epsilon \mapsto \frac{\epsilon}{\delta}$ is strictly decreasing.

(vi) We have

$$\tau(\delta, \epsilon) \geq \frac{2}{\epsilon} \int_{\epsilon/2}^{\epsilon} \overline{T}(\delta, \epsilon) \, \mathrm{d}x + \frac{\delta}{\epsilon} \geq \overline{T}(\delta, \epsilon) + \frac{\delta}{\epsilon},$$

as claimed. ▼

Now, for $(\delta, s) \in [0, \delta_0] \times \mathbb{R}_{>0}$, define $\sigma(\delta, s) \in \mathbb{R}_{\geq 0}$ by asking that $\sigma(\delta, \tau(\delta, \epsilon)) = \epsilon$, i.e., $s \mapsto \sigma(\delta, s)$ is the inverse of $\epsilon \mapsto \tau(\delta, \epsilon)$. We have the following properties of σ.

[9] There is a fussy little point here about whether $\overline{T}$ is locally integrable in ϵ. This follows since $\overline{T}$ is nonincreasing, and so of "bounded variation."

3 Lemma

(i) $\delta \mapsto \sigma(\delta, s)$ *is strictly increasing for every* $s \in \mathbb{R}_{>0}$;

(ii) $s \mapsto \sigma(\delta, s)$ *is strictly decreasing for every* $\delta \in [0, \delta_0]$;

(iii) $s \mapsto \sigma(\delta, s)$ *is continuous for every* $\delta \in [0, \delta_0]$;

(iv) $\lim_{s\to\infty} \sigma(\delta, s) = 0$ *for* $\delta \in [0, \delta_0]$;

(v) $s = \tau(\delta, \sigma(\delta, s)) > \overline{T}(\delta, \sigma(\delta, s))$ *for every* $\delta \in [0, \delta_0]$.

Proof (i) and (ii) follows from parts (iv) and (v) of Lemma 2.
(iii) This follows from Lemma 2(ii).
(iv) This follows from Lemma 2(iii).
(v) This follows from Lemma 2(vi). ▼

To complete the proof, we let $\delta_0 \in \mathbb{R}_{>0}$ be as above and define

$$\begin{aligned} \beta\colon [0,\delta) \times [t_0, \infty) &\to \mathbb{R}_{\geq 0} \\ (\delta, t) &\mapsto \sqrt{\alpha(\delta)\sigma(\tfrac{\delta_0}{2}, t - t_0)}. \end{aligned}$$

The following lemma gives the essential feature of β.

4 Lemma $\beta \in \mathcal{KL}([0, \frac{\delta_0}{2}) \times [t_0, \infty); \mathbb{R}_{\geq 0})$.

Proof For fixed $t \in [t_0, \infty)$, the function

$$\delta \mapsto \sqrt{\alpha(\delta)\sigma(\tfrac{\delta_0}{2}, t - t_0)}$$

is in $\mathcal{K}([0, \frac{\delta_0}{2}); \mathbb{R}_{\geq 0})$ because:

1. $\delta \mapsto \alpha(\delta)$ is continuous and strictly increasing since $\alpha \in \mathcal{K}([0, \delta_0); \mathbb{R}_{\geq 0})$;
2. the product of strictly increasing functions is and strictly increasing;
3. $x \mapsto \sqrt{x}$ is continuous and strictly increasing on $\mathbb{R}_{\geq 0}$;
4. the composition of continuous strictly increasing functions is continuous and strictly increasing;
5. $\alpha(0) = 0$ since $\alpha \in \mathcal{K}([0, \delta_0); \mathbb{R}_{\geq 0})$.

For fixed $\delta \in [0, \frac{\delta_0}{2})$, the function

$$t \mapsto \sqrt{\alpha(\delta)\sigma(\tfrac{\delta_0}{2}, t - t_0)}$$

is in $\mathcal{L}([t_0, \infty); \mathbb{R}_{\geq 0})$ because:

1. $t \mapsto \sigma(\delta, t - t_0)$ is continuous and strictly decreasing by parts (ii) and (iii) of Lemma 3;
2. $\lim_{t\to\infty} \sigma(\delta, t - t_0) = 0$ by Lemma 3(iv). ▼

Now let $\boldsymbol{x} \in U$ satisfy $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \frac{\delta_0}{2}$ and let $\boldsymbol{\xi}$ be the solution to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x}.$$

Then

$$\|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| \leq \alpha(\|\boldsymbol{x} - \boldsymbol{x}_0\|), \qquad t \geq t_0.$$

Also, for $t > t_0$ and $\delta \in [0, \frac{\delta_0}{2}]$, if $x \in U$ satisfies $\|x - x_0\| < \delta$, and if ξ is the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

then we have

$$t - t_0 = \tau(\delta, \sigma(\delta, t - t_0)) \geq \overline{T}(\delta, \sigma(\delta, t - t_0)) + \frac{\delta}{t - t_0} > \overline{T}(\delta, \sigma(\delta, t - t_0)).$$

By definition of $\overline{T}$, this means that

$$\|\xi(t) - x_0\| \leq \sigma(\delta, t - t_0).$$

Continuity of σ in the second argument means that this relation holds, not just for $t > t_0$, but for $t \geq t_0$. Combining the inequalities

$$\|\xi(t) - x_0\| \leq \alpha(\|x - x_0\|), \quad \|\xi(t) - x_0\| \leq \sigma(\delta, t - t_0) < \sigma(\tfrac{\delta_0}{2}, t - t_0)$$

which we have shown to hold for $(t_0, x) \in \mathbb{T} \times U$ satisfying $\|x - x_0\| < \frac{\delta_0}{2}$ and for $t \geq t_0$, we have

$$\|\xi(t) - x_0\| \leq \sqrt{\alpha(\|x - x_0\|)\sigma(\tfrac{\delta_0}{2}, t - t_0)} = \beta(\|x - x_0\|, t - t_0),$$

which gives this part of the lemma.

(iv) First suppose that there exist $\delta \in \mathbb{R}_{>0}$ and $\beta \in \mathcal{KL}([0, \delta) \times [0, \infty); \mathbb{R}_{\geq 0})$ such that, if $(t_0, x) \in \mathbb{T} \times U$ satisfy $\|x - x_0\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t) - \xi_0(t)\| \leq \beta(\|x - x_0\|, t - t_0)$ for $t \geq t_0$. Let δ and β be as above. If $(t_0, x) \in \mathbb{T} \times U$ satisfies $\|x - x_0\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies

$$\|\xi(t) - \xi_0(t)\| \leq \beta(\|x - x_0\|, t - t_0) \leq \beta(\|x - x_0\|, 0)$$

for $t \geq t_0$. By (ii) we conclude that x_0 is uniformly stable. Also, let $\epsilon \in \mathbb{R}_{>0}$ and let $T \in \mathbb{R}_{>0}$ be sufficiently large that $\beta(\frac{\delta}{2}, T) < \epsilon$. Then, if $(t_0, x) \in \mathbb{T} \times U$ satisfy $\|x - x_0\| < \frac{\delta}{2}$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies

$$\|\xi(t) - x_0\| \leq \beta(\tfrac{\delta}{2}, t - t_0) \leq \beta(\tfrac{\delta}{2}, T) < \epsilon$$

for $t \geq t_0 + T$. This gives uniform asymptotic stability of x_0.

Next suppose that x_0 is uniformly asymptotically stable. Since x_0 is uniformly stable (by definition), by part (ii) there exists $\delta_0 \in \mathbb{R}_{>0}$ and $\alpha \in \mathcal{K}([0, \delta_0); \mathbb{R}_{\geq 0})$ such that, for $\delta \in [0, \delta_0]$, if $(t_0, x) \in \mathbb{T} \times U$ satisfies $\|x - x_0\| < \delta$, then the solution ξ of the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies

$$\|\xi(t) - x_0\| \leq \alpha(\|x - x_0\|) < \alpha(r).$$

Now, if $(\delta, \epsilon) \in [0, \delta_0] \times \mathbb{R}_{>0}$, then let $A(\delta, \epsilon) \subseteq \mathbb{R}_{>0}$ be the set of $T \in \mathbb{R}_{>0}$ such that, if $(t_0, x) \in \mathbb{T} \times U$ satisfies $\|x - x_0\| < \delta$, then the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies $\|\xi(t) = x_0\| < \epsilon$ for $t \geq t_0 + T$, this being possible by uniform asymptotic stability. Then define $\overline{T}(\delta, \epsilon) = \inf A(\delta, \epsilon)$.

The properties of Lemma 1 also hold for $\overline{T}$ in this case. For $(\delta, \epsilon) \in [0, \delta_0] \times \mathbb{R}_{>0}$, define

$$\tau(\delta, \epsilon) = \frac{2}{\epsilon} \int_{\epsilon/2}^{\epsilon} \overline{T}(\delta, x), \mathrm{d}x + \frac{\delta}{\epsilon}.$$

The properties of Lemma 2 also hold for τ in this case. Now, for $(\delta, s) \in [0, \delta_0] \times \mathbb{R}_{>0}$, define $\sigma(\delta, s) \in \mathbb{R}_{\geq 0}$ by asking that $\sigma(\delta, \tau(\delta, \epsilon)) = \epsilon$, i.e., $s \mapsto \sigma(\delta, s)$ is the inverse of $\epsilon \mapsto \tau(\delta, \epsilon)$. The properties of Lemma 3 also hold for σ in this case.

To complete the proof, we let $\delta_0 \in \mathbb{R}_{>0}$ be as above and define

$$\begin{aligned} \beta \colon [0, \delta) \times \mathbb{R}_{\geq 0} &\to \mathbb{R}_{\geq 0} \\ (\delta, s) &\mapsto \sqrt{\alpha(\delta)\sigma(\tfrac{\delta_0}{2}, s)}. \end{aligned}$$

The following lemma gives the essential feature of β.

5 Lemma $\beta \in \mathcal{KL}([0, \frac{\delta_0}{2}) \times \mathbb{R}_{\geq 0}; \mathbb{R}_{\geq 0})$.

Proof For fixed $s \in \mathbb{R}_{\geq 0}$, the function

$$\delta \mapsto \sqrt{\alpha(\delta)\sigma(\tfrac{\delta_0}{2}, s)}$$

is in $\mathcal{K}([0, \frac{\delta_0}{2}); \mathbb{R}_{\geq 0})$ because:

1. $\delta \mapsto \alpha(\delta)$ is continuous and strictly increasing since $\alpha \in \mathcal{K}([0, \delta_0); \mathbb{R}_{\geq 0})$;
2. the product of strictly increasing functions is and strictly increasing;
3. $x \mapsto \sqrt{x}$ is continuous and strictly increasing on $\mathbb{R}_{\geq 0}$;
4. the composition of continuous strictly increasing functions is continuous and strictly increasing;
5. $\alpha(0) = 0$ since $\alpha \in \mathcal{K}([0, \delta_0); \mathbb{R}_{\geq 0})$.

For fixed $\delta \in [0, \frac{\delta_0}{2})$, the function

$$s \mapsto \sqrt{\alpha(\delta)\sigma(\tfrac{\delta_0}{2}, s)}$$

is in $\mathcal{L}(\mathbb{R}_{\geq 0}; \mathbb{R}_{\geq 0})$ because:

1. $s \mapsto \sigma(\delta, s)$ is continuous and strictly decreasing by parts (ii) and (iii) of Lemma 3;
2. $\lim_{s \to \infty} \sigma(\delta, s) = 0$ by Lemma 3(iv). ▼

Now let $(t_0, \boldsymbol{x}) \in \mathbb{T} \times U$ satisfy $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \frac{\delta_0}{2}$ and let $\boldsymbol{\xi}$ be the solution to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x}.$$

Then

$$\|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| \leq \alpha(\|\boldsymbol{x} - \boldsymbol{x}_0\|), \qquad t \geq t_0.$$

Also, for $t > t_0$ and $\delta \in [0, \frac{\delta_0}{2}]$, if $(t_0, \boldsymbol{x}) \in \mathbb{T} \times U$ satisfies $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta$, and if $\boldsymbol{\xi}$ is the solution to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x},$$

then we have

$$t - t_0 = \tau(\delta, \sigma(\delta, t - t_0)) \geq \overline{T}(\delta, \sigma(\delta, t - t_0)) + \frac{\delta}{t - t_0} > \overline{T}(\delta, \sigma(\delta, t - t_0)).$$

By definition of $\overline{T}$, this means that

$$\|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| \leq \sigma(\delta, t - t_0).$$

Continuity of σ in the second argument means that this relation holds, not just for $t > t_0$, but for $t \geq t_0$. Combining the inequalities

$$\|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| \leq \alpha(\|\boldsymbol{x} - \boldsymbol{x}_0\|), \quad \|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| \leq \sigma(\delta, t - t_0) < \sigma(\tfrac{\delta_0}{2}, t - t_0)$$

which we have shown to hold for $(t_0, \boldsymbol{x}) \in \mathbb{T} \times U$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \frac{\delta_0}{2}$ and for $t \geq t_0$, we have

$$\|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| \leq \sqrt{\alpha(\|\boldsymbol{x} - \boldsymbol{x}_0\|)\sigma(\tfrac{\delta_0}{2}, t - t_0)} = \beta(\|\boldsymbol{x} - \boldsymbol{x}_0\|, t - t_0),$$

which gives this part of the lemma. ■

Section 10.7

Lyapunov's Second Method: Stability theorems

Much of the basic stability theory used in practice originates with the work of Aleksandr Mikhailovich Lyapunov (1857–1918). In this section and the next we shall cover what are commonly called "Lyapunov's First Method" (also "Lyapunov's Indirect Method") and "Lyapunov's Second Method" (also "Lyapunov's Direct Method"). The First Method is a useful one in that it allows one to deduce stability from the linearisation, and often the stability of the linearisation can be determined by computing a polynomial (Section 10.3.1) and performing computations with its coefficients (Section 10.4). The Second Method, on the other hand, involves hypothesising a function—called a "Lyapunov function"—with certain properties. In practice and in general, it is to be regarded as impossible to find a Lyapunov function. However, the true utility of the Second Method is that, once one has a Lyapunov function, there is a great deal one can say about the differential equation. However, such matters lie beyond the scope of the present text, and we refer to the references for further discussion.

It goes without saying that we shall discuss the Second Method first. Lyapunov's Second Method, or Direct Method, is a little. . . er. . . indirect, since it has to do with considering functions with certain properties. We shall consider in the text four settings for Lyapunov's Second Method. We shall treat each of the four cases in a self-contained manner, so a reader does not have to understand the (somewhat complicated) most general setting in order to understand the (less complicated) less general settings. Therefore, let us provide a roadmap for these cases.

10.7.1 Road map for Lyapunov's Second Method We list the four settings for Lyapunov's Second Method, and what should be read to comprehend them, together or separately.

1. *General nonautonomous equations.* The most general setting is that of equations that are nonautonomous, i.e., time-varying, and not necessarily linear. Here one needs to carefully discriminate between uniform and nonuniform stability notions. The material required to access the result on these equations is:
   (a) class $\mathcal{K}$- and class $\mathcal{KL}$-functions in Section 10.6.1;
   (b) time-invariant definite and semidefinite functions in Section 10.6.2;
   (c) time-varying definite and semidefinite functions in Section 10.6.3;
   (d) characterisations of stability using class $\mathcal{K}$- and class $\mathcal{KL}$-functions in Section 10.6.6;
   (e) the results on Lyapunov's Second Method in Section 10.7.1;
   (f) the theorems of Sections 10.7.2, 10.7.3, and 10.7.4 are corollaries of the more general theorems, although we also give independent proofs.

2. *General autonomous equations.* Here we consider autonomous ordinary differential that are not necessarily linear. The simplifications assumed by not having to discriminate between uniform and nonuniform stability make the results here significantly simpler than those for nonautonomous equations. The material needed to understand the results in this case is:
   (a) understand Definition 10.6.5;
   (b) the results on Lyapunov's Second Method in Section 10.7.2;
   (c) the theorems of Section 10.7.4 are corollaries of the more general theorems, although we also give independent proofs. •
3. *Time-varying linear equations.* The next class of equations one can consider are linear homogeneous time-varying ordinary differential equations. Note that it is necessary to understand the results on Lyapunov's Second Method here in order to prove the results on Lyapunov's First Method for nonautonomous equations. In order to understand this material, the following material needs to be read:
   (a) time-invariant quadratic functions in Section 10.6.4;
   (b) time-varying quadratic functions in Section 10.6.5;
   (c) the results on Lyapunov's Second Method in Section 10.7.3.
4. *Time-invariant linear equations.* Our final setting concerns linear homogeneous time-invariant ordinary differential equations. Note that these results are required to understand the results on Lyapunov's First Method for autonomous equations. In this setting, one needs to read the following material:
   (a) time-invariant quadratic functions in Section 10.6.4;
   (b) the result on Lyapunov's Second Method in Section 10.7.4;
   (c) the theorems of Section 10.7.4 are corollaries of the more general theorems, although we also give independent proofs. •

10.7.1 The Second Method for nonautonomous equations

Now, after that lengthy diversion concerning sort of elementary properties of functions, we come to Lyapunov's Section Method. We shall consider this method in four settings, nonautonomous/autonomous and nonlinear/linear. We begin with the most general setting, that for nonautonomous nonlinear equations.

In Lyapunov's Second Method, we will need to evaluate the derivative of a function along the solutions of an ordinary differential equation. To facilitate this, we make the following definition.

10.7.2 Definition (Lie derivative of a function along an ordinary differential equation) Let $\boldsymbol{F}$ be an ordinary differential equation with right-hand side

$$\widehat{\boldsymbol{F}} \colon \mathbb{T} \times U \to \mathbb{R}^n$$

and let $f\colon \mathbb{T}\times U\to\mathbb{R}$ be of class C^1. The ***Lie derivative*** of f along $\boldsymbol{F}$ is

$$\begin{aligned}\mathscr{L}_{\boldsymbol{F}}f\colon \mathbb{T}\times U&\to\mathbb{R}\\ (t,x)&\mapsto \frac{\partial f}{\partial t}(t,x)+\sum_{j=1}^{n}\widehat{F}_j(t,x)\frac{\partial f}{\partial x_j}(t,x).\end{aligned}$$

•

10.7.3 Lemma (Essential property of the Lie derivative I) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U}\to\mathbb{R}^n$$

and let $\mathrm{f}\colon \mathbb{T}\times\mathrm{U}\to\mathbb{R}$ *be of class* C^1. *If* $\xi\colon \mathbb{T}'\to\mathrm{U}$ *is a solution for* $\mathbf{F}$, *then*

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathrm{f}(\mathrm{t},\xi(\mathrm{t}))=\mathscr{L}_{\mathrm{F}}\mathrm{f}(\mathrm{t},\xi(\mathrm{t})).$$

Proof Using the Chain Rule and the fact that

$$\dot{\xi}(t)=\widehat{\boldsymbol{F}}(t,\xi(t)),$$

we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}f(t,\xi(t))&=\frac{\partial f}{\partial t}(t,\xi(t))+\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(t,\xi(t))\frac{\mathrm{d}\xi_j}{\mathrm{d}t}(t)\\ &=\frac{\partial f}{\partial t}(t,\xi(t))+\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(t,\xi(t))\widehat{F}_j(t,\xi(t))\\ &=\mathscr{L}_{\boldsymbol{F}}f(t,\xi(t)),\end{aligned}$$

as desired. ■

We collect our basic results on Lyapunov's Second Method in this case in the following result.

10.7.4 Theorem (Lyapunov's Second Method for nonautonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U}\to\mathbb{R}^n$$

and let $\mathbf{x}_0\in\mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$. *Assume that* $\sup\mathbb{T}=\infty$ *and that* $\mathbf{F}$ *satisfies Assumption 10.2.1. Then the following statements hold.*

(i) The equilibrium point $\mathbf{x}_0$ *is stable if there exists* $\mathrm{V}\colon \mathbb{T}\times\mathrm{U}\to\mathbb{R}$ *with the following properties:*

(a) V *is of class* C^1;

(b) $\mathrm{V}\in\mathrm{TVLPD}(\mathbf{x}_0)$;

(c) $-\mathscr{L}_{\mathbf{F}}\mathrm{V}\in\mathrm{TVLPSD}(\mathbf{x}_0)$.

(ii) *The equilibrium point* $\mathbf{x}_0$ *is uniformly stable if there exists* $\mathrm{V}\colon \mathbb{T}\times\mathrm{U}\to\mathbb{R}$ *with the following properties:*

(a) V *is of class* C^1*;*

(b) $\mathrm{V}\in\mathrm{TVLPD}(\mathbf{x}_0)$*;*

(c) $\mathrm{V}\in\mathrm{TVLD}(\mathbf{x}_0)$*;*

(d) $-\mathscr{L}_{\mathbf{F}}\mathrm{V}\in\mathrm{TVLPSD}(\mathbf{x}_0)$.

(iii) *The equilibrium point* $\mathbf{x}_0$ *is asymptotically stable if there exists* $\mathrm{V}\colon \mathbb{T}\times\mathrm{U}\to\mathbb{R}$ *with the following properties:*

(a) V *is of class* C^1*;*

(b) $\mathrm{V}\in\mathrm{TVLPD}(\mathbf{x}_0)$*;*

(c) $-\mathscr{L}_{\mathbf{F}}\mathrm{V}\in\mathrm{TVLPD}(\mathbf{x}_0)$.

(iv) *The equilibrium point* $\mathbf{x}_0$ *is uniformly asymptotically stable if there exists* $\mathrm{V}\colon \mathbb{T}\times\mathrm{U}\to\mathbb{R}$ *with the following properties:*

(a) V *is of class* C^1*;*

(b) $\mathrm{V}\in\mathrm{TVLPD}(\mathbf{x}_0)$*;*

(c) $\mathrm{V}\in\mathrm{TVLD}(\mathbf{x}_0)$*;*

(d) $-\mathscr{L}_{\mathbf{F}}\mathrm{V}\in\mathrm{TVLPD}(\mathbf{x}_0)$.

Proof (i) Let $t_0\in\mathbb{T}$. Let $r\in\mathbb{R}_{>0}$ be such that

1. $\overline{\mathsf{B}}(2r,x_0)\subseteq U$,
2. $V\in\mathrm{TVLPD}_{2r}(x_0)$, and
3. $-\mathscr{L}_{\boldsymbol{F}}V\in\mathrm{TVLPSD}_{2r}(x_0)$.

By definition of time-varying locally positive, let $f\in\mathrm{LPD}_{2r}(x_0)$ be such that

$$f(\boldsymbol{x})\le V(t,\boldsymbol{x}) \tag{10.23}$$

for all $(t,\boldsymbol{x})\in\mathbb{T}\times\overline{\mathsf{B}}(r,x_0)$. Also let $g\in\mathrm{LPSD}_r(x_0)$ be such that

$$\mathscr{L}_{\boldsymbol{F}}V(t,\boldsymbol{x})\le -g(\boldsymbol{x})\le 0$$

for $(t,\boldsymbol{x})\in\mathbb{T}\times\overline{\mathsf{B}}(r,x_0)$. Let $c\in\mathbb{R}_{>0}$ be such that

$$c<\inf\{f(\boldsymbol{x})\mid \|\boldsymbol{x}-x_0\|=r\}$$

and then define

$$f^{-1}(\le c)=\{\boldsymbol{x}\in\overline{\mathsf{B}}(r,x_0)\mid f(\boldsymbol{x})\le c\}.$$

Also, for $t\in\mathbb{T}$, denote

$$V_t^{-1}(\le c)=\{\boldsymbol{x}\in\overline{\mathsf{B}}(r,x_0)\mid V(t,\boldsymbol{x})\le c\}.$$

By (10.23), we have

$$V_t^{-1}(\le c)\subseteq f^{-1}(\le c)\subseteq\mathsf{B}(r,x_0),\qquad t\in\mathbb{T}.$$

Define $\alpha_2 \colon [0, 2r] \to \mathbb{R}$ by

$$\beta(s) = \sup\{V(t_0, \boldsymbol{x}) \mid \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq s\}.$$

A reference to the proof of Lemma 10.6.6(i) gives $\alpha_2 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ such that and

$$V(t_0, \boldsymbol{x}) \leq \beta(\|\boldsymbol{x} - \boldsymbol{x}_0\|) \leq \alpha_2(\|\boldsymbol{x} - \boldsymbol{x}_0\|), \qquad \boldsymbol{x} \in \overline{\mathsf{B}}(r, \boldsymbol{x}_0).$$

Note that $\lim_{s \to 0} \alpha_2(s) = 0$, and so there exists $\delta \in \mathbb{R}_{>0}$ such that $\alpha_2(s) < c$ for $s \in [0, \delta]$. Note that

$$\boldsymbol{x} \in \mathsf{B}(\delta, \boldsymbol{x}_0) \implies V(t_0, \boldsymbol{x}) \leq c.$$

Let $\boldsymbol{x} \in \mathsf{B}(\delta, \boldsymbol{x}_0)$ and let $\boldsymbol{\xi}$ be the solution to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}(t, \boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(t_0) = \boldsymbol{x}. \tag{10.24}$$

The following technical lemmata are required to proceed with the proof, and will recur a number of times for proofs relating to Lyapunov's Second Method.

1 Lemma *The solution $\boldsymbol{\xi}$ satisfies $\boldsymbol{\xi}(\mathrm{t}) \in \overline{\mathsf{B}}(\mathrm{r}, \mathbf{x}_0)$ for $\mathrm{t} \geq \mathrm{t}_0$.*

Proof Suppose this is not true. Then, by continuity of $\boldsymbol{\xi}$, there exists a largest $T \in \mathbb{R}_{>0}$ such that $\boldsymbol{\xi}(t) \in \overline{\mathsf{B}}(r, \boldsymbol{x}_0)$ for all $t \in [t_0, t_0 + T]$. This implies, by continuity of $t \mapsto V(t, \boldsymbol{\xi}(t))$, that

$$\|\boldsymbol{\xi}(T) - \boldsymbol{x}_0\| = r. \tag{10.25}$$

Using the facts that

$$\boldsymbol{x} \in \mathsf{B}(\delta, \boldsymbol{x}_0) \subseteq V_{t_0}^{-1}(\leq c) \subseteq \mathsf{B}(r, \boldsymbol{x}_0),$$

and that

$$\frac{\mathrm{d}}{\mathrm{d}t} V(t, \boldsymbol{\xi}(t)) = \mathscr{L}_{\boldsymbol{F}} V(t, \boldsymbol{\xi}(t)) \leq 0, \qquad t \in [t_0, t_0 + T]$$

(the leftmost equality by Lemma 10.7.3), we have

$$\begin{aligned} V(T, \boldsymbol{\xi}(T)) &= V(t_0, \boldsymbol{\xi}(t_0)) + \int_{t_0}^{T} V(t, \boldsymbol{\xi}(t)) \, \mathrm{d}t \\ &= V(t_0, \boldsymbol{\xi}(t_0)) + \int_{t_0}^{T} \mathscr{L}_{\boldsymbol{F}} V(t, \boldsymbol{\xi}(t)) \, \mathrm{d}t < c. \end{aligned} \tag{10.26}$$

However, this contradicts (10.25) and the definition of c, and so we conclude the lemma. ▼

The next lemma we state in some generality, since it asserts a generally useful fact.

2 Lemma *Let* $\mathbf{F}$ *be an ordinary differential equation whose right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}^{\mathrm{n}}$$

satisfies $\sup \mathbb{T} = \infty$ *and Assumption 10.2.1. Let* $\mathrm{K} \subseteq \mathrm{U}$ *be compact and assume that, for every* $(\mathrm{t}_0, \mathbf{x}) \in \mathbb{T} \times \mathrm{K}$, *the solution* ξ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \widehat{\mathbf{F}}(\mathrm{t}, \xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathbf{x}_0,$$

satisfies $\xi(\mathrm{t}) \in \mathrm{K}$ *for* $\mathrm{t} \geq \mathrm{t}_0$.

Then, for every $(\mathrm{t}_0, \mathbf{x}) \in \mathbb{T} \times \mathrm{K}$, *the solution* ξ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \widehat{\mathbf{F}}(\mathrm{t}, \xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathbf{x}_0,$$

is defined on $[\mathrm{t}_0, \infty)$.

Proof Suppose the hypotheses of the lemma hold, but the conclusions do not. Thus there exists $(t_0, \boldsymbol{x}) \in \mathbb{T} \times K$ for which the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = \boldsymbol{x}_0, \tag{10.27}$$

in not defined for all $t \in [t_0, \infty)$. Then there exists a largest $T \in \mathbb{R}_{>0}$ such that the solution of the initial value problem is defined on $[t_0, t_0 + T)$. Let $(t_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $[t_0, t_0 + T)$ converging to $t_0 + T$. By the Bolzano–Weierstrass Theorem, the sequence $(\xi(t_j))_{j \in \mathbb{Z}_{>0}}$ has a convergent subsequence $(\xi(t_{j_k}))_{k \in \mathbb{Z}_{>0}}$:

$$\lim_{k \to \infty} \xi(t_{j_k}) = \boldsymbol{y} \in K.$$

Now, by Theorem 3.2.8(ii), there exists $\epsilon \in \mathbb{R}_{>0}$ such that the solution η to the initial value problem

$$\dot{\eta}(t) = \widehat{\boldsymbol{F}}(t, \eta(t)), \quad \eta(t_0 + T) = \boldsymbol{y},$$

is defined on $t \in [t_0 + T - \epsilon, t_0 + T + \epsilon]$. Moreover, by assumption, $\eta(t) \in K$ for every $t \in [t_0 + T - \epsilon, t_0 + T + \epsilon]$. Define $\overline{\xi}\colon [t_0, t_0 + T + \epsilon] \to K$ by

$$\overline{\xi}(t) = \begin{cases} \xi(t), & t \in [t_0, t_0 + T), \\ \eta(t), & t \in [t_0 + T, t_0 + T + \epsilon]. \end{cases}$$

Note, then, that $\overline{\xi}$ is a solution to the differential equation and satisfies the initial condition $\overline{\xi}(t_0) = \boldsymbol{x}$. Thus we have arrived at a contradiction to the solution to the initial value problem (10.27) being defined only on $[t_0, t_0 + T)$. ▼

By combining the preceding two lemmata, we conclude that the solution ξ to the initial value problem (10.24) with $\boldsymbol{x} \in \mathsf{B}(\delta, \boldsymbol{x}_0)$ satisfies (1) $\xi(t) \in \overline{\mathsf{B}}(r, \boldsymbol{x}_0)$ for all $t \geq t_0$ and (2) it is defined on $[t_0, \infty)$. Moreover, by the computation (10.26),

$$\xi(t) \in V_t^{-1}(\leq c) \subseteq f^{-1}(\leq c).$$

By Lemma 10.6.6, there exists $\alpha_1 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ such that

$$\alpha_1(\|\boldsymbol{x} - \boldsymbol{x}_0\|) \leq f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \overline{\mathsf{B}}(r, \boldsymbol{x}_0).$$

Let $x \in \mathsf{B}(\delta, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x.$$

Since $x \in \mathsf{B}(\delta, x_0)$, our arguments above imply that ξ is defined on $[t_0, \infty)$ and that $\xi(t) \in \overline{\mathsf{B}}(r, x_0)$ for $t \geq t_0$. Moreover,

$$\alpha_1(\|\xi(t) - x_0\|) \leq f_1(\xi(t)) \leq V(t, \xi(t)) \leq V(t_0, \xi(t_0)) \leq \alpha_2(\|\xi(t_0) - x_0\|)$$

for $t \geq t_0$. Thus

$$\|\xi(t) - x_0\| \leq \alpha_1^{-1} \circ \alpha_2(\|\xi(t_0) - x_0\|)$$

for $t \geq t_0$. Since $\alpha_1^{-1} \circ \alpha_2 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ by Lemma 10.6.2, we can now conclude uniform stability from Lemma 10.6.18(ii).

(ii) Let $r \in \mathbb{R}_{>0}$ be such that

1. $\overline{\mathsf{B}}(2r, x_0) \subseteq U$,
2. $V \in \mathrm{TVLPD}_{2r}(x_0)$,
3. $V \in \mathrm{TVLD}_{2r}(x_0)$, and
4. $-\mathscr{L}_F V \in \mathrm{TVLPSD}_{2r}(x_0)$.

By definition of time-varying locally positive and locally decrescent, let $f_1, f_2 \in \mathrm{LPD}_{2r}(x_0)$ be such that

$$f_1(x) \leq V(t, x) \leq f_2(x) \tag{10.28}$$

for all $(t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0)$. Also let $g \in \mathrm{LPSD}_r(x_0)$ be such that

$$\mathscr{L}_F V(t, x) \leq -g(x) \leq 0$$

for $(t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0)$. Let $c \in \mathbb{R}_{>0}$ be such that

$$c < \inf\{f_1(x) \mid \|x - x_0\| = r\}$$

and then define

$$f_1^{-1}(\leq c) = \{x \in \overline{\mathsf{B}}(r, x_0) \mid f_1(x) \leq c\}$$

and

$$f_2^{-1}(\leq c) = \{x \in \overline{\mathsf{B}}(r, x_0) \mid f_2(x) \leq c\}.$$

Also, for $t \in \mathbb{T}$, denote

$$V_t^{-1}(\leq c) = \{x \in \overline{\mathsf{B}}(r, x_0) \mid V(t, x) \leq c\}.$$

By (10.28), we have

$$f_2^{-1}(\leq c) \subseteq V_t^{-1}(\leq c) \subseteq f_1^{-1}(\leq c) \subseteq \mathsf{B}(r, x_0), \qquad t \in \mathbb{T}.$$

Let $x \in f_2^{-1}(\leq c)$, let $t_0 \in \mathbb{T}$, and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x. \tag{10.29}$$

The following lemma is an adaptation of Lemma 1 to our current setting.

3 Lemma *The solution ξ satisfies $\xi(t) \in \overline{B}(r, x_0)$ for $t \geq t_0$.*

Proof Suppose this is not true. Then, by continuity of ξ, there exists a largest $T \in \mathbb{R}_{>0}$ such that $\xi(t) \in \overline{B}(r, x_0)$ for all $t \in [t_0, t_0 + T]$. This implies, by continuity of $t \mapsto V(t, \xi(t))$, that

$$\|V(T, \xi(T)) - x_0\| = r. \tag{10.30}$$

Using the facts that

$$x \in f_2^{-1}(t_0, \leq c) \subseteq V_{t_0}^{-1}(\leq c) \subseteq B(r, x_0),$$

and that

$$\frac{\mathrm{d}}{\mathrm{d}t} V(t, \xi(t)) = \mathscr{L}_F V(t, \xi(t)) \leq 0, \qquad t \in [t_0, t_0 + T]$$

(the leftmost equality by Lemma 10.7.3), we have

$$\begin{aligned} V(T, \xi(T)) &= V(t_0, \xi(t_0)) + \int_{t_0}^{T} V(t, \xi(t)) \, \mathrm{d}t \\ &= V(t_0, \xi(t_0)) + \int_{t_0}^{T} \mathscr{L}_F V(t, \xi(t)) \, \mathrm{d}t < c. \end{aligned} \tag{10.31}$$

However, this contradicts (10.30) and the definition of c, and so we conclude the lemma. ▼

By combining the preceding lemma with Lemma 2, we conclude that the solution ξ to the initial value problem (10.29) with $x \in f_2^{-1}(\leq c)$ satisfies (1) $\xi(t) \in \overline{B}(r, x_0)$ for all $t \geq t_0$ and (2) it is defined on $[t_0, \infty)$. Moreover, by the computation (10.31),

$$\xi(t) \in V_t^{-1}(\leq c) \subseteq f_1^{-1}(\leq c).$$

By Lemma 10.6.6, there exist $\alpha_1, \alpha_2 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ such that

$$\alpha_1(\|x - x_0\|) \leq f_1(x), \quad f_2(x) \leq \alpha_2(\|x - x_0\|), \qquad x \in \overline{B}(r, x_0).$$

Now let $\delta \in (0, r]$ be sufficiently small that $\alpha_2(s) \leq c$ for $s \in [0, \delta]$. Note that

$$x \in B(\delta, x_0) \implies \alpha_2(\|x - x_0\|) \leq c \implies x \in f_2^{-1}(\leq c).$$

Let $(t_0, x) \in \mathbb{T} \times B(\delta, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x.$$

Since $x \in f_2^{-1}(\leq c)$, our arguments above imply that ξ is defined on $[t_0, \infty)$ and that $\xi(t) \in \overline{B}(r, x_0)$ for $t \geq t_0$. Moreover,

$$\alpha_1(\|\xi(t) - x_0\|) \leq f_1(\xi(t)) \leq V(t, \xi(t)) \leq V(t_0, \xi(t_0)) \leq f_2(\xi(t_0)) \leq \alpha_2(\|\xi(t_0) - x_0\|)$$

for $t \geq t_0$. Thus

$$\|\xi(t) - x_0\| \leq \alpha_1^{-1} \circ \alpha_2(\|\xi(t_0) - x_0\|)$$

for $t \geq t_0$. Since $\alpha_1^{-1} \circ \alpha_2 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ by Lemma 10.6.2, we can now conclude uniform stability from Lemma 10.6.18(ii).

(iii) Let $t_0 \in \mathbb{T}$. Let $r \in \mathbb{R}_{>0}$ be such that

1. $\overline{\mathsf{B}}(2r, x_0) \subseteq U$,
2. $V \in \mathrm{TVLPD}_{2r}(x_0)$,
3. $-\mathscr{L}_F V \in \mathrm{TVLPD}_{2r}(x_0)$.

As in the proof of part (i), we let $f_1 \in \mathrm{LPD}_{2r}(x_0)$ and $\alpha_1 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ be such that

$$\alpha_1(\|x - x_0\|) \leq f_1(x) \leq V(t, x)$$

for $(t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0)$. Also as in the proof of part (i), let $\alpha_2 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ be such that

$$V(t_0, x) \leq \alpha_2(\|x - x_0\|), \qquad x \in \overline{\mathsf{B}}(r, x_0).$$

Also let $f_3 \in \mathrm{LPD}_{2r}(x_0)$ and $\alpha_3 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ be such that

$$\alpha_3(\|x - x_0\|) \leq f_3(x) \leq -\mathscr{L}_F V(t, x)$$

for $(t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0)$.

Of course, we then conclude stability of x_0 from part (i). We then have

$$V(t, x) \leq \alpha_2(\|x - x_0\|) \implies \alpha_3 \circ \alpha_2^{-1} \circ V(t, x) \leq \alpha_3(\|x - x_0\|) \tag{10.32}$$

for $(t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0)$. By Lemma 10.6.2 we have $\alpha_3 \circ \alpha_2^{-1} \in \mathcal{K}([0, r); \mathbb{R}_{\geq 0})$ and, therefore, by Lemma 10.6.3, there exists a locally Lipschitz $\alpha \in \mathcal{K}([0, r); \mathbb{R}_{\geq 0})$ such that $\alpha(s) \leq \alpha_3 \circ \alpha_2^{-1}(x)$ for all $x \in [0, r)$. Now let δ be as in the proof of part (i). Let $x \in \overline{\mathsf{B}}(\delta, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x.$$

Recall that

1. $\xi(t) \in \overline{\mathsf{B}}_r(x_0,)$ for all $t \in [t_0, \infty)$ by Lemma 1,
2. $V(t, \xi(t)) \leq c$ for all $t \in [t_0, \infty)$ by definition of δ.

Using Lemma 10.7.3 and (10.32), we then have

$$\frac{\mathrm{d}}{\mathrm{d}t} V(t, \xi(t)) = \mathscr{L}_F V(t, \xi(t)) \leq -\alpha_3(\|\xi(t) - x_0\|) \leq -\alpha \circ V(t, \xi(t)).$$

The following technical lemma is now required.

4 Lemma *Let* F *be a scalar ordinary differential equation with right-hand side*

$$\widehat{\mathrm{F}} \colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}$$

where $\mathrm{U} \subseteq \mathbb{R}$ *is open. For* $(\mathrm{t}_0, \mathrm{y}_0) \in \mathbb{T} \times \mathrm{U}$*, let* $\xi, \eta \colon \mathbb{T}' \to \mathrm{U}$ *be of class* C^1 *and satisfy*

$$\dot{\xi}(\mathrm{t}) = \widehat{\mathrm{F}}(\mathrm{t}, \xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathrm{y}_0$$

and

$$\dot{\eta}(\mathrm{t}) < \widehat{\mathrm{F}}(\mathrm{t}, \eta(\mathrm{t})), \quad \eta(\mathrm{t}_0) = \mathrm{y}_0.$$

Then $\eta(\mathrm{t}) < \xi(\mathrm{t})$ *for* $\mathrm{t} > \mathrm{t}_0$.

Proof We have

$$\dot\eta(t_0) < \widehat{F}(t_0, y_0) = \dot\xi(t_0).$$

Therefore, by continuity of the derivatives, there exists $\epsilon \in \mathbb{R}_{>0}$ such that

$$\dot\eta(t) < \dot\xi(t), \qquad t \in [t_0, t_0 + \epsilon].$$

Therefore, for $t \in (t_0, t_0 + \epsilon]$,

$$\eta(t) = \int_{t_0}^{t} \dot\eta(\tau)\,\mathrm{d}\tau < \int_{t_0}^{t} \dot\xi(\tau)\,\mathrm{d}\tau = \xi(t).$$

Now suppose that it does not hold that $\eta(t) < \xi(t)$ for all $t \geq t_0$. Then let

$$T = \inf\{t \geq t_0 \mid \eta(t) \geq \xi(t)\} > t_0 + \epsilon.$$

By continuity, $\eta(T) = \xi(T)$. Thus

$$\begin{aligned}\dot\eta(T) &= \underbrace{\dot\eta(T) - \widehat{F}(T, \eta(T))}_{<0} + \widehat{F}(T, \eta(T)) \\ &< \underbrace{\dot\xi(t) - \widehat{F}(T, \xi(T))}_{=0} + \widehat{F}(T, \xi(T)) = \dot\xi(T).\end{aligned}$$

On the other hand, for $h \in \mathbb{R}_{>0}$ (sufficiently small for the expression to be defined) we have

$$\frac{\eta(T) - \eta(T-h)}{h} > \frac{\xi(T) - \xi(T-h)}{h},$$

and taking the limit as $h \to 0$ gives $\dot\eta(T) \geq \dot\xi(T)$, contradicting our computation just proceeding. ▼

By Lemma 10.6.4, there exists $\psi \in \mathcal{KL}([0, r) \times [t_0, \infty); \mathbb{R}_{\geq 0})$ such that, if $y \in [0, r)$, then the solution to the initial value problem

$$\dot\eta(t) = -\alpha(\eta(t)), \quad \eta(t_0) = y,$$

is $\psi(y, t)$ for $t \geq t_0$. By Lemma 4 we have

$$V(t, \boldsymbol{\xi}(t)) \leq \psi(V(t_0, \boldsymbol{x}), t), \qquad t \geq t_0.$$

Therefore,

$$\begin{aligned}\|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| &\leq \alpha_1^{-1} \circ \psi(V(t_0, \boldsymbol{x}), t) \\ &\leq \alpha_1^{-1} \circ \psi(\alpha_2(\|\boldsymbol{x} - \boldsymbol{x}_0\|), t).\end{aligned}$$

By Lemma 10.6.2(iii), the mapping

$$\begin{aligned}\beta\colon [0, r) \times [t_0, \infty) &\to \mathbb{R} \\ (s, \tau) &\mapsto \alpha_1^{-1} \circ \psi(\alpha_2(s), \tau)\end{aligned}$$

is of class $\mathcal{KL}$. The asymptotic stability of $\boldsymbol{x}_0$ now follows from Lemma 10.6.18(iii).

(iv) Let $r \in \mathbb{R}_{>0}$ be such that

1. $\overline{\mathsf{B}}(2r, x_0) \subseteq U$,
2. $V \in \text{TVLPD}_{2r}(x_0)$,
3. $V \in \text{TVLD}_{2r}(x_0)$, and
4. $-\mathscr{L}_F V \in \text{TVLPD}_{2r}(x_0)$.

As in the proof of part (ii), we let $f_1, f_2 \in \text{LPD}_{2r}(x_0)$ and $\alpha_1, \alpha_2 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ be such that

$$\alpha_1(\|x - x_0\|) \leq f_1(x) \leq V(t, x) \leq f_2(x) \leq \alpha_2(\|x - x_0\|)$$

for $(t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0)$. Also let $f_3 \in \text{LPD}_{2r}(x_0)$ and $\alpha_3 \in \mathcal{K}([0, 2r); \mathbb{R}_{\geq 0})$ be such that

$$\alpha_3(\|x - x_0\|) \leq f_3(x) \leq -\mathscr{L}_F V(t, x)$$

for $(t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0)$.

Of course, we then conclude uniform stability of x_0 from part (ii). We then have

$$V(t, x) \leq \alpha_2(\|x - x_0\|) \implies \alpha_3 \circ \alpha_2^{-1} \circ V(t, x) \leq \alpha_3(\|x - x_0\|) \tag{10.33}$$

for $(t, x) \in \mathbb{T} \times \overline{\mathsf{B}}(r, x_0)$. By Lemma 10.6.2 we have $\alpha_3 \circ \alpha_2^{-1} \in \mathcal{K}([0, r); \mathbb{R}_{\geq 0})$ and, therefore, by Lemma 10.6.3, there exists a locally Lipschitz $\alpha \in \mathcal{K}([0, r); \mathbb{R}_{\geq 0})$ such that $\alpha(s) \leq \alpha_3 \circ \alpha_2^{-1}(x)$ for all $x \in [0, r)$. Now let δ be as in the proof of part (ii). Let $(t_0, x) \in \mathbb{T} \times \overline{\mathsf{B}}(\delta, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}(t, \xi(t)), \quad \xi(t_0) = x.$$

Recall that

1. $\xi(t) \in \overline{\mathsf{B}}_r(x_0,)$ for all $t \in [t_0, \infty)$ by Lemma 3,
2. $V(t, \xi(t)) \leq c$ for all $t \in [t_0, \infty)$ by definition of δ.

Using Lemma 10.7.3 and (10.33), we then have

$$\frac{\mathrm{d}}{\mathrm{d}t} V(t, \xi(t)) = \mathscr{L}_F V(t, \xi(t)) \leq -\alpha_3(\|\xi(t) - x_0\|) \leq -\alpha \circ V(t, \xi(t)).$$

By Lemma 10.6.4, there exists $\psi \in \mathcal{KL}([0, r) \times \mathbb{R}_{\geq 0}; \mathbb{R}_{\geq 0})$ such that, if $y \in [0, r)$ and $t_0 \in \mathbb{R}$, then the solution to the initial value problem

$$\dot{\eta}(t) = -\alpha(\eta(t)), \quad \eta(t_0) = y,$$

is $\psi(y, t - t_0)$ for $t \geq t_0$. By Lemma 4 we have

$$V(t, \xi(t)) \leq \psi(V(t_0, x), t - t_0), \qquad t \geq t_0.$$

Therefore,

$$\begin{aligned} \|\xi(t) - x_0\| &\leq \alpha_1^{-1} \circ \psi(V(t_0, x), t - t_0) \\ &\leq \alpha_1^{-1} \circ \psi(\alpha_2(\|x - x_0\|), t - t_0). \end{aligned}$$

By Lemma 10.6.2(iii), the mapping

$$\begin{aligned} \beta \colon [0, r) \times \mathbb{R}_{\geq 0} &\to \mathbb{R} \\ (s, \tau) &\mapsto \alpha_1^{-1} \circ \psi(\alpha_2(s), \tau) \end{aligned}$$

is of class $\mathcal{KL}$. The uniform asymptotic stability of x_0 now follows from Lemma 10.6.18(iv). ■

10.7.5 Terminology The function V in the statement of the preceding theorem is typically called a ***Lyapunov function***. It is not uncommon for this terminology to be used imprecisely, in the sense that when one sees the expression "Lyapunov function," it is clear only from context whether one is in case (i), (ii), (iii), or (iv) of the preceding theorem. Typically this is not to be thought of as confusing, as the context indeed makes this clear. •

We also have the following sufficient condition for exponential stability (as opposed to mere asymptotic stability) which comes with the flavour of Lyapunov's Second Method.

10.7.6 Theorem (Lyapunov's Second Method for exponential stability of nonautonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathsf{U}\to \mathbb{R}^n$$

and let $\mathbf{x}_0\in \mathsf{U}$ *be an equilibrium point for* $\mathbf{F}$*. Assume that* $\sup\mathbb{T}=\infty$ *and* $\mathbf{F}$ *satisfies Assumption 10.2.1. Then* $\mathbf{x}_0$ *is uniformly exponentially stable if there exists* $\mathsf{V}\colon \mathbb{T}\times\mathsf{U}\to\mathbb{R}$ *with the following properties:*

(i) V *is of class* C^1*;*

(ii) there exists $\mathsf{C}_1,\alpha_1,\mathsf{r}_1\in\mathbb{R}_{>0}$ *such that*

$$\mathsf{C}_1\|\mathbf{x}-\mathbf{x}_0\|^{\alpha_1}\leq \mathsf{V}(\mathsf{t},\mathbf{x})\leq \mathsf{C}_1^{-1}\|\mathbf{x}-\mathbf{x}_0\|^{\alpha_1}$$

for all $(\mathsf{t},\mathbf{x})\in\mathbb{T}\times\mathsf{B}(\mathsf{r}_1,\mathbf{x}_0)$*;*

(iii) there exists $\mathsf{C}_2,\alpha_2,\mathsf{r}_2\in\mathbb{R}_{>0}$ *such that*

$$\mathscr{L}_{\mathbf{F}}\mathsf{V}(\mathsf{t},\mathbf{x})\leq -\mathsf{C}_2\|\mathbf{x}-\mathbf{x}_0\|^{\alpha_2}$$

for all $(\mathsf{t},\mathbf{x})\in\mathbb{T}\times\mathsf{B}(\mathsf{r}_2,\mathbf{x}_0)$.

Proof Let $r,\alpha\in\mathbb{R}_{>0}$ be such that

1. $C_1\|x-x_0\|^\alpha\leq V(t,x)\leq C_1^{-1}\|x-x_0\|^\alpha$ for all $(t,x)\in\mathbb{T}\times\mathsf{B}(2r,x_0)$ and
2. $-\mathscr{L}_F V(t,x)\geq C_2\|x-x_0\|^\alpha$ for all $(t,x)\in\mathbb{T}\times\mathsf{B}(r,x_0)$.

Let $c\in\mathbb{R}_{>0}$ be such that

$$c<\inf\{C_1\|x-x_0\|^\alpha\mid \|x-x_0\|=r\}.$$

We then let $\delta\in\mathbb{R}_{>0}$ be such that, if $x\in\mathsf{B}(\delta,x_0)$, then $C_2\|x-x_0\|\leq c$. Let $(t_0,x)\in\mathbb{T}\times\mathsf{B}(\delta,x_0)$ and let ξ be the solution to the initial value problem

$$\dot\xi(t)=\widehat{F}(t,\xi(t)),\quad \xi(t_0=x.$$

We then argue as in Lemmata 3 and 2 from the proof of Theorem 10.7.4 that $\xi(t)\in\overline{\mathsf{B}}(r,x_0)$ for $t\geq t_0$ and that ξ is defined for all $t\in[t_0,\infty)$. Now compute, using Lemma 10.7.3 and the definitions of C_1, C_2, and α,

$$\frac{\mathrm{d}}{\mathrm{d}t}V(t,\xi(t))=\mathscr{L}_F V(t,\xi(t))\leq -C_2\|\xi(t)-x_0\|^\alpha\leq -C_1C_2V(t,\xi(t)).$$

By Lemma 4 of Theorem 10.7.4,

$$V(t, \xi(t)) \leq V(t_0, \xi(t_0))\mathrm{e}^{-C_1C_2(t-t_0)}$$

for $t \geq t_0$. Now, again using the definition of C_1 and α,

$$\begin{aligned}\|\xi(t) - x_0\| &\leq \left(\frac{V(t, \xi(t))}{C_1}\right)^{1/\alpha} \leq \left(\frac{V(t_0, \xi(t_0))\mathrm{e}^{-C_1C_2(t-t_0)}}{C_1}\right)^{1/\alpha} \\ &\leq \frac{\|x - x_0\|}{C_1^{2\alpha}}\mathrm{e}^{-C_1C_2(t-t_0)/\alpha}\end{aligned}$$

for all $t \geq t_0$. Recalling that the preceding estimates are valid for any $(t_0, x) \in \mathbb{T}\times\mathsf{B}(\delta, x_0)$, we conclude uniform exponential stability of x_0. ■

10.7.2 The Second Method for autonomous equations

In the preceding section we gave a quite general version of Lyapunov's Second Method applied to nonautonomous ordinary differential equations. As can be seen, the proofs are lengthy and a little detailed. Here we consider the simpler autonomous case, for which we give a self-contained proof for a reader wishing for a "light" alternative. In stating the result in this case, we recall from Proposition 10.2.5 that "stability" and "uniform stability" are equivalent, and that "asymptotic stability" and "uniform asymptotic stability" are equivalent for nonautonomous ordinary differential equations.

Before we get to the statement of the main result, we first give the non-time-varying version of the definition of Lie derivative.

10.7.7 Definition (Lie derivative of a function along an autonomous ordinary differential equation) Let F be an autonomous ordinary differential equation with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{T}\times U &\to \mathbb{R}^n \\ (t, x) &\mapsto \widehat{F}_0(x),\end{aligned}$$

and let $f\colon U \to \mathbb{R}$ be of class C^1. The ***Lie derivative*** of f along F is

$$\begin{aligned}\mathscr{L}_{F_0}f\colon U &\to \mathbb{R} \\ x &\mapsto \sum_{j=1}^{n}\widehat{F}_{0,j}(x)\frac{\partial f}{\partial x_j}(x).\end{aligned}$$

•

10.7.8 Lemma (Essential property of the Lie derivative II) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U} &\to \mathbb{R}^{\mathrm{n}} \\ (\mathrm{t}, \mathbf{x}) &\mapsto \widehat{\mathbf{F}}_0(\mathrm{t}, \mathbf{x}),\end{aligned}$$

and let $f\colon U \to \mathbb{R}$ *be of class* C^1. *If* $\xi\colon \mathbb{T}' \to U$ *is a solution for* $\mathbf{F}$, *then*

$$\frac{d}{dt} f(\xi(t)) = \mathscr{L}_{F_0} f(\xi(t)).$$

Proof Using the Chain Rule and the fact that

$$\dot{\xi}(t) = \widehat{F}_0(\xi(t)),$$

we have

$$\begin{aligned}\frac{d}{dt} f(\xi(t)) &= \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\xi(t)) \frac{d\xi_j}{dt}(t) \\ &= \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\xi(t)) \widehat{F}_{0,j}(\xi(t)) \\ &= \mathscr{L}_{F_0} f(\xi(t)),\end{aligned}$$

as desired. ∎

We can now state the main concerning Lyapunov's Second Method in the nonautonomous case.

10.7.9 Theorem (Lyapunov's Second Method for autonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{\mathbf{F}}\colon \mathbb{T} \times U &\to \mathbb{R}^n \\ (t, \mathbf{x}) &\mapsto \widehat{\mathbf{F}}_0(\mathbf{x}),\end{aligned}$$

and let $\mathbf{x}_0 \in U$ *be an equilibrium point for* $\mathbf{F}$. *Assume that* $\sup \mathbb{T} = \infty$ *and that* $\mathbf{F}$ *satisfies Assumption 10.2.1. Then the following statements hold.*

(i) The equilibrium point $\mathbf{x}_0$ *is stable if there exists* $V\colon U \to \mathbb{R}$ *with the following properties:*

(a) V *is of class* C^1*;*

(b) $V \in \mathrm{LPD}(\mathbf{x}_0)$*;*

(c) $-\mathscr{L}_{\mathbf{F}_0} V \in \mathrm{LPSD}(\mathbf{x}_0)$.

(ii) The equilibrium point $\mathbf{x}_0$ *is asymptotically stable if there exists* $V\colon U \to \mathbb{R}$ *with the following properties:*

(a) V *is of class* C^1*;*

(b) $V \in \mathrm{LPD}(\mathbf{x}_0)$*;*

(c) $-\mathscr{L}_{\mathbf{F}_0} V \in \mathrm{LPD}(\mathbf{x}_0)$.

We shall give two proofs of Theorem 10.7.9, one assuming Theorem 10.7.4 and one independent of that more general theorem.

Proof of Theorem 10.7.9, assuming Theorem 10.7.4 In this case, the theorem is an easy corollary of the more general Theorem 10.7.4. Indeed, the hypotheses of parts (i) and (ii) of Theorem 10.7.9 immediately imply those of parts (ii) and (iv), respectively, of Theorem 10.7.4. ■

Independent proof of Theorem 10.7.9 (i) Let $\epsilon \in \mathbb{R}_{>0}$. Let $r \in (0, \frac{\epsilon}{2}]$ be chosen so that

1. $\overline{\mathsf{B}}(2r, x_0) \subseteq U$,
2. $V \in \mathrm{LPD}_{2r}(x_0)$, and
3. $-\mathscr{L}_{F_0} V \in \mathrm{LPSD}_{2r}(x_0)$.

Let $c \in \mathbb{R}_{>0}$ be such that

$$c < \inf\{V(x) \mid \|x - x_0\| = r\}$$

and define

$$V^{-1}(\leq c) = \{x \in \overline{\mathsf{B}}(r, x_0) \mid V(x) \leq c\}.$$

Then $V^{-1}(\leq c) \subseteq \mathsf{B}(r, x_0)$ by continuity of V and the definition of c. By continuity of V, let $\delta \in \mathbb{R}_{>0}$ be such that, if $x \in \mathsf{B}(\delta, x_0)$, then $V(x) < c$. Therefore, we have

$$\mathsf{B}(\delta, x_0) \subseteq V^{-1}(\leq c) \subseteq \mathsf{B}(r, x_0).$$

Let $(t_0, x) \in \mathbb{T} \times \mathsf{B}(\delta, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}_0(\xi(t)), \quad \xi(t_0) = x.$$

The following lemmata, which essentially appear in the proof of Theorem 10.7.4, are repeated here for the purposes of making the proof self-contained.

1 Lemma *The solution ξ satisfies $\xi(t) \in \overline{\mathsf{B}}(r, x_0)$ for $t \geq t_0$.*

Proof Suppose this is not true. Then, by continuity of ξ, there exists a largest $T \in \mathbb{R}_{>0}$ such that $\xi(t) \in \overline{\mathsf{B}}(r, x_0)$ for all $t \in [t_0, t_0 + T]$. This implies, by continuity of $t \mapsto V(\xi(t))$, that

$$\|V(\xi(T)) - x_0\| = r. \tag{10.34}$$

Using the facts that

$$x \in \mathsf{B}(\delta, x_0) \subseteq V^{-1}(\leq c) \subseteq \mathsf{B}(r, x_0),$$

and that

$$\frac{\mathrm{d}}{\mathrm{d}t} V(\xi(t)) = \mathscr{L}_{F_0} V(\xi(t)) \leq 0, \qquad t \in [t_0, t_0 + T]$$

(the leftmost equality by Lemma 10.7.8), we have

$$\begin{aligned} V(\xi(T)) &= V(\xi(t_0)) + \int_{t_0}^{T} V(\xi(t))\,\mathrm{d}t \\ &= V(\xi(t_0)) + \int_{t_0}^{T} \mathscr{L}_{F_0} V(\xi(t))\,\mathrm{d}t < c. \end{aligned} \tag{10.35}$$

However, this contradicts (10.34) and the definition of c, and so we conclude the lemma. ▼

2 Lemma *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation whose right-hand side*

$$\begin{aligned}\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U} &\to \mathbb{R}^{\mathrm{n}}\\ (\mathrm{t},\mathbf{x}) &\mapsto \widehat{\mathbf{F}}_0(\mathbf{x})\end{aligned}$$

satisfies $\sup\mathbb{T}=\infty$ *and Assumption 10.2.1. Let* $\mathrm{K}\subseteq\mathrm{U}$ *be compact and assume that, for every* $(\mathrm{t}_0,\mathbf{x})\in\mathbb{T}\times\mathrm{K}$*, the solution* ξ *to the initial value problem*

$$\dot{\xi}(\mathrm{t})=\widehat{\mathbf{F}}_0(\xi(\mathrm{t})),\quad \xi(\mathrm{t}_0)=\mathbf{x}_0,$$

satisfies $\xi(\mathrm{t})\in\mathrm{K}$ *for* $\mathrm{t}\geq\mathrm{t}_0$.

Then, for every $(\mathrm{t}_0,\mathbf{x})\in\mathbb{T}\times\mathrm{K}$*, the solution* ξ *to the initial value problem*

$$\dot{\xi}(\mathrm{t})=\widehat{\mathbf{F}}_0(\xi(\mathrm{t})),\quad \xi(\mathrm{t}_0)=\mathbf{x}_0,$$

is defined on $[\mathrm{t}_0,\infty)$.

Proof Suppose the hypotheses of the lemma hold, but the conclusions do not. Thus there exists $(t_0,\boldsymbol{x})\in\mathbb{T}\times K$ for which the solution ξ to the initial value problem

$$\dot{\xi}(t)=\widehat{\boldsymbol{F}}_0(\xi(t)),\quad \xi(t_0)=\boldsymbol{x}_0, \tag{10.36}$$

in not defined for all $t\in[t_0,\infty)$. Then there exists a largest $T\in\mathbb{R}_{>0}$ such that the solution of the initial value problem is defined on $[t_0,t_0+T)$. Let $(t_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $[t_0,t_0+T)$ converging to t_0+T. By the Bolzano–Weierstrass Theorem, the sequence $(\xi(t_j))_{j\in\mathbb{Z}_{>0}}$ has a convergent subsequence $(\xi(t_{j_k}))_{k\in\mathbb{Z}_{>0}}$:

$$\lim_{k\to\infty}\xi(t_{j_k})=\boldsymbol{y}\in K.$$

Now, by Theorem 3.2.8(ii), there exists $\epsilon\in\mathbb{R}_{>0}$ such that the solution $\boldsymbol{\eta}$ to the initial value problem

$$\dot{\boldsymbol{\eta}}(t)=\widehat{\boldsymbol{F}}_0(\boldsymbol{\eta}(t)),\quad \boldsymbol{\eta}(t_0+T)=\boldsymbol{y},$$

is defined on $t\in[t_0+T-\epsilon,t_0+T+\epsilon]$. Moreover, by assumption, $\boldsymbol{\eta}(t)\in K$ for every $t\in[t_0+T-\epsilon,t_0+T+\epsilon]$. Define $\overline{\xi}\colon[t_0,t_0+T+\epsilon]\to K$ by

$$\overline{\xi}(t)=\begin{cases}\xi(t), & t\in[t_0,t_0+T),\\ \boldsymbol{\eta}(t), & t\in[t_0+T,t_0+T+\epsilon].\end{cases}$$

Note, then, that $\overline{\xi}$ is a solution to the differential equation and satisfies the initial condition $\overline{\xi}(t_0)=\boldsymbol{x}$. Thus we have arrived at a contradiction to the solution to the initial value problem (10.36) being defined only on $[t_0,t_0+T)$. ▼

Since $r\leq\frac{\epsilon}{2}<\epsilon$, the preceding lemma immediately proves stability of $\boldsymbol{x}_0$.

(ii) Let $r,\delta\in\mathbb{R}_{>0}$ be chosen so that

1. $\overline{\mathsf{B}}(2r,\boldsymbol{x}_0)\subseteq U$,
2. $V\in\mathrm{LPD}_{2r}(\boldsymbol{x}_0)$,
3. $-\mathscr{L}_{\mathbf{F}_0}V\in\mathrm{LPD}(\boldsymbol{x}_0)$, and

4. if $(t_0, x) \in \mathbb{T} \times \mathsf{B}(\delta, x_0)$ and if ξ is the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}_0(\xi(t)), \quad \xi(t_0) = x,$$

then $\xi(t) \in \overline{\mathsf{B}}(r, x_0)$ for $t \geq t_0$ and ξ is defined on $[t_0, \infty)$.

The last condition is possible by virtue of our arguments in part (i).

Let $(t_0, x) \in \mathbb{T} \times \mathsf{B}(\delta, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}_0(\xi(t)), \quad \xi(t_0) = x.$$

Since $\frac{\mathrm{d}}{\mathrm{d}t} V(\xi(t)) < 0$ for all $t \geq t_0$, it follows that $t \mapsto V(\xi(t))$ is strictly decreasing. Thus, since V is nonnegative, there exists $\gamma \in \mathbb{R}_{\geq 0}$ such that

$$\lim_{t \to \infty} V(\xi(t)) = \gamma.$$

We claim that $\gamma = 0$. Suppose otherwise, and that $\gamma \in \mathbb{R}_{>0}$. Let $\alpha \in \mathbb{R}_{>0}$ be such that, if $x \in \mathsf{B}(\alpha, x_0)$, then $V(x) < \gamma$. Therefore, $\xi(t) \in \overline{\mathsf{B}}(r, x_0) \setminus \mathsf{B}(\alpha, x_0)$. Denote

$$\beta = \inf\{-\mathscr{L}_{F_0} V(x) \mid \|x - x_0\| \in [\alpha, r]\},$$

ref the infimum existing because it is over a compact set by . Moreover, since $\mathscr{L}_{F_0} V$ is negative definite, $\beta \in \mathbb{R}_{>0}$. Now we calculate

$$\begin{aligned} V(\xi(t)) &= V(\xi(t_0)) + \int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}\tau} V(\xi(\tau))\, \mathrm{d}\tau \\ &= V(\xi(t_0)) + \int_{t_0}^{t} \mathscr{L}_{F_0} V(\xi(\tau))\, \mathrm{d}\tau \\ &\leq V(\xi(t_0)) - \beta(t - t_0). \end{aligned}$$

This implies that $\lim_{t \to \infty} V(\xi(t)) = -\infty$. This contradiction leads us to conclude that $\gamma = 0$.

Finally, we must show that this implies that

$$\lim_{t \to \infty} \|\xi(t) - x_0\| = 0$$

(still supposing ξ to be the solution for initial condition $(t_0, x) \in \mathbb{T} \times \mathsf{B}(\delta, x_0)$). To this end, let $\epsilon \in \mathbb{R}_{>0}$ and let $b \in \mathbb{R}_{>0}$ be such that

$$b < \inf\{V(x) \mid \|x - x_0\| = \epsilon\}.$$

Then, as we argued above that $V^{-1}(\leq c) \subseteq \mathsf{B}(r, x_0)$, here we conclude that $V^{-1}(\leq b) \subseteq \mathsf{B}(\epsilon, x_0)$. Therefore, if we let $T \in \mathbb{R}_{>0}$ be sufficiently large that $V(\xi(t)) \leq b$ for $t \geq T$, then $\xi(t) \in \mathsf{B}(\epsilon, x_0)$ for all $t \geq T$. ■

10.7.10 Terminology The function V in the statement of the preceding theorem is typically called a ***Lyapunov function***. It is not uncommon for this terminology to be used imprecisely, in the sense that when one sees the expression "Lyapunov function," it is clear only from context whether one is in case (i) or (ii) of the preceding theorem. Typically this is not to be thought of as confusing, as the context indeed makes this clear. •

10.7.11 Remark (Automatic implications of Theorem 10.7.9) We recall from Proposition 10.2.5 that uniform stability and stability are equivalent for autonomous ordinary differential equations, and similarly that uniform asymptotic stability and asymptotic stability are equivalent. •

10.7.12 Theorem (Lyapunov's Second Method for exponential stability of autonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U} &\to \mathbb{R}^n\\ (t,\mathbf{x}) &\mapsto \widehat{\mathbf{F}}_0(\mathbf{x}),\end{aligned}$$

and let $\mathbf{x}_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$*. Assume that* $\sup\mathbb{T} = \infty$ *and* $\mathbf{F}$ *satisfies Assumption 10.2.1. Then* $\mathbf{x}_0$ *is exponentially stable if there exists* $\mathrm{V}\colon \mathrm{U}\to\mathbb{R}$ *with the following properties:*

(i) V *is of class* C^1*;*

(ii) there exist $\mathrm{C}_1, \alpha_1, \mathrm{r}_1 \in \mathbb{R}_{>0}$ *such that*

$$\mathrm{C}_1\|\mathbf{x}-\mathbf{x}_0\|^{\alpha_1} \le \mathrm{V}(\mathbf{x}) \le \mathrm{C}_1^{-1}\|\mathbf{x}-\mathbf{x}_0\|^{\alpha_1}$$

for all $\mathbf{x} \in \mathsf{B}(\mathrm{r}_1, \mathbf{x}_0)$*;*

(iii) there exist $\mathrm{C}_2, \alpha_2, \mathrm{r}_2 \in \mathbb{R}_{>0}$ *such that*

$$\mathscr{L}_{\mathbf{F}_0}\mathrm{V}(\mathbf{x}) \le -\mathrm{C}_2\|\mathbf{x}-\mathbf{x}_0\|^{\alpha_2}$$

for all $\mathbf{x} \in \mathsf{B}(\mathrm{r}_2, \mathbf{x}_0)$.

Proof Let $r, \alpha \in \mathbb{R}_{>0}$ be such that

1. $C_1\|x-x_0\|^\alpha \le V(x) \le C_1^{-1}\|x-x_0\|^\alpha$ for all $x \in \mathsf{B}(2r, x_0)$ and
2. $-\mathscr{L}_{F_0}V(x) \ge C_2\|x-x_0\|^\alpha$ for all $x \in \mathsf{B}(r, x_0)$.

Let $c \in \mathbb{R}_{>0}$ be such that

$$c < \inf\{C_1\|x-x_0\|^\alpha \mid \|x-x_0\| = r\}.$$

We then let $\delta \in \mathbb{R}_{>0}$ be such that, if $x \in \mathsf{B}(\delta, x_0)$, then $C_2\|x-x_0\| \le c$. Let $(t_0, x) \in \mathbb{T}\times\mathsf{B}(\delta, x_0)$ and let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}_0(\xi(t)), \quad \xi(t_0) = x.$$

We then argue as in Lemmata 1 and 2 from the proof of Theorem 10.7.9 that $\xi(t) \in \overline{\mathsf{B}}(r, x_0)$ for $t \ge t_0$ and that ξ is defined for all $t \in [t_0, \infty)$. Now compute, using Lemma 10.7.8 and the definitions of C_1, C_2, and α,

$$\frac{\mathrm{d}}{\mathrm{d}t}V(\xi(t)) = \mathscr{L}_F V(\xi(t)) \le -C_2\|\xi(t)-x_0\|^\alpha \le -C_1C_2V(\xi(t)).$$

The following technical lemma is now required.

1 Lemma *Let* F *be an autonomous scalar ordinary differential equation with right-hand side*

$$\widehat{\mathrm{F}}\colon \mathbb{T}\times \mathrm{U} \to \mathbb{R}$$
$$(\mathrm{t}, \mathrm{x}) \mapsto \widehat{\mathrm{F}}_0(\mathrm{x}),$$

where $\mathrm{U} \subseteq \mathbb{R}$ *is open. For* $(\mathrm{t}_0, \mathrm{y}_0) \in \mathbb{T}\times \mathrm{U}$*, let* $\xi, \eta\colon \mathbb{T}' \to \mathrm{U}$ *be of class* C^1 *and satisfy*

$$\dot{\xi}(\mathrm{t}) = \widehat{\mathrm{F}}_0(\xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathrm{y}_0$$

and

$$\dot{\eta}'(\mathrm{t}) < \widehat{\mathrm{F}}_0(\boldsymbol{\eta}(\mathrm{t})), \quad \eta(\mathrm{t}_0) = \mathrm{y}_0.$$

Then $\eta(\mathrm{t}) < \xi(\mathrm{t})$ *for* $\mathrm{t} > \mathrm{t}_0$.

Proof We have

$$\dot{\eta}(t_0) < \widehat{F}(y_0) = \dot{\xi}(t_0).$$

Therefore, by continuity of the derivatives, there exists $\epsilon \in \mathbb{R}_{>0}$ such that

$$\dot{\eta}(t) < \dot{\xi}(t), \qquad t \in [t_0, t_0 + \epsilon].$$

Therefore, for $t \in (t_0, t_0 + \epsilon]$,

$$\eta(t) = \int_{t_0}^{t} \dot{\eta}(\tau)\,\mathrm{d}\tau < \int_{t_0}^{t} \dot{\xi}(\tau)\,\mathrm{d}\tau = \xi(t).$$

Now suppose that it does not hold that $\eta(t) < \xi(t)$ for all $t \geq t_0$. Then let

$$T = \inf\{t \geq t_0 \mid \eta(t) \geq \xi(t)\} > t_0 + \epsilon.$$

By continuity, $\eta(T) = \xi(T)$. Thus

$$\begin{aligned}\eta'(T) &= \underbrace{\eta'(T) - \widehat{F}(\eta(T))}_{<0} + \widehat{F}(T, \eta(T)) \\ &< \underbrace{\xi'(t) - \widehat{F}(\xi(T))}_{=0} + \widehat{F}(\xi(T)) = \xi'(T).\end{aligned}$$

On the other hand, for $h \in \mathbb{R}_{>0}$ (sufficiently small for the expression to be defined) we have

$$\frac{\eta(T) - \eta(T-h)}{h} > \frac{\xi(T) - \xi(T-h)}{h},$$

and taking the limit as $h \to 0$ gives $\eta'(T) \geq \xi'(T)$, contradicting our computation just proceeding. ▼

By the lemma,

$$V(\boldsymbol{\xi}(t)) \leq V(\boldsymbol{\xi}(t_0))\mathrm{e}^{-C_1C_2(t-t_0)}$$

for $t \geq t_0$. Now, again using the definition of C_1 and α,

$$\begin{aligned}\|\boldsymbol{\xi}(t) - \boldsymbol{x}_0\| &\leq \left(\frac{V(\boldsymbol{\xi}(t))}{C_1}\right)^{1/\alpha} \leq \left(\frac{V(\boldsymbol{\xi}(t_0))\mathrm{e}^{-C_1C_2(t-t_0)}}{C_1}\right)^{1/\alpha} \\ &\leq \frac{\|\boldsymbol{x} - \boldsymbol{x}_0\|}{C_1^{2\alpha}}\mathrm{e}^{-C_1C_2(t-t_0)/\alpha}\end{aligned}$$

for all $t \geq t_0$. Recalling that the preceding estimates are valid for any $(t_0, \boldsymbol{x}) \in \mathbb{T}\times\mathsf{B}(\delta, \boldsymbol{x}_0)$, we conclude exponential stability of $\boldsymbol{x}_0$. ■

10.7.3 The Second Method for time-varying linear equations

The next two sections will be concerned with Lyapunov's Second Method for systems of linear homogeneous ordinary differential equations. In this section we treat the time-varying case, and in the next we treat the time-invariant case. While it is relatively easy to prove the theorems in this case using the general, not for linear equations, results of Sections 10.7.1 and 10.7.2, we instead give self-contained proofs that illustrate the special character of stability for linear differential equations that we studied in Section 10.3.

In the study of Lyapunov's Second Method for linear equations, one works with Lyapunov functions that are especially adapted to the linear structure of the equations, namely the quadratic functions of Sections 10.6.4 and 10.6.5. In working with such functions, the derivatives along solutions, called the "Lie derivative" in Definitions 10.7.2 and 10.7.7, take a particular form that leads to the following definition and associated following result.

10.7.13 Definition (Lyapunov pair for time-varying linear ordinary differential equations) Let F be a system of linear homogeneous ordinary differential equations in an n-dimensional $\mathbb{R}$-vector space V and with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{T}\times\mathsf{V} &\to \mathsf{V}\\ (t,x) &\mapsto A(t)(x)\end{aligned}$$

for $A\colon \mathbb{T}\to \mathrm{L}(\mathsf{V};\mathsf{V})$. Suppose that V has an inner product $\langle\cdot,\cdot\rangle$. A ***Lyapunov pair*** for F is a pair (P,Q) where

(i) $P,Q\colon \mathbb{T}\to \mathrm{L}(\mathsf{V};\mathsf{V})$ are such that P is of class C^1, Q is continuous, and $P(t)$ and $Q(t)$ are symmetric, and

(ii) $\dot{P}(t)+P(t)\circ A(t)+A^T(t)\circ P(t)=-Q(t)$ for all $t\in\mathbb{T}$. •

Note that, with the notion of a Lyapunov pair, one can think of (1) P as being given, and part (ii) of the definition prescribing Q or (2) Q as being given, in which case part (ii) prescribing a linear differential equation for P. Both ways of thinking about this will be useful.

At first encounter, such a definition seems to come from nowhere. However, the motivation for it is straightforward, as the following lemma shows, and its proof makes clear.

10.7.14 Lemma (Derivative of quadratic function along solutions of a linear ordinary differential equation) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *and with right-hand side*

$$\begin{aligned}\widehat{\mathrm{F}}\colon \mathbb{T}\times\mathsf{V} &\to \mathsf{V}\\ (\mathrm{t},\mathrm{x}) &\mapsto \mathrm{A}(\mathrm{t})(\mathrm{x})\end{aligned}$$

for $A\colon \mathbb{T} \to L(V;V)$. *Suppose that* V *has an inner product* $\langle\cdot,\cdot\rangle$. *Let* $P\colon \mathbb{T} \to L(V;V)$ *be of class* C^1 *and such that* $P(t)$ *is symmetric for every* $t \in \mathbb{T}$ *and let* f_P *be the corresponding time-varying quadratic function as in Definition 10.6.14. Then, for any solution* $\xi\colon \mathbb{T} \to V$ *for* F, *we have*

$$\frac{d}{dt} f_P(t, \xi(t)) = -f_Q(t, \xi(t)),$$

where (P, Q) *is a Lyapunov pair for* F.

Proof We shall represent solutions using the state transition map as in Section 5.2.1.2. Thus, if $(t_0, x) \in \mathbb{T} \times V$, the solution to the initial value problem

$$\dot{\xi}(t) = A(t)(\xi(t)), \quad \xi(t_0) = x,$$

is $\xi(t) = \Phi_A(t, t_0)(x)$. Now we directly compute

$$\begin{aligned}
\frac{d}{dt} f_P(t, \Phi_A(t, t_0)(x)) &= \frac{d}{dt}\langle P(t) \circ \Phi_A(t, t_0)(x), \Phi_A(t, t_0)(x)\rangle \\
&= \langle \dot{P}(t) \circ \Phi_A(t, t_0)(x), \Phi_A(t, t_0)(x)\rangle + \langle P(t) \circ \tfrac{d}{dt}\Phi_A(t, t_0)(x), \Phi_A(t, t_0)(x)\rangle \\
&\quad + \langle P(t) \circ \Phi_A(t, x_0)(x), \tfrac{d}{dt}\Phi_A(t, t_0)(x)\rangle \\
&= \langle \dot{P}(t) \circ \Phi_A(t, t_0)(x), \Phi_A(t, t_0)(x)\rangle + \langle P(t) \circ A(t) \circ \Phi_A(t, t_0)(x), \Phi_A(t, t_0)(x)\rangle \\
&\quad + \langle P(t) \circ \Phi_A(t, x_0)(x)), A(t) \circ \Phi_A(t, t_0)(x)\rangle \\
&= -\langle Q(t) \circ \Phi_A(t, t_0)(x), \Phi_A(t, t_0)(x)\rangle,
\end{aligned}$$

where (P, Q) is a Lyapunov pair, i.e.,

$$Q(t) = -\dot{P}(t) - P(t) \circ A(t) - A^T(t) \circ P(t), \qquad t \in \mathbb{T}.$$

■

The lemma allows us to provide the following connection to the Lie derivative characterisations of Lemmata 10.7.3 and 10.7.8.

10.7.15 Corollary (Lie derivative of quadratic function along a linear ordinary differential equation) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *and with right-hand side*

$$\begin{aligned}
\widehat{F}\colon \mathbb{T} \times V &\to V \\
(t, x) &\mapsto A(t)(x)
\end{aligned}$$

for $A\colon \mathbb{T} \to L(V;V)$. *Suppose that* V *has an inner product* $\langle\cdot,\cdot\rangle$. *Let* $P\colon \mathbb{T} \to L(V;V)$ *be of class* C^1 *and such that* $P(t)$ *is symmetric for every* $t \in \mathbb{T}$ *and let* f_P *be the corresponding time-varying quadratic function as in Definition 10.6.14. Then,*

$$\mathscr{L}_F f_P(t, x) = -f_Q(t, x), \qquad (t, x) \in \mathbb{T} \times V.$$

Proof From the proof of the preceding lemma we have

$$\frac{d}{dt} f_P(t, \Phi_A(t, t_0)(x)) = -f_Q(t, \Phi_A(t, t_0)(x)).$$

Evaluating at $t = t_0$ gives the result. ■

We can now state and prove our main result concerning Lyapunov's Second Method for time-varying linear ordinary differential equations.

10.7.16 Theorem (Lyapunov's Second Method for linear time-varying ordinary differential equations) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *and with right-hand side*

$$\begin{aligned}\widehat{\mathrm{F}}\colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V}\\ (t, x) &\mapsto \mathrm{A}(t)(x)\end{aligned}$$

for $\mathrm{A}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$*. Suppose that* A *is continuous and that* $\sup \mathbb{T} = \infty$*. Suppose that* V *has an inner product* $\langle \cdot, \cdot \rangle$*. Then the following statements hold.*

(i) The equation F *is stable if there exists* $\mathrm{P}, \mathrm{Q}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$ *with the following properties:*
   (a) P *is of class* C^1 *and* Q *is continuous;*
   (b) $\mathrm{P}(t)$ *and* $\mathrm{Q}(t)$ *are symmetric for every* $t \in \mathbb{T}$*;*
   (c) (P, Q) *is a Lyapunov pair for* F*;*
   (d) P *is positive-definite;*
   (e) Q *is positive-semidefinite.*

(ii) The equation F *is uniformly stable if there exists* $\mathrm{P}, \mathrm{Q}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$ *with the following properties:*
   (a) P *is of class* C^1 *and* Q *is continuous;*
   (b) $\mathrm{P}(t)$ *and* $\mathrm{Q}(t)$ *are symmetric for every* $t \in \mathbb{T}$*;*
   (c) (P, Q) *is a Lyapunov pair for* F*;*
   (d) P *is positive-definite;*
   (e) P *is decrescent;*
   (f) Q *is positive-semidefinite.*

(iii) The equation F *is asymptotically stable if there exists* $\mathrm{P}, \mathrm{Q}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$ *with the following properties:*
   (a) P *is of class* C^1 *and* Q *is continuous;*
   (b) $\mathrm{P}(t)$ *and* $\mathrm{Q}(t)$ *are symmetric for every* $t \in \mathbb{T}$*;*
   (c) (P, Q) *is a Lyapunov pair for* F*;*
   (d) P *is positive-definite;*
   (e) Q *is positive-definite.*

(iv) The equation F *is uniformly asymptotically stable if there exists* $\mathrm{P}, \mathrm{Q}\colon \mathbb{T} \to \mathrm{L}(\mathsf{V}; \mathsf{V})$ *with the following properties:*
   (a) P *is of class* C^1 *and* Q *is continuous;*
   (b) $\mathrm{P}(t)$ *and* $\mathrm{Q}(t)$ *are symmetric for every* $t \in \mathbb{T}$*;*
   (c) (P, Q) *is a Lyapunov pair for* F*;*
   (d) P *is positive-definite;*
   (e) P *is decrescent;*

(f) Q *is positive-definite.*

We shall give two proofs of Theorem 10.7.16, one assuming Theorem 10.7.4 and the other an independent proof. The independent proof is interesting in and of itself because it makes use of methods particular to linear equations.

Proof of Theorem 10.7.16, assuming Theorem 10.7.4 If we collect together the conclusions of Lemma 10.6.16 and Corollary 10.7.15, we see that the hypotheses of parts (i)–(iv) of Theorem 10.7.16 imply those of the corresponding parts of Theorem 10.7.4, and thus the conclusions also correspond. ■

Independent proof of Theorem 10.7.16 (i) Let $t_0 \in \mathbb{T}$. Since P is positive-definite, by definition and by Lemma 10.6.13, there exists $C_1 \in \mathbb{R}_{>0}$ such that

$$C_1\|x\|^2 \leq f_P(t,x), \qquad t \in \mathbb{T},\ x \in \mathsf{V}.$$

By Lemma 10.6.13, there exists $C_2 \in \mathbb{R}_{>0}$ such that

$$f_P(t_0,x) \leq C_1\|x\|^2, \qquad x \in \mathsf{V}.$$

Since Q is positive-semidefinite, by Lemma 10.7.14 we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f_P(t, \Phi_A(t,t_0)(x)) \leq 0$$

for all $x \in \mathsf{V}$ and $t \geq t_0$. Therefore, we have

$$C_1\|\Phi_A(t,t_0)(x)\|^2 \leq f_P(t,\Phi_A(t,t_0)(x)) \leq f_P(t_0,x) \leq C_2\|x\|^2$$

for every $x \in \mathsf{V}$ and $t \geq t_0$. Thus

$$\|\Phi_A(t,t_0)(x)\| \leq \sqrt{C_2/C_1}\|x\|,$$

which gives stability.

(ii) Here, since P is positive-definite and decrescent, by definition and by Lemma 10.6.13, we have $C_1, C_2 \in \mathbb{R}_{>0}$ such that

$$C_1\|x\|^2 \leq f_P(t,x) \leq C_2\|x\|^2, \qquad t \in \mathbb{T}.$$

As in the proof of part (i),

$$\frac{\mathrm{d}}{\mathrm{d}t} f_P(t, \Phi_A(t,t_0)(x)) \leq 0$$

for all $(t_0,x) \in \mathbb{T} \times \mathsf{V}$ and $t \geq t_0$. Therefore,

$$C_1\|\Phi_A(t,t_0)(x)\|^2 \leq f_P(t,\Phi_A(t,t_0)(x)) \leq f_P(t_0,x) \leq C_2\|x\|^2$$

for all $(t_0,x) \in \mathbb{T} \times \mathsf{V}$ and $t \geq t_0$. Thus,

$$\|\Phi_A(t,t_0)(x)\| \leq \sqrt{C_2/C_1}\|x\|$$

for every $(t_0,x) \in \mathbb{T} \times \mathsf{V}$ and $t \geq t_0$. This gives uniform stability, as desired.

(iii) Let $t_0 \in \mathbb{T}$. Here we have stability from part (i). From that part of the proof we also have $C_1, C_2 \in \mathbb{R}_{>0}$ (with C_2 possibly depending on t_0) such that

$$C_1\|\Phi_A(t, t_0)(x)\|^2 \leq f_P(t, \Phi_A(t, t_0)(x)) \leq f_P(t_0, x) \leq C_2\|x\|^2$$

for every $x \in \mathsf{V}$ and $t \geq t_0$. Since Q is positive-definite, by definition and by Lemma 10.6.13, there exists $C_3 \in \mathbb{R}_{>0}$ such that

$$C_3\|x\|^2 \leq f_Q(t, x), \qquad (t, x) \in \mathbb{T} \times \mathsf{V}.$$

Thus, by Lemma 10.7.14, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f_P(t, \Phi_A(t, t_0)(x)) = -f_Q(t, \Phi_A(t, t_0)) \leq -C_3\|\Phi_A(t, t_0)(x)\|^2.$$

for all $x \in \mathsf{V}$ and $t \geq t_0$. Therefore, there exists $\gamma \in \mathbb{R}_{\geq 0}$ such that

$$\lim_{t\to\infty} f_P(t, \Phi_A(t, t_0)(x)) = \gamma.$$

We claim that $\gamma = 0$. Suppose otherwise, and that $\gamma \in \mathbb{R}_{>0}$. We then have

$$\begin{aligned}
f_P(t, \Phi_A(t, t_0)(x)) &= f_P(t_0, x) + \int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}\tau} f_P(\tau, \Phi_A(\tau, t_0)(x))\,\mathrm{d}\tau \\
&= f_P(t_0, x) - \int_{t_0}^{t} f_Q(\tau, \Phi_A(\tau, t_0))\,\mathrm{d}\tau \\
&\leq f_P(t_0, x) - C_3 \int_{t_0}^{t} \|\Phi_A(\tau, t_0)(x)\|^2\,\mathrm{d}\tau \\
&\leq f_P(t_0, x) - \frac{C_3}{C_1} \int_{t_0}^{t} f_P(\tau, \Phi_A(\tau, t_0)(x))\,\mathrm{d}\tau \\
&\leq f_P(t_0, x) - \frac{C_3}{C_1}\gamma(t - t_0).
\end{aligned}$$

This implies that $\lim_{t\to\infty} f_P(t, \Phi_A(t, t_0)(x)) = -\infty$. This contradiction leads us to conclude that $\gamma = 0$. Finally, we then have

$$\lim_{t\to\infty} \|\Phi_A(t, t_0)(x)\|^2 \leq \lim_{t\to\infty} C_1^{-1} f_P(t, \Phi_A(t, t_0)(x)) = 0,$$

which gives asymptotic stability.

(iv) Here we have uniform stability from part (i). From that part of the proof we also have $C_1, C_2 \in \mathbb{R}_{>0}$ such that

$$C_1\|\Phi_A(t, t_0)(x)\|^2 \leq f_P(t, \Phi_A(t, t_0)(x)) \leq f_P(t_0, x) \leq C_2\|x\|^2$$

for every $(t_0, x) \in \mathbb{T} \times \mathsf{V}$ and $t \geq t_0$. Since Q is positive-definite by definition and by Lemma 10.6.13, there exists $C_3 \in \mathbb{R}_{>0}$ such that

$$C_3\|x\|^2 \leq f_Q(t, x), \qquad (t, x) \in \mathbb{T} \times \mathsf{V}.$$

Thus, by Lemma 10.7.14, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f_P(t, \Phi_A(t, t_0)(x)) = -f_Q(t, \Phi_A(t, t_0)) \leq -C_3 \|\Phi_A(t, t_0)(x)\|^2.$$

for all $(t_0, x) \in \mathbb{T} \times \mathsf{V}$ and $t \geq t_0$. Therefore, there exists $\gamma \in \mathbb{R}_{\geq 0}$ such that

$$\lim_{t \to \infty} f_P(t, \Phi_A(t, t_0)(x)) = \gamma.$$

We claim that $\gamma = 0$. Suppose otherwise, and that $\gamma \in \mathbb{R}_{>0}$. We then have

$$\begin{aligned}
f_P(t, \Phi_A(t, t_0)(x)) &= f_P(t_0, x) + \int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}\tau} f_P(\tau, \Phi_A(\tau, t_0)(x)) \, \mathrm{d}\tau \\
&= f_P(t_0, x) - \int_{t_0}^{t} f_Q(\tau, \Phi_A(\tau, t_0)) \, \mathrm{d}\tau \\
&\leq f_P(t_0, x) - C_3 \int_{t_0}^{t} \|\Phi_A(\tau, t_0)(x)\|^2 \, \mathrm{d}\tau \\
&\leq f_P(t_0, x) - \frac{C_3}{C_1} \int_{t_0}^{t} f_P(\tau, \Phi_A(\tau, t_0)(x)) \, \mathrm{d}\tau \\
&\leq f_P(t_0, x) - \frac{C_3}{C_1} \gamma (t - t_0).
\end{aligned}$$

This implies that $\lim_{t \to \infty} f_P(t, \Phi_A(t, t_0)(x)) = -\infty$. This contradiction leads us to conclude that $\gamma = 0$. Finally, we then have

$$\lim_{t \to \infty} \|\Phi_A(t, t_0)(x)\|^2 \leq \lim_{t \to \infty} C_1^{-1} f_P(t, \Phi_A(t, t_0)(x)) = 0,$$

which gives uniform asymptotic stability, since C_1, C_2, and C_3 are independent of t_0. ■

10.7.17 Remark (Automatic implications of Theorem 10.7.16) We recall from Theorem 10.2.9 that the conclusions of stability, uniform stability, asymptotic stability, and uniform asymptotic stability are actually of the global sort given in Definition 10.2.7. Moreover, from Proposition 10.3.1 we see that uniform stability and stability are equivalent for linear homogeneous equations with constant coefficients, and similarly that uniform asymptotic stability and asymptotic stability are equivalent. •

10.7.4 The Second Method for linear equations with constant coefficients

The final setting in which we consider conditions for stability using Lyapunov's Second Method is that for linear homogeneous ordinary differential equations with constant coefficients.

As in the time-varying setting of the preceding section, in this section we work with Lyapunov functions that are especially adapted to the linear structure of the equations, namely the quadratic functions of Sections 10.6.4. In working with

such functions, the derivatives along solutions, called the "Lie derivative" in Definitions 10.7.2 and 10.7.7, take a particular form that leads to the following definition and associated following result. What we have, of course, is a specialisation Definition 10.7.13.

10.7.18 Definition (Lyapunov pair for linear ordinary differential equations with constant coefficients) Let F be a system of linear homogeneous ordinary differential equations in an n-dimensional $\mathbb{R}$-vector space V with constant coefficients and with right-hand side

$$\begin{aligned}\widehat{F}\colon \mathbb{T}\times\mathsf{V} &\to \mathsf{V}\\ (t,x) &\mapsto A(x)\end{aligned}$$

for $A \in \mathrm{L}(\mathsf{V};\mathsf{V})$. Suppose that V has an inner product $\langle\cdot,\cdot\rangle$. A ***Lyapunov pair*** for F is a pair (P,Q) where

(i) $P, Q \in \mathrm{L}(\mathsf{V};\mathsf{V})$ are symmetric, and

(ii) $P \circ A + A^T \circ P = -Q$. •

As in the time-varying case, one can think of (1) P as being given, and part (ii) of the definition prescribing Q or (2) Q as being given, and (ii) of the definition giving a linear algebraic equation for P. Both ways of thinking about this will be useful.

Let us indicate the significance of the notion of a Lyapunov pair in this context.

10.7.19 Lemma (Derivative of quadratic function along solutions of a linear ordinary differential equation with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *with constant coefficients and with right-hand side*

$$\begin{aligned}\widehat{\mathrm{F}}\colon \mathbb{T}\times\mathsf{V} &\to \mathsf{V}\\ (\mathrm{t},\mathrm{x}) &\mapsto \mathrm{A}(\mathrm{x})\end{aligned}$$

for $\mathrm{A} \in \mathrm{L}(\mathsf{V};\mathsf{V})$. *Suppose that* V *has an inner product* $\langle\cdot,\cdot\rangle$. *Let* $\mathrm{P} \in \mathrm{L}(\mathsf{V};\mathsf{V})$ *be symmetric and let* $\mathrm{f_P}$ *be the corresponding quadratic function as in Definition 10.6.10. Then, for any solution* $\xi\colon \mathbb{T} \to \mathsf{V}$ *for* F*, we have*

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathrm{f_P}(\xi(\mathrm{t})) = -\mathrm{f_Q}(\xi(\mathrm{t})),$$

where (P,Q) *is a Lyapunov pair for* F.

Proof We shall represent solutions using the state transition map as in Section 5.2.1.2. Thus, if $(t_0,x) \in \mathbb{T}\times\mathsf{V}$, the solution to the initial value problem

$$\dot{\xi}(t) = A(\xi(t)), \quad \xi(t_0) = x,$$

is $\xi(t) = \Phi_A(t, t_0)(x)$. Now we directly compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} f_P(\Phi_A(t,t_0)(x)) &= \frac{\mathrm{d}}{\mathrm{d}t}\langle P \circ \Phi_A(t,t_0)(x), \Phi_A(t,t_0)(x)\rangle \\ &= \langle P \circ \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi_A(t,t_0)(x), \Phi_A(t,t_0)(x)\rangle \\ &\quad + \langle P \circ \Phi_A(t,x_0)(x), \tfrac{\mathrm{d}}{\mathrm{d}t}\Phi_A(t,t_0)(x)\rangle \\ &= \langle P \circ A \circ \Phi_A(t,t_0)(x), \Phi_A(t,t_0)(x)\rangle \\ &\quad + \langle P \circ \Phi_A(t,x_0)(x)), A \circ \Phi_A(t,t_0)(x)\rangle \\ &= -\langle Q \circ \Phi_A(t,t_0)(x), \Phi_A(t,t_0)(x)\rangle,\end{aligned}$$

where (P, Q) is a Lyapunov pair, i.e.,

$$Q = -P \circ A - A^T \circ P.$$ ∎

The lemma allows us to provide the following connection to the Lie derivative characterisation Lemma 10.7.8.

10.7.20 Corollary (Lie derivative of quadratic function along a linear ordinary differential equation with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *with constant coefficients and with right-hand side*

$$\begin{aligned}\widehat{\mathrm{F}}\colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V}\\ (\mathrm{t}, \mathrm{x}) &\mapsto \mathrm{A}(\mathrm{x})\end{aligned}$$

for $\mathrm{A} \in \mathrm{L}(\mathsf{V};\mathsf{V})$*. Suppose that* V *has an inner product* $\langle\cdot,\cdot\rangle$*. Let* $\mathrm{P} \in \mathrm{L}(\mathsf{V};\mathsf{V})$ *be symmetric and let* $\mathrm{f_P}$ *be the corresponding quadratic function as in Definition 10.6.10. Then,*

$$\mathscr{L}_{\mathrm{F}}\mathrm{f_P}(\mathrm{x}) = -\mathrm{f_Q}(\mathrm{x}), \qquad \mathrm{x} \in \mathsf{V}.$$

Proof From the proof of the preceding lemma we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f_P(\Phi_A(t,t_0)(x)) = -f_Q(\Phi_A(t,t_0)(x)).$$

Evaluating at $t = t_0$ gives the result. ∎

We may now state our first result.

10.7.21 Theorem (Lyapunov's Second Method for linear ordinary differential equations with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *with constant coefficients and with right-hand side*

$$\begin{aligned}\widehat{\mathrm{F}}\colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V}\\ (\mathrm{t}, \mathrm{x}) &\mapsto \mathrm{A}(\mathrm{x})\end{aligned}$$

for $\mathrm{A} \in \mathrm{L}(\mathsf{V};\mathsf{V})$*. Suppose that* $\sup \mathbb{T} = \infty$*. Suppose that* V *has an inner product* $\langle\cdot,\cdot\rangle$*. Then the following statements hold.*

(i) *The equation* F *is stable if there exists* $\mathrm{P}, \mathrm{Q} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ *with the following properties:*
  (a) P *and* Q *are symmetric;*
  (b) (P, Q) *is a Lyapunov pair for* F*;*
  (c) P *is positive-definite;*
  (d) Q *is positive-semidefinite.*

(ii) *The equation* F *is asymptotically stable if there exists* $\mathrm{P}, \mathrm{Q} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ *with the following properties:*
  (a) P *and* Q *are symmetric;*
  (b) (P, Q) *is a Lyapunov pair for* F*;*
  (c) P *is positive-definite;*
  (d) Q *is positive-definite.*

We shall give two proofs of Theorem 10.7.21, one assuming Theorem 10.7.9 and the other an independent proof. The independent proof is interesting in and of itself because it makes use of methods particular to linear equations.

Proof of Theorem 10.7.21, assuming Theorem 10.7.9 If we collect together the conclusions of Lemma 10.6.12 and Corollary 10.7.20, we see that the hypotheses of parts (i) and (ii) of Theorem 10.7.21 imply those of the corresponding parts of Theorem 10.7.9, and thus the conclusions also correspond. ■

Independent proof of Theorem 10.7.21 (i) Let $t_0 \in \mathbb{T}$. Since P is positive-definite, by Lemma 10.6.13, there exists $C_1, C_2 \in \mathbb{R}_{>0}$ such that

$$C_1\|x\|^2 \leq f_P(x) \leq C_2\|x\|^2, \qquad x \in \mathsf{V}.$$

Since Q is positive-semidefinite, by Lemma 10.7.19 we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f_P(\Phi_A(t, t_0)(x)) \leq 0$$

for all $x \in \mathsf{V}$ and $t \geq t_0$. Therefore, we have

$$C_1\|\Phi_A(t, t_0)(x)\|^2 \leq f_P(t, \Phi_A(t, t_0)(x)) \leq f_P(t_0, x) \leq C_2\|x\|^2$$

for every $x \in \mathsf{V}$ and $t \geq t_0$. Thus

$$\|\Phi_A(t, t_0)(x)\| \leq \sqrt{C_2/C_1}\|x\|,$$

which gives stability.

(ii) Let $t_0 \in \mathbb{T}$. Here we have stability from part (i). From that part of the proof we also have $C_1, C_2 \in \mathbb{R}_{>0}$ such that

$$C_1\|\Phi_A(t, t_0)(x)\|^2 \leq f_P(\Phi_A(t, t_0)(x)) \leq f_P(x) \leq C_2\|x\|^2$$

for every $x \in \mathsf{V}$ and $t \geq t_0$. Since Q is positive-definite, by Lemma 10.6.13, there exists $C_3 \in \mathbb{R}_{>0}$ such that

$$C_3\|x\|^2 \leq f_Q(x), \qquad x \in \mathsf{V}.$$

Thus, by Lemma 10.7.19, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f_P(\Phi_A(t, t_0)(x)) = -f_Q(\Phi_A(t, t_0)) \leq -C_3 \|\Phi_A(t, t_0)(x)\|^2.$$

for all $x \in \mathsf{V}$ and $t \geq t_0$. Therefore, there exists $\gamma \in \mathbb{R}_{\geq 0}$ such that

$$\lim_{t \to \infty} f_P(\Phi_A(t, t_0)(x)) = \gamma.$$

We claim that $\gamma = 0$. Suppose otherwise, and that $\gamma \in \mathbb{R}_{>0}$. We then have

$$\begin{aligned}
f_P(\Phi_A(t, t_0)(x)) &= f_P(x) + \int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}\tau} f_P(\Phi_A(\tau, t_0)(x)) \, \mathrm{d}\tau \\
&= f_P(x) - \int_{t_0}^{t} f_Q(\Phi_A(\tau, t_0)) \, \mathrm{d}\tau \\
&\leq f_P(x) - C_3 \int_{t_0}^{t} \|\Phi_A(\tau, t_0)(x)\|^2 \, \mathrm{d}\tau \\
&\leq f_P(x) - \frac{C_3}{C_1} \int_{t_0}^{t} f_P(\Phi_A(\tau, t_0)(x)) \, \mathrm{d}\tau \\
&\leq f_P(x) - \frac{C_3}{C_1} \gamma (t - t_0).
\end{aligned}$$

This implies that $\lim_{t \to \infty} f_P(\Phi_A(t, t_0)(x)) = -\infty$. This contradiction leads us to conclude that $\gamma = 0$. Finally, we then have

$$\lim_{t \to \infty} \|\Phi_A(t, t_0)(x)\|^2 \leq \lim_{t \to \infty} C_1^{-1} f_P(\Phi_A(t, t_0)(x)) = 0,$$

which gives asymptotic stability. ■

10.7.22 Remark (Automatic implications of Theorem 10.7.21) We recall from Theorem 10.2.9 that the conclusions of stability, uniform stability, asymptotic stability, and uniform asymptotic stability are actually of the global sort given in Definition 10.2.7. •

10.7.23 Example (Example 10.3.5 cont'd) We again look at the linear homogeneous ordinary differential equation $\boldsymbol{F}$ on $\mathsf{V} = \mathbb{R}^2$ defined by the 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix}.$$

The inner product we use is the standard one:

$$\langle (u_1, u_2), (v_1, v_2) \rangle_{\mathbb{R}^2} = u_1 v_1 + u_2 v_2.$$

In this case, the induced norm is the standard norm for $\mathbb{R}^2$. Note that, if $\boldsymbol{L} \in \mathrm{L}(\mathbb{R}^2; \mathbb{R}^2)$, then the transpose with respect to the standard inner product is just the usual matrix transpose.

For this example, there are various cases to consider, and we look at them separately in view of Theorem 10.7.21. In the following discussion, the reader should compare the conclusions with those of Example 10.3.5.

1. $a = 0$ and $b = 0$: In this case, we know the system is unstable. Thus we will certainly not be able to find a Lyapunov pair $(\boldsymbol{P}, \boldsymbol{Q})$ for $\boldsymbol{F}$ with $\boldsymbol{P}$ positive-definite and $\boldsymbol{Q}$ positive-semidefinite. Note, however, that without knowing more, just the lack of existence of such a $(\boldsymbol{P}, \boldsymbol{Q})$ does not allow us to conclude anything about stability in this case. We shall have more to say about this case in Example 10.9.3–1.
2. $a = 0$ and $b > 0$: The matrices
$$\boldsymbol{P} = \begin{bmatrix} b & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$
have the property that $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$. Since $\boldsymbol{P}$ is positive-definite and $\boldsymbol{Q}$ is positive-semidefinite, stability follows from part (i) of Theorem 10.7.21. Note that asymptotic stability cannot be concluded from this $\boldsymbol{P}$ and $\boldsymbol{Q}$ using part (ii) (and indeed asymptotic stability does not hold in this case).
3. $a = 0$ and $b < 0$: We shall consider this case in Example 10.9.3–2, where we will be able to use Lyapunov methods to conclude instability.
4. $a > 0$ and $b = 0$: Here we take
$$\boldsymbol{P} = \begin{bmatrix} a^2 & a \\ a & 2 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 2a \end{bmatrix}$$
and verify that $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$. The eigenvalues of $\boldsymbol{P}$ are $\{\frac{1}{2}(a^2 + 2 \pm \sqrt{a^4 + 4})\}$. One may verify that $a^2 + 2 > \sqrt{a^4 + 4}$, thus $\boldsymbol{P}$ is positive-definite. Since $\boldsymbol{Q}$ is positive-semidefinite, we conclude stability of $\boldsymbol{F}$ from part (i) of Theorem 10.7.21. However, we cannot conclude asymptotic stability from part (ii); indeed, asymptotic stability does not hold.
5. $a > 0$ and $b > 0$: Here we take
$$\boldsymbol{P} = \begin{bmatrix} b & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 2a \end{bmatrix}$$
having the property that $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$. Since $\boldsymbol{P}$ is positive-definite and $\boldsymbol{Q}$ is positive-semidefinite, from part (i) of Theorem 10.7.21 we can conclude stability for $\boldsymbol{F}$. However, we cannot conclude asymptotic stability using part (ii). However, we *do* have asymptotic stability in this case. We can rectify this in one of two ways.
   (a) By choosing a different $\boldsymbol{P}$ and $\boldsymbol{Q}$ with both positive-semidefinite, we can ensure asymptotic stability by part (ii) of Theorem 10.7.21. Theorem 10.9.2 guarantees that this is possible.
   (b) By resorting to an invariance principle, we can rescue things for this particular $\boldsymbol{P}$ and $\boldsymbol{Q}$. This is explained in Theorem 10.8.8, and in Example 10.8.9 for this example particularly.

6. $a > 0$ and $b < 0$: We shall consider this case in Example 10.9.3–3, where we will be able to use Lyapunov methods to conclude instability.
7. $a < 0$ and $b = 0$: This case is much like case 1 in that the system is unstable; thus we cannot find a Lyapunov pair $(\boldsymbol{P}, \boldsymbol{Q})$ for $\boldsymbol{F}$ with $\boldsymbol{P}$ positive-definite and $\boldsymbol{Q}$ positive-semidefinite. In Example 10.9.3–4 we shall have more to say about this case.
8. $a < 0$ and $b > 0$: We shall consider this case in Example 10.9.3–5, where we will be able to use Lyapunov methods to conclude instability.
9. $a < 0$ and $b < 0$: We shall consider this case in Example 10.9.3–6, where we will be able to use Lyapunov methods to conclude instability. •

The reader can see from this example that, even for a simple linear ordinary differential equation with constant coefficients, the sufficient conditions of Lyapunov's Second Method leave a great deal of room for improvement. In the subsequent sections we shall address this somewhat, although it is still the case that the method is one that is difficult to apply conclusively.

Notes and references

[Liapunov 1893]
[Bacciotti and Rosier 2005] for Lyapunov's Second Method.
[Kellett 2014] for comparison functions.
The original reference for this work is [LaSalle 1968].
[Barbashin and Krasovskiĭ 1952]
Theorem 10.9.1 is due to Chetaev.

Exercises

10.7.1 Determine whether the following functions are or are not of class $\mathcal{K}$:
(a) $[0, \infty) \ni x \mapsto \tan^{-1}(x) \in \mathbb{R}_{\geq 0}$;
(b) $[0, b) \ni x \mapsto x^\alpha \in \mathbb{R}_{\geq 0}$ for $b \in \mathbb{R}_{>0} \cup \{\infty\}$ and $\alpha \in \mathbb{R}_{>0}$;
(c) $[0, b) \ni x \mapsto \min\{\phi_1(x), \phi_2(x)\} \in \mathbb{R}_{\geq 0}$ for $b \in \mathbb{R}_{>0} \cup \{\infty\}$ and $\phi_1, \phi_2 \colon [0, a) \to \mathbb{R}_{\geq 0}$ of class $\mathcal{K}$;
(d) $[0, \pi) \ni x \mapsto \cos(x - \frac{\pi}{2}) + 1 \in \mathbb{R}_{\geq 0}$;
(e) $[0, 2\pi) \ni x \mapsto \cos(x - \frac{\pi}{2}) + 1 \in \mathbb{R}_{\geq 0}$;
(f) $[0, b) \ni x \mapsto \begin{cases} \ln(x), & x > 0, \\ 0, & x = 0 \end{cases}$ for $b \in \mathbb{R}_{>0} \cup \{\infty\}$.

10.7.2 Prove Lemma 10.6.2.

10.7.3 Determine whether the following functions are or are not of class $\mathcal{L}$:
(a) $[a, \infty) \ni y \mapsto \mathrm{e}^{-\sigma y} \in \mathbb{R}_{\geq 0}$ for $a \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$;
(b) $[a, \infty) \ni y \mapsto y^\alpha$ for $a \in \mathbb{R}$ and α 1;
(c) $(-\frac{\pi}{2}, \frac{\pi}{2}) \ni y \mapsto \tan^{-1}(y)$;

(d) $[a, \infty) \ni y \mapsto -\ln(y)$ for $a \in \mathbb{R}$.

10.7.4 Determine whether the following functions are or are not of class $\mathcal{KL}$:

(a) $[0, b) \times [a, \infty) \ni (x, y) \mapsto \phi(x)\psi(y)$, where ϕ is one of the functions from Exercise 10.7.1 and ψ is one of the functions from Exercise 10.7.3;

(b) $[0, b) \times [0, \infty) \ni (x, y) \mapsto \dfrac{x}{\alpha x y + 1}$ for $b \in \mathbb{R}_{>0} \cup \{\infty\}$ and $\alpha \in \mathbb{R}_{>0}$;

(c) $[0, b) \times [0, \infty) \ni (x, y) \mapsto \dfrac{x}{\sqrt{2x^2 y + 1}}$.

10.7.5 Let $\boldsymbol{F}$ be the system of linear ordinary differential equations in $\mathbb{R}^2$ defined by the 2×2-matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & a \end{bmatrix},$$

for $a \geq 0$. Show that if $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$ for which $\boldsymbol{Q}$ is positive-semidefinite, then $(\boldsymbol{A}, \boldsymbol{Q})$ is not observable.

Section 10.8

Invariance principles

We shall see in Section 10.10 that the sufficient conditions for asymptotic stability of some of the flavours of Lyapunov's Second Method are actually also necessary. However, in practice, one often produces a locally positive-definite function whose Lie derivative is merely negative-semidefinite, not negative-definite as one needs for asymptotic stability. In order to deal with this commonly encountered situation, we provide in this section a strategy that falls under a general umbrella of what are know of as "invariance principles." We prove two associated theorems, one for autonomous, not necessarily linear, ordinary differential equations and one for linear ordinary differential equations with constant coefficients.

10.8.1 Invariant sets and limit sets

In order to prove our result for general autonomous ordinary differential equations, we need a collection of preliminary definitions and results.

10.8.1 Definition (Invariant set) Let $\boldsymbol{F}$ be an ordinary differential equation with right-hand side

$$\widehat{F}\colon \mathbb{T} \times U \to \mathbb{R}^n.$$

A subset $A \subseteq U$ is:

(i) **F-*invariant*** if, for all $(t_0, x) \in \mathbb{T} \times A$, the solution $\xi\colon \mathbb{T}' \to U$ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies $\xi(t) \in A$ for every $t \in \mathbb{T}'$;

(ii) ***positively* F-*invariant*** if, for all $(t_0, x) \in \mathbb{T} \times A$, the solution $\xi\colon \mathbb{T}' \to U$ to the initial value problem

$$\dot{\xi}(t) = \widehat{\boldsymbol{F}}(t, \xi(t)), \quad \xi(t_0) = x,$$

satisfies $\xi(t) \in A$ for every $t \geq t_0$. •

10.8.2 Definition (Positive limit set) Let $\boldsymbol{F}$ be an autonomous ordinary differential equation with right-hand side

$$\begin{aligned}\widehat{\boldsymbol{F}}\colon \mathbb{T} \times U &\to \mathbb{R}^n\\ (t, x) &\mapsto \widehat{\boldsymbol{F}}_0(x).\end{aligned}$$

Suppose that $\sup \mathbb{T} = \infty$ and that $0 \in \mathbb{T}$. Let $x_0 \in U$ and let $\xi\colon \mathbb{T}' \to U$ be the solution to the initial value problem

$$\dot{\xi} = \widehat{\boldsymbol{F}}_0(\xi(t)), \quad \xi(0) = x_0,$$

and suppose that $\sup \mathbb{T}' = \infty$.

(i) A point $x \in U$ is a ***positive limit point of*** $\mathbf{x_0}$ if there exists a sequence $(t_j)_{j \in \mathbb{Z}_{>0}} \subseteq \mathbb{R}$ such that
   (a) $t_j < t_{j+1}$, $j \in \mathbb{Z}_{>0}$,
   (b) $\lim_{j \to \infty} t_j = \infty$, and
   (c) $\lim_{j \to \infty} \xi(t_j) = x$.

(ii) The ***positive limit set of*** $\mathbf{x_0}$, denoted by $\Omega(F, x_0)$, is the set of positive limit points of x_0. •

Positive limit sets have many interesting properties.

10.8.3 Lemma (Properties of the positive limit set) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\begin{aligned} \widehat{\mathbf{F}} \colon \mathbb{T} \times \mathrm{U} &\to \mathbb{R}^n \\ (t, \mathbf{x}) &\mapsto \widehat{\mathbf{F}}_0(\mathbf{x}). \end{aligned}$$

Let $\mathrm{A} \subseteq \mathrm{U}$ *be compact and positively* $\mathbf{F}$*-invariant. If* $\mathbf{x}_0 \in \mathrm{A}$*, then* $\Omega(\mathbf{F}, \mathbf{x}_0)$ *is a nonempty, compact, and positively* $\mathbf{F}$*-invariant subset of* A*. Furthermore, if* $\xi \colon \mathbb{T}' \to \mathrm{U}$ *is the solution to the initial value problem*

$$\dot{\xi}(t) = \widehat{\mathbf{F}}_0(\xi(t)), \quad \xi(0) = \mathbf{x}_0,$$

then

$$\lim_{t \to \infty} \mathrm{d}_{\Omega(\mathbf{F}, \mathbf{x}_0)}(\xi(t)) = 0.$$

Proof Let $(t_j)_{j \in \mathbb{Z}_{>0}} \subseteq \mathbb{R}_{>0}$ satisfy $t_j < t_{j+1}$, $j \in \mathbb{Z}_{>0}$, and $\lim_{j \to \infty} t_j = \infty$. The sequence $(\xi(t_j))_{j \in \mathbb{Z}_{>0}} \subseteq A$ has a convergence subsequence by the Bolzano–Weierstrass Theorem. ref By definition, the limit x will be in $\Omega(F, x_0)$. Since A is closed and positively-invariant, $x \in A$. Thus $\Omega(F, x_0)$ is a nonempty subset of A.

If $x \in A \setminus \Omega(F, x_0)$, then there exists $\epsilon \in \mathbb{R}_{>0}$ and $T \in \mathbb{R}_{>0}$ such that $\mathsf{B}(\epsilon, x) \cap \{\xi(t) \mid t \geq T\} = \emptyset$. Therefore, $A \setminus \Omega(F, x_0)$ is open, and thus $\Omega(F, x_0)$ is closed and so compact since A is compact. ref

Let $x \in \Omega(F, x_0)$ and let $t \in \mathbb{R}_{\geq 0}$. There then exists a sequence $(t_j)_{j \in \mathbb{Z}_{>0}}$ such that

$$\lim_{j \to \infty} \xi(t_j) = x.$$

Let η_j, $j \in \mathbb{Z}_{>0}$, be the solution to the initial value problem

$$\dot{\eta}_j(t) = \widehat{F}_0(\eta_j(t)), \quad \eta(0) = \xi(t_j).$$

Then

$$\lim_{j \to \infty} \xi(t + t_j) = \lim_{j \to \infty} \eta_j(t) = \xi(t),$$

by continuity of solutions with respect to initial conditions. This shows that $\xi(t) \in \Omega(F, x_0)$, and so that $\Omega(F, x)$ is positively X-invariant.

Lastly, suppose that there exists $\epsilon \in \mathbb{R}_{>0}$ and a sequence $(t_j)_{j \in \mathbb{Z}_{>0}}$ in $\mathbb{R}_{>0}$ such that

1. $t_j < t_{j+1}$, $j \in \mathbb{Z}_{>0}$,
2. $\lim_{j\to\infty} t_j = \infty$, and
3. $\mathrm{d}_{\Omega(F,x_0)}(\xi(t_j)) \geq \epsilon$, $j \in \mathbb{Z}_{>0}$.

By the Bolzano–Weierstrass Theorem, since A is compact there exists a convergent subsequence $(t_{j_k})_{k\in\mathbb{Z}_{>0}}$ such that $(\xi(t_{j_k}))_{k\in\mathbb{Z}_{>0}}$ converges to, say $x \in A$. Note that $x \in \Omega(F, x_0)$. However, we also have $\mathrm{d}_{\Omega(F,x_0)}(x) \geq \epsilon$. This contradiction means that we must have

$$\lim_{t\to\infty} \mathrm{d}_{\Omega(F,x_0)}(\xi(t)) = 0,$$

as claimed. ■

10.8.2 Invariance principle for autonomous equations

We are now ready to present the LaSalle Invariance Principle on the asymptotic behavior of the integral curves of vector fields.

10.8.4 Theorem (LaSalle Invariance Principle) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}^n$$
$$(t, \mathbf{x}) \mapsto \widehat{\mathbf{F}}_0(\mathbf{x}).$$

Suppose that $\sup \mathbb{T} = \infty$ *and that* $0 \in \mathbb{T}$. *Let* $\mathrm{A} \subseteq \mathrm{U}$ *be compact and positively* $\mathbf{F}$*-invariant. Let* $\mathrm{V}\colon \mathrm{U} \to \mathbb{R}$ *be continuously differentiable and satisfy* $\mathscr{L}_{\mathbf{F}_0}\mathrm{V}(\mathbf{x}) \leq 0$ *for all* $\mathbf{x} \in \mathrm{A}$, *and let* B *be the largest positively* $\mathbf{F}$*-invariant set contained in* $\{\mathbf{x} \in \mathrm{A} \mid \mathscr{L}_{\mathbf{F}_0}\mathrm{V}(\mathbf{x}) = 0\}$. *Then the following statements hold:*

(i) for every $\mathbf{x} \in \mathrm{A}$, *the solution* ξ *to the initial value problem*

$$\dot{\xi}(t) = \widehat{\mathbf{F}}_0(\xi(t)), \quad \xi(0) = \mathbf{x},$$

satisfies $\lim_{t\to\infty} \mathrm{d}_B(\xi(t)) = 0$;

(ii) if B *consists of a finite number of isolated points, then, for every* $\mathbf{x} \in \mathrm{A}$, *there exists* $\mathbf{y} \in \mathrm{B}$ *such that the solution* ξ *to the initial value problem*

$$\dot{\xi}(t) = \widehat{\mathbf{F}}_0(\xi(t)), \quad \xi(0) = \mathbf{x},$$

satisfies $\lim_{t\to\infty} \xi(t) = \mathbf{y}$.

Proof (i) The function $V|A$ is bounded from below, because it is continuous on the ref compact set A. For $x \in A$, let ξ be the solution to the initial value problem

$$\dot{\xi}(t) = \widehat{F}_0(\xi(t)), \quad \xi(0) = x.$$

The function $t \mapsto V \circ \xi(t)$ is nonincreasing and bounded from below. Therefore, $\lim_{t\to\infty} V \circ \xi(t)$ exists and is equal to, say, $\alpha \in \mathbb{R}$. Now, let $y \in \Omega(F, x)$ and let $(t_j)_{j\in\mathbb{Z}_{>0}}$ satisfy $\lim_{j\to\infty} \xi(t_k) = y$. By continuity of V, $\alpha = \lim_{j\to\infty} V \circ \xi(t_k) = V(y)$. This proves

that $V(\boldsymbol{y}) = \alpha$ for all $\boldsymbol{y} \in \Omega(\boldsymbol{F}, \boldsymbol{x})$. Because $\Omega(\boldsymbol{F}, \boldsymbol{x})$ is positively $\boldsymbol{F}$-invariant, if $\boldsymbol{y} \in \Omega(\boldsymbol{F}, \boldsymbol{x})$ and if $\boldsymbol{\eta}$ is the solution to the initial value problem

$$\dot{\boldsymbol{\eta}}(t) = \widehat{\boldsymbol{F}}_0(\boldsymbol{\eta}(t)), \quad \boldsymbol{\eta}(0) = \boldsymbol{y},$$

then $\boldsymbol{\eta}(t) \in \Omega(\boldsymbol{F}, \boldsymbol{x})$ for all $t \in \mathbb{R}_{>0}$. Therefore, $V \circ \boldsymbol{\eta}(t) = \alpha$ for all $t \in \mathbb{R}_{>0}$ and, therefore, by Lemma 10.7.8, $\mathscr{L}_{\boldsymbol{F}_0} V(y) = 0$. Now, because $\mathscr{L}_{\boldsymbol{F}_0} V(\boldsymbol{y}) = 0$ for all $\boldsymbol{y} \in \Omega(\boldsymbol{F}, \boldsymbol{x})$, we know that

$$\Omega(\boldsymbol{F}, \boldsymbol{x}) \subseteq \{\boldsymbol{x} \in A \mid \mathscr{L}_{\boldsymbol{F}_0} V(x) = 0\}.$$

This implies that $\Omega(\boldsymbol{F}, \boldsymbol{x}) \subseteq B$, and this proves this part of the theorem.

(ii) Let $\boldsymbol{x} \in A$ and let $\boldsymbol{\xi}$ be the solution to the initial value problem

$$\dot{\boldsymbol{\xi}}(t) = \widehat{\boldsymbol{F}}_0(\boldsymbol{\xi}(t)), \quad \boldsymbol{\xi}(0) = \boldsymbol{x}.$$

Since $B = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_k\}$ is comprised of isolated points, there exists $\epsilon \in \mathbb{R}_{>0}$ such that

$$\overline{\mathsf{B}}(2\epsilon, \boldsymbol{y}_{j_1}) \cap \overline{\mathsf{B}}(2\epsilon, \boldsymbol{y}_{j_2}) = \varnothing$$

for all $j_1, j_2 \in \{1, \dots, k\}$. By assumption and by part (i), there exists $T \in \mathbb{R}_{>0}$ such that

$$\boldsymbol{\xi}(t) \in \cup_{j=1}^{k} \mathsf{B}(\epsilon, \boldsymbol{y}_j), \qquad t \geq T.$$

Since $\boldsymbol{\xi}$ is continuous, $\boldsymbol{\xi}([T, \infty))$ is connected by . This, however, implies that there must exist $\boldsymbol{y} \in B$ such that $\boldsymbol{\xi}([T, \infty)) \subseteq \mathsf{B}(\epsilon, \boldsymbol{y})$, giving this part of the theorem. ■ ref

The following more or less immediate corollary provides a common situation where the LaSalle Invariance Principle is used.

10.8.5 Corollary (Barbashin–Krasovskiĭ criterion) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}} \colon \mathbb{T} \times \mathsf{U} \to \mathbb{R}^n$$
$$(t, \mathbf{x}) \mapsto \widehat{\mathbf{F}}_0(\mathbf{x}).$$

Suppose that $\sup \mathbb{T} = \infty$ *and that* $0 \in \mathbb{T}$. *Let* $\mathbf{x}_0 \in \mathsf{U}$ *be an equilibrium point for* $\mathbf{F}$. *Assume tha there exists a function* $\mathrm{V} \colon \mathsf{U} \to \mathbb{R}$ *with the following properties:*

(i) V *is of class* C^1*;*

(ii) $\mathrm{V} \in \mathrm{LPD}(\mathbf{x}_0)$*;*

(iii) $-\mathscr{L}_{\mathbf{F}_0} \mathrm{V} \in \mathrm{LPSD}(\mathbf{x}_0)$.

Let $\mathsf{C} = \{\mathbf{x} \in \mathsf{U} \mid \mathscr{L}_{\mathbf{F}_0} \mathrm{V}(\mathrm{x}) = 0\}$. *If there exists* $\mathrm{r} \in \mathbb{R}_{>0}$ *such that the only positively* $\mathbf{F}$*-invariant subset of* $\mathsf{C} \cap \mathsf{B}(\mathrm{r}, \mathbf{x}_0)$ *is* $\{\mathbf{x}_0\}$*, then* $\mathbf{x}_0$ *is asymptotically stable.*

Proof As in the proof of Theorem 10.7.9(i), the fact that $V \in \mathrm{LPD}(\boldsymbol{x}_0)$ ensures that there is a closed subset of some ball about $\boldsymbol{x}_0$ that is $\boldsymbol{F}$-positively invariant. The corollary then follows from Theorem 10.8.4. ■

10.8.3 Invariance principle for linear equations with constant coefficients

Next we turn to an invariance principle specifically adapted to linear ordinary differential equations with constant coefficients. Unsurprisingly, the construction is linear algebraic in nature. The key to the construction is the following definition.

10.8.6 Definition (Observability operator, observable pair) Let V and W be finite-dimensional $\mathbb{R}$-vector spaces, and let $A \in \mathrm{L}(\mathsf{V};\mathsf{V})$ and $C \in \mathrm{L}(\mathsf{V};\mathsf{W})$.

(i) The ***observability operator*** for the pair (A, C) is the linear map

$$\begin{aligned} O(A,C)\colon \mathsf{V} &\to \mathsf{U}^{\dim_{\mathbb{R}}(\mathsf{V})} \\ v &\mapsto (C(v), C\circ A(v), \dots, C^{\dim_{\mathbb{R}}(\mathsf{V})-1}\circ A(v)). \end{aligned}$$

(ii) The pair (A, C) is ***observable*** if $\operatorname{rank}(O(A,C)) = \dim_{\mathbb{R}}(\mathsf{V})$. •

This definition, while clear, does not capture the essence of the attribute of observability. The following result goes towards clarifying this.

10.8.7 Lemma (Characterisation of observability) *Let* V *and* W *be finite-dimensional* $\mathbb{R}$*-vector spaces, and let* $\mathrm{A} \in \mathrm{L}(\mathsf{V};\mathsf{V})$ *and* $\mathrm{C} \in \mathrm{L}(\mathsf{V};\mathsf{W})$*. Let* $\mathbb{T} \subseteq \mathbb{R}$ *be a time-domain for which* $0 \in \mathbb{T}$ *and* $\operatorname{int}(\mathbb{T}) \neq \emptyset$*. Then* (A, C) *is observable if and only if, given* $\mathrm{x}_1, \mathrm{x}_2 \in \mathsf{V}$ *with* $\xi_1, \xi_2\colon \mathbb{T} \to \mathsf{V}$ *the solutions to the initial value problems*

$$\dot{\xi}_{\mathrm{a}}(\mathrm{t}) = \mathrm{A}(\xi_{\mathrm{a}}(\mathrm{t})), \quad \xi_{\mathrm{a}}(0) = \mathrm{x}_{\mathrm{a}}, \qquad \mathrm{a} \in \{1, 2\},$$

we have $\mathrm{C}\circ\xi_1 = \mathrm{C}\circ\xi_2$ *if and only if* $\mathrm{x}_1 = \mathrm{x}_2$.

Moreover, $\ker(\mathrm{O}(\mathrm{A},\mathrm{C}))$ *is the largest* A*-invariant subspace contained in* $\ker(\mathrm{C})$.

Proof Let $n = \dim_{\mathbb{R}}(\mathsf{V})$.

First suppose that (A, C) is observable and that $C\circ\xi_1 = C\circ\xi_2$. Then, differentiating successively with respect to t,

$$\frac{\mathrm{d}^j(C\circ\xi_a)}{\mathrm{d}t^j}(0) = C\circ A^j(x_a), \qquad j \in \mathbb{Z}_{\geq 0},\ a \in \{1, 2\}.$$

Thus we have

$$C\circ A^j(x_1) = C\circ A^j(x_2), \qquad j \in \mathbb{Z}_{\geq 0}.$$

Thus $x_1 - x_2 \in \ker(O(A,C))$, and so $x_1 = x_2$ since $O(A, C)$ is observable.

Next suppose that (A, C) is not observable, and so $O(A, C)$ is not injective. Thus there exists a nonzero $x_0 \in \ker(O(A,C))$, meaning that $C\circ A^j(x_0) = 0$, $j \in \{0, 1, \dots, n-1\}$. By the Cayley–Hamilton Theorem, this implies that $C\circ A^j(x_0) = 0$ for $j \in \mathbb{Z}_{\geq 0}$. Therefore, for any $t \geq 0$

$$\sum_{j=0}^{\infty} \frac{C\circ A^j}{j!}(x_0) = C\circ \mathrm{e}^{At}(x_0) = 0.$$

Therefore, taking $x_1 = x_0$ and $x_2 = 0$, $C\circ\xi_1 = C\circ\xi_2$ while $x_1 \neq x_2$.

Now we prove the final assertion of the lemma. First let us show that $\ker(O(A,C)) \subseteq \ker(C)$. If $x \in \ker(O(A,C))$, then $C\circ A^j(x) = 0$ for $j \in \{0, 1, \dots, n-1\}$. This holds in particular for $j = 0$, giving the desired conclusion in this case.

Next we show that the kernel of $O(A, C)$ is A-invariant. Let $x \in \ker(O(A,C))$ and compute

$$O(A,C)\circ A(x) = (C\circ A(x), \dots, C\circ A^n(x)).$$

Since $x \in \ker(O(A, C))$, we have

$$C \circ A(x) = 0, \dots, C \circ A^{n-1}(x) = 0.$$

Also, by the Cayley–Hamilton Theorem, $C \circ A^n(x) = 0$. This shows that

$$O(A, c) \circ A(x) = 0,$$

or that $A(x) \in \ker(O(A, C))$.

Finally, we show that, if S is an A-invariant subspace contained in $\ker(C)$, then S is a subspace of $\ker(O(A, C))$. Given such an S and $x \in \mathsf{S}$, $C(x) = 0$. Since S is A-invariant, $A(x) \in V$, and since $\mathsf{S} \subseteq \ker(C)$, $C \circ A(x) = 0$. Proceeding in this way we see that

$$C \circ A^2(x) = \cdots = C \circ A^{n-1}(x) = 0.$$

But this means exactly that x is in $\ker(O(A, C))$. ■

The idea of observability is this. The linear map C we view as providing us with "measurements" in W of the states in V. The pair (A, C) is observable if we can deduce the state behaviour of the system merely by observing the measurements via C.

With this brief discussion of observability, we can now state a version of Theorem 10.8.4 adapted specially for linear differential equations.

10.8.8 Theorem (Invariance principle for linear ordinary differential equations with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *with constant coefficients and with right-hand side*

$$\begin{aligned} \widehat{\mathrm{F}} \colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V} \\ (\mathrm{t}, \mathrm{x}) &\mapsto \mathrm{A}(\mathrm{x}) \end{aligned}$$

for $\mathrm{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$. *Suppose that* $\sup \mathbb{T} = \infty$. *Suppose that* V *has an inner product* $\langle \cdot, \cdot \rangle$. *Then* F *is asymptotically stable if there exists* $\mathrm{P}, \mathrm{Q} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ *with the following properties:*

(i) P *and* Q *are symmetric;*

(ii) (P, Q) *is a Lyapunov pair for* F*;*

(iii) P *is positive-definite;*

(iv) Q *is positive-semidefinite;*

(v) (A, Q) *is observable.*

We shall offer two proofs of the preceding theorem, one assuming the more general Theorem 10.8.4 and the other an independent proof.

Proof of Theorem 10.8.8, assuming Theorem 10.8.4 Under the hypotheses of Theorem 10.8.8, the function $V = f_P$ satisfies the hypotheses of Corollary 10.8.5. The subset C from the statement of Corollary 10.8.5 is then exactly the subspace $\ker(Q)$. Since (A, Q) is observable, by Lemma 10.8.7 $\{0\}$ is the largest A-invariant subspace of $\ker(Q)$. Since any invariant subset is contained in an invariant subspace—namely the subspace generated by the subset—it follows that the only F-invariant subset of C is $\{0\}$. Thus Theorem 10.8.8 follows from Theorem 10.8.4, specifically its Corollary 10.8.5. ■

Independent proof of Theorem 10.8.8 We suppose that P is positive-definite, Q is positive-semidefinite, (A, Q) is observable, and that F is not asymptotically stable. By Theorem 10.7.9(i) we know that F is stable, so it must be the case that A has at least one eigenvalue on the imaginary axis, and, therefore, a nontrivial periodic solution ξ. From our characterisation of the operator exponential in Procedures 5.2.23 and 5.2.26, we know that this periodic solution takes values in a two-dimensional subspace that we shall denote by L. What's more, every solution of F with initial condition in L is periodic and remains in L, i.e., L is F-invariant. Indeed, if $x \in L$, then

$$A(x) = \lim_{t \to 0} \frac{\mathrm{e}^{At}(x) - x}{t} \in \mathsf{L}$$

since $x, \mathrm{e}^{At}(x) \in \mathsf{L}$. We also claim that the subspace L is in $\ker(Q)$. To see this, suppose that the solutions in L have period T. We have, for any solution $\xi \colon [0, T] \to \mathsf{L}$ for F, by Lemma 10.7.19,

$$0 = f_P \circ \xi(T) - f_P \circ \xi(0) = \int_0^T \frac{\mathrm{d} f_P \circ \xi}{\mathrm{d}t}(t)\,\mathrm{d}t = -\int_0^T f_Q \circ \xi(t)\,\mathrm{d}t.$$

Since Q is positive-semidefinite, this implies that $f_Q \circ \xi(t) = 0$ for $t \in [0, T]$. Thus $\mathsf{L} \subseteq \ker(Q)$, as claimed. Thus, with our initial assumptions, we have shown the existence of a nontrivial A-invariant subspace of $\ker(Q)$. This is a contradiction, however, since (A, Q) is observable. It follows, therefore, that F is asymptotically stable. ■

Let us resume our Example 10.7.23 to conclude asymptotic stability in the case where this is possible.

10.8.9 Example (Example 10.7.23 cont'd) We continue with the linear homogeneous ordinary differential equation $\boldsymbol{F}$ on $\mathsf{V} = \mathbb{R}^2$ defined by the 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix}.$$

Again, we use the standard inner product.

We consider the case where $a > 0$ and $b > 0$, since we know that $\boldsymbol{A}$ is Hurwitz in this case. We take

$$\boldsymbol{P} = \begin{bmatrix} b & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 2a \end{bmatrix},$$

noting that $\boldsymbol{P}$ is positive-definite, $\boldsymbol{Q}$ is positive-semidefinite, and $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$. Using Theorem 10.7.21, we can only conclude stability, and not asymptotic stability. But we can compute

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{Q}) = \begin{bmatrix} 0 & 0 \\ 0 & 2a \\ 0 & 0 \\ -2ab & -2a^2 \end{bmatrix},$$

implying that $(\boldsymbol{A}, \boldsymbol{Q})$ is observable. We can thus conclude from Theorem 10.8.8 that $\boldsymbol{F}$ is asymptotically stable. •

Section 10.9

Lyapunov's Second Method: Instability theorems

In this section we provide two so-called instability theorems. While our results above in this section give sufficient conditions for various flavours of stability, instability theorems give sufficient conditions for instability. The instability results we give fit under the umbrella of Lyapunov's Second method since the characterisations we give involve functions having certain properties. While our sufficient conditions for stability using Lyapunov's Second Method in Sections 10.7.1, 10.7.2, 10.7.3, and 10.7.4 are quite comprehensive, we shall back off from this level of exhaustiveness here, and only give two theorems, both for autonomous ordinary differential equations, one in the linear case and one in the not necessarily linear case.

10.9.1 Instability theorem for autonomous equations

Let us state the more general result first. Some notation is useful. Let $U \subseteq \mathbb{R}^n$ be open, let $f\colon U \to \mathbb{R}$ be continuous, and let $x_0 \in U$. We suppose that $r \in \mathbb{R}_{>0}$ is such that $\overline{\mathsf{B}}(r, x_0) \subseteq U$ and define, for $a \in \mathbb{R}$,

$$f^{-1}(r, > a) = \{x \in \overline{\mathsf{B}}(r, x_0) \mid \ f(x) > a\}.$$

With this simple piece of notation, we then have the following result.

10.9.1 Theorem (An instability for autonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\begin{aligned} \widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} &\to \mathbb{R}^{\mathrm{n}} \\ (\mathrm{t}, \mathbf{x}) &\mapsto \widehat{\mathbf{F}_0}(\mathbf{x}). \end{aligned}$$

Suppose that $\sup \mathbb{T} = \infty$. *Then an equilibrium state* $\mathbf{x}_0 \in \mathrm{U}$ *is unstable if there exists a function* $\mathrm{V}\colon \mathrm{U} \to \mathbb{R}$ *and* $\mathrm{r} \in \mathbb{R}_{>0}$ *with the following properties:*

(i) V *is of class* C^1*;*

(ii) $\mathrm{V}(\mathbf{x}_0) = 0$*;*

(iii) $\overline{\mathsf{B}}(\mathrm{r}, \mathbf{x}_0) \subseteq \mathrm{U}$*;*

(iv) $\mathrm{V}^{-1}(\mathrm{s}, > 0) \neq \varnothing$ *for every* $\mathrm{s} \in (0, \mathrm{r})$*;*

(v) $\mathscr{L}_{\mathbf{F}_0}\mathrm{V}(\mathbf{x}) \in \mathbb{R}_{>0}$ *for* $\mathbf{x} \in \mathsf{B}(\mathrm{r}, \mathbf{x}_0)$.

Proof Let $\epsilon = \frac{r}{2}$ and let $\delta \in \mathbb{R}_{>0}$. We show that there exists $(t_0, x) \in \mathbb{T} \times \mathsf{B}(\delta, x_0)$ such that the solution ξ to the initial value problem

$$\dot{\xi}(t) = \widehat{F}_0(\xi(t)), \quad \xi(t_0) = x$$

satisfies $\xi(T) \notin \mathsf{B}(\epsilon, x_0)$ for some $T \geq t_0$. Indeed, let $\delta \in \mathbb{R}_{>0}$ and choose $x \in V^{-1}(s, >0)$ for $s \leq \min\{\epsilon, \delta\}$. We claim that $\xi(T) \notin \mathsf{B}(\epsilon, x_0)$ for some $T \geq t_0$. Suppose otherwise and let

$$\beta = \inf\{\mathscr{L}_{F_0} V(x') \mid x' \in \overline{\mathsf{B}}(\epsilon, x_0), V(x') \geq V(x)\}.$$

Note that $\beta \in \mathbb{R}_{>0}$ since it is the infimum of a positive-valued function over the compact set

$$\overline{\mathsf{B}}(\epsilon, x_0) \cap \{x' \in \overline{\mathsf{B}}(\epsilon, x_0) \mid V(x') \geq V(x)\}.$$

ref Now we calculate, using Lemma 10.7.8,

$$\begin{aligned} V(\xi(t)) &= V(\xi(t_0)) + \int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}\tau} V(\xi(\tau))\, \mathrm{d}\tau \\ &= V(x) + \int_{t_0}^{t} \mathscr{L}_{F_0} V(\xi(\tau))\, \mathrm{d}\tau \\ &\geq V(x) + \beta(t - t_0). \end{aligned}$$

Thus $t \mapsto V(\xi(t))$ is unbounded as $t \to \infty$, which is a contradiction since $x \mapsto V(x)$ is bounded on $\overline{\mathsf{B}}(\epsilon, x_0)$. Thus we conclude that $\xi(T) \notin \overline{\mathsf{B}}(\epsilon, x_0)$ for some $T \geq t_0$. This gives the desired instability. ■

10.9.2 Instability theorem for linear equations with constant coefficients

Next we consider an instability theorem for linear homogeneous ordinary differential equations with constant coefficients. The result we give is one that makes use of very particular attributes of linear ordinary differential equations, and, in particular, makes use of the notion of observability introduced in Definition 10.8.6.

10.9.2 Theorem (An instability theorem for linear ordinary differential equations with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *with constant coefficients and with right-hand side*

$$\begin{aligned} \widehat{\mathrm{F}} \colon \mathbb{T} \times \mathsf{V} &\to \mathsf{V} \\ (\mathrm{t}, \mathrm{x}) &\mapsto \mathrm{A}(\mathrm{x}) \end{aligned}$$

for $\mathrm{A} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$. *Suppose that* $\sup \mathbb{T} = \infty$. *Suppose that* V *has an inner product* $\langle \cdot, \cdot \rangle$. *Then* F *is unstable if there exists* $\mathrm{P}, \mathrm{Q} \in \mathrm{L}(\mathsf{V}; \mathsf{V})$ *with the following properties:*

(i) P *and* Q *are symmetric;*

(ii) (P, Q) *is a Lyapunov pair for* F*;*

(iii) P *is not positive-semidefinite;*

(iv) Q *is positive-semidefinite;*

(v) (A, Q) *is observable.*

Proof Since Q is positive-semidefinite and (A, Q) is observable, the argument from the proof of Theorem 10.8.8 shows that there are no nontrivial periodic solutions for F. Thus this part of the theorem will follow if we can show that F is not asymptotically

stable. By hypothesis, there exists $x_0 \in \mathsf{V}$ so that $f_P(x_0) < 0$. Let $\xi(t) = \mathrm{e}^{At}(x_0)$ be the solution of F with initial condition x_0 at $t = 0$. As in the proof of Theorem 10.7.21(i), we have $f_P \circ \xi(t) \leq f_P(x_0) < 0$ for all $t \geq 0$ since Q is positive-semidefinite. Denote

$$r = \inf\{\|x\| \mid f_P(x) \leq f_P(x_0)\},$$

and observe that $r \in \mathbb{R}_{>0}$. We have shown that $\|\xi(t)\| \geq r$ for all $t \geq 0$. This prohibits internal asymptotic stability, and in this case, internal stability. ■

Let us use this theorem to fill in a few gaps left by our treatment of Example 10.7.23.

10.9.3 Example (Example 10.7.23 (cont'd)) We continue with the linear homogeneous ordinary differential equation $\boldsymbol{F}$ on $\mathsf{V} = \mathbb{R}^2$ defined by the 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix}.$$

Again, we use the standard inner product.

We consider here the unstable cases.

1. $a = 0$ and $b = 0$: In this case, by Exercise 10.7.5, if $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$ with $\boldsymbol{Q}$ positive-semidefinite, then $(\boldsymbol{A}, \boldsymbol{Q})$ is not observable. This means that we cannot conclude instability using Theorem 10.9.2.
2. $a = 0$ and $b < 0$: If we define

$$\boldsymbol{P} = \frac{1}{2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} -b & 0 \\ 0 & 1 \end{bmatrix},$$

then one verifies that $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$. However, $\boldsymbol{P}$ is not positive-semidefinite (its eigenvalues are $\{\pm\frac{1}{2}\}$), while $\boldsymbol{Q}$ is positive-definite. Since $\boldsymbol{Q}$ is invertible, one can immediately conclude observability, and, therefore, conclude instability from Theorem 10.9.2.
3. $a > 0$ and $b < 0$: We use the Lyapunov pair $(\boldsymbol{P}, \boldsymbol{Q})$ with

$$\boldsymbol{P} = \begin{bmatrix} b & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 2a \end{bmatrix}.$$

Here we compute

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{Q}) = \begin{bmatrix} 0 & 0 \\ 0 & 2a \\ 0 & 0 \\ -2ab & -2a^2 \end{bmatrix}.$$

Since $\boldsymbol{P}$ is not positive-semidefinite, since $\boldsymbol{Q}$ is positive-semidefinite, and since $(\boldsymbol{A}, \boldsymbol{Q})$ is observable we conclude from Theorem 10.9.2 that $\boldsymbol{F}$ is unstable.

4. $a < 0$ and $b = 0$: In this case, as in case 1, if $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$ with $\boldsymbol{Q}$ positive-semidefinite, then $(\boldsymbol{A}, \boldsymbol{Q})$ is not observable. Thus the instability that holds in this case cannot be determined from Theorem 10.9.2.
5. $a < 0$ and $b > 0$: We note that, if

$$\boldsymbol{P} = \begin{bmatrix} -b & 0 \\ 0 & -1 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & -2a \end{bmatrix},$$

then $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$. We also have

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{Q}) = \begin{bmatrix} 0 & 0 \\ 0 & -2a \\ 0 & 0 \\ 2ab & 2a^2 \end{bmatrix}.$$

Thus $(\boldsymbol{A}, \boldsymbol{Q})$ is observable. Since $\boldsymbol{P}$ is not positive-definite and since $\boldsymbol{Q}$ is positive-semidefinite, we conclude from Theorem 10.9.2 that $\boldsymbol{F}$ is unstable.
6. Here we again take

$$\boldsymbol{P} = \begin{bmatrix} -b & 0 \\ 0 & -1 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & -2a \end{bmatrix}.$$

The same argument as in the previous case will tell us that $\boldsymbol{F}$ is unstable. •

Section 10.10

Lyapunov's Second Method: Converse theorems

The results of Sections 10.7.1, 10.7.2, 10.7.3, and 10.7.4 provide useful sufficient conditions for stability and asymptotic stability of equilibria. However, if there are lots of examples of ordinary differential equations that are stable, but for which the hypotheses of these theorems do not hold, then this reduces their potential effectiveness in practice. For this reason, in this section we give six so-called "converse theorems," i.e., theorems that assert the manner in which the converses of conditions like those in the preceding sections also hold. One is for general, nonautonomous, not necessarily linear ordinary differential equations. The next is for exponential stability for nonautonomous ordinary differential equations. Both of these results are mirrored for autonomous systems, with self-contained proof for readers wishing to sidestep time dependence. The other two are results for linear homogeneous ordinary differential equations, one a result for time-varying equations and the other a result for equations with constant coefficients.

10.10.1 Converse theorems for nonautonomous equations

We begin with the most general result.

10.10.1 Theorem (A converse theorem for nonautonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}^n$$

and let $\mathbf{x}_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$*. Assume that* $\sup \mathbb{T} = \infty$, $\mathrm{T}_- \triangleq \inf \mathbb{T} > -\infty$, *and that* $\mathbf{F}$ *satisfies Assumption 10.2.1. If* $\mathbf{x}_0$ *is uniformly asymptotically stable, then there exists* $\mathrm{V}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}$ *such that*

(i) V *is of class* C^1,

(ii) $\mathrm{V} \in \mathrm{TVLPD}_{s_0}(\mathbf{x}_0)$,

(iii) $\mathrm{V} \in \mathrm{TVLD}_{s_0}(\mathbf{x}_0)$,

(iv) $(t, \mathbf{x}) \mapsto \frac{\partial \mathrm{V}}{\partial x_j}(t, \mathbf{x})$ *is in* $\mathrm{TVLD}_{s_0}(\mathbf{x}_0)$, *and*

(v) $-\mathscr{L}_{\mathbf{F}}\mathrm{V} \in \mathrm{TVLPD}_{s_0}(\mathbf{x}_0)$.

Proof ∎

Next we specialise the preceding result to exponential stability, not just asymptotic stability.

10.10.2 Theorem (A converse theorem for exponential stability of nonautonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}^{\mathrm{n}}$$

and let $\mathbf{x}_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$*. Assume that* $\sup \mathbb{T} = \infty$, $\mathrm{T}_- \triangleq \inf \mathbb{T} > -\infty$, *and that there exists* $\mathrm{M}, \mathrm{r} \in \mathbb{R}_{>0}$ *such that*

$$\left| \frac{\partial \widehat{\mathrm{F}_{\mathrm{j}}}}{\partial \mathrm{x}_{\mathrm{k}}}(\mathrm{t}, \mathbf{x}) \right| \leq \mathrm{M}, \qquad \mathrm{j}, \mathrm{k} \in \{1, \dots, \mathrm{n}\},\ (\mathrm{t}, \mathbf{x}) \in \mathbb{T} \times \overline{\mathsf{B}}(\mathrm{r}, \mathbf{x}_0).$$

If there exist $\mathrm{L}, \delta, \sigma \in \mathbb{R}_{>0}$ *such that, if* $\mathbf{x} \in \mathrm{U}$ *satisfies* $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, *then* $\mathrm{t} \mapsto \Phi^{\mathbf{F}}(\mathrm{t}, \mathrm{t}_0, \mathbf{x}_0)$ *is defined on* $[\mathrm{t}_0, \infty)$ *and satisfies*

$$\|\Phi^{\mathbf{F}}(\mathrm{t}, \mathrm{t}_0, \mathbf{x}) - \mathbf{x}_0\| \leq \mathrm{L} \mathrm{e}^{-\sigma(\mathrm{t}-\mathrm{t}_0)} \|\mathbf{x} - \mathbf{x}_0\|,$$

then there exist $\mathrm{V}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}$ *and* $\mathrm{r}_0 \in \mathbb{R}_{>0}$ *such that*

(i) V *is of class* C^1*;*

(ii) there exists $\mathrm{C}_1 \in \mathbb{R}_{>0}$ *such that*

$$\left\| \frac{\partial \mathrm{V}}{\partial \mathrm{x}_{\mathrm{j}}}(\mathrm{t}, \mathbf{x}) \right\| \leq \mathrm{C}_1 \|\mathbf{x} - \mathbf{x}_0\|, \qquad \mathrm{j} \in \{1, \dots, \mathrm{n}\},\ (\mathrm{t}, \mathbf{x}) \in \mathbb{T} \times \mathsf{B}(\mathrm{r}_0, \mathbf{x}_0);$$

(iii) there exists $\mathrm{C}_2 \in \mathbb{R}_{>0}$ *such that*

$$\mathrm{C}_2 \|\mathbf{x} - \mathbf{x}_0\|^2 \leq \mathrm{V}(\mathrm{t}, \mathbf{x}) \leq \mathrm{C}_2^{-1} \|\mathbf{x} - \mathbf{x}_0\|^2, \qquad (\mathrm{t}, \mathbf{x}) \in \mathbb{T} \times \mathsf{B}(\mathrm{r}_0, \mathbf{x}_0);$$

(iv) there exists $\mathrm{C}_3 \in \mathbb{R}_{>0}$ *such that*

$$\mathscr{L}_{\mathbf{F}} \mathrm{V}(\mathrm{t}, \mathbf{x}) \leq -\mathrm{C}_3 \|\mathbf{x} - \mathbf{x}_0\|^2, \qquad (\mathrm{t}, \mathbf{x}) \in \mathbb{T} \times \mathsf{B}(\mathrm{r}_0, \mathbf{x}_0).$$

Proof We start with a few technical lemmata.

Are we using this?

1 Lemma *If* $\mathbb{T}$ *is an interval and if* $\gamma\colon \mathbb{T} \to \mathbb{R}^{\mathrm{n}}$ *is of class* C^1*, then*

$$\frac{\mathrm{d}}{\mathrm{dt}} \|\gamma(\mathrm{t})\| \leq \left\| \frac{\mathrm{d}\gamma}{\mathrm{dt}}(\mathrm{t}) \right\|.$$

Proof The first thing we need to do is understand what we mean by $\frac{\mathrm{d}}{\mathrm{d}t}\|\gamma(t)\|$, since it may be that $t \mapsto \|\gamma(t)\|$ is not differentiable. We shall use the notion of weak differentiability from ref . First let us suppose that $\boldsymbol{\gamma(t)} \neq 0$. Then, by continuity, $\gamma(\tau) \neq 0$ for τ nearby t. Then,

$$2\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\|\gamma(\tau)\|\right)\|\gamma(t)\| = \frac{\mathrm{d}}{\mathrm{d}\tau}\|\gamma(\tau)\|^2 = 2\left\langle \frac{\mathrm{d}}{\mathrm{d}\tau}\gamma(\tau), \gamma(\tau) \right\rangle_{\mathbb{R}^n}.$$

Then, by the Cauchy–Bunyakovsky–Schwarz inequality,

$$2\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\|\gamma(\tau)\|\right)\|\gamma(t)\| = 2\left\langle \frac{\mathrm{d}}{\mathrm{d}\tau}\gamma(\tau), \gamma(\tau) \right\rangle_{\mathbb{R}^n} \leq 2\left\| \frac{\mathrm{d}}{\mathrm{d}\tau}\gamma(\tau) \right\| \|\gamma(\tau)\|.$$

Thus, when $\gamma(t) \neq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\gamma(t)\| \leq \left\|\frac{\mathrm{d}}{\mathrm{d}\tau}\gamma(\tau)\right\|.$$

We need to account for the possibility that $\gamma(t)$ may be zero. Note that

$$\Gamma \triangleq \{t \in \mathbb{T} \mid \|\gamma(t)\| > 0\}$$

is open. Thus, by , there exists a countable set J and a collection I_j, $j \in J$, of open ref intervals such that $\Gamma = \cup_{j=1}^{\infty} I_j$. Let $\phi \in \mathscr{D}(\mathbb{T};\mathbb{R})$. Then

$$\begin{aligned}\int_{\mathbb{T}} \|\gamma(t)\|\dot{\phi}(t)\,\mathrm{d}t &= \sum_{j\in J}\int_{I_j} \|\gamma(t)\|\dot{\phi}(t)\,\mathrm{d}t \\ &= -\sum_{j\in J}\int_{I_j} \frac{\langle \frac{\mathrm{d}}{\mathrm{d}t}\gamma(t), \gamma(t)\rangle}{\|\gamma(t)\|}\phi(t)\,\mathrm{d}t.\end{aligned}$$

using integration by parts. Thus $t \mapsto \|\gamma(t)\|$ is differentiable in the sense of distributions, and its derivative in this sense is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\gamma(t)\| = \begin{cases} \frac{\langle \frac{\mathrm{d}}{\mathrm{d}t}\gamma(t), \gamma(t)\rangle}{\|\gamma(t)\|}, & \gamma(t) \neq \mathbf{0}, \\ 0, & \gamma(t) = 0. \end{cases}$$

The lemma now follows from our estimates above. ▼

2 Lemma *Let $\mathbb{T}$ be an interval and let $\alpha, \beta, \xi \colon \mathbb{T} \to \mathbb{R}$ be such that*

(i) *α and β are continuous,*

(ii) *ξ is continuously differentiable, and*

(iii) *$\alpha(\mathrm{t})\xi(\mathrm{t}) \leq \dot{\xi}(\mathrm{t}) \leq \beta(\mathrm{t})\xi(\mathrm{t})$, $\mathrm{t} \in \mathbb{T}$.*

Then, for any $\mathrm{t}_0 \in \mathbb{T}$,

$$\xi(\mathrm{t}_0)\mathrm{e}^{\int_{\mathrm{t}_0}^{\mathrm{t}} \alpha(\tau)\,\mathrm{d}\tau} \leq \xi(\mathrm{t}) \leq \xi(\mathrm{t}_0)\mathrm{e}^{\int_{\mathrm{t}_0}^{\mathrm{t}} \beta(\tau)\,\mathrm{d}\tau}, \qquad \mathrm{t} \geq \mathrm{t}_0.$$

Proof Denote $\eta \colon \mathbb{T} \to \mathbb{R}$ by

$$\eta(t) = \exp^{\int_{t_0}^{t} \beta(\tau)\,\mathrm{d}\tau}.$$

A direct computation gives

$$\frac{\mathrm{d}\eta}{\mathrm{d}t}(t) = \beta(t)\eta(t), \qquad t \in \mathbb{T}.$$

Noting that $\eta(t) > 0$ for every $t \in \mathbb{T}$, we then have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\xi(t)}{\eta(t)}\right) = \frac{\eta(t)\frac{\mathrm{d}}{\mathrm{d}t}\xi(t) - \xi(t)\frac{\mathrm{d}}{\mathrm{d}t}\eta(t)}{\eta(t)^2} = \frac{1}{\eta(t)}\left(\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) - \beta(t)\xi(t)\right) \leq 0$$

for $t \geq t_0$. Thus we have

$$\frac{\xi(t)}{\eta(t)} \leq \frac{\xi(t_0)}{\eta(t_0)} = \xi(t_0), \qquad t \geq t_0.$$

This gives the rightmost inequality in the statement of the lemma. The leftmost inequality follows from this by replacing "ξ" with "$-\xi$" and "β" with "α." ▼

3 Lemma *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T} \times \mathrm{U} \to \mathbb{R}^{\mathrm{n}}$$

and let $\mathbf{x}_0 \in \mathrm{U}$. *If there exists* $\mathrm{L} \in \mathbb{R}_{>0}$ *such that*

$$\|\widehat{\mathbf{F}}(\mathrm{t}, \mathbf{x})\| \leq \mathrm{L}\|\mathbf{x} - \mathbf{x}_0\|, \qquad (\mathrm{t}, \mathbf{x}) \in \mathbb{T} \times \mathrm{U},$$

then, for $(\mathrm{t}_0, \mathbf{x}) \in \mathbb{T} \times \mathrm{U}$, *the solution* ξ *to the initial value problem*

$$\dot{\xi}(\mathrm{t}) = \widehat{\mathbf{F}}(\mathrm{t}, \xi(\mathrm{t})), \quad \xi(\mathrm{t}_0) = \mathbf{x},$$

satisfies

(i) $|\frac{\mathrm{d}}{\mathrm{dt}}\|\xi(\mathrm{t}) - \mathbf{x}_0\|^2| \leq 2\mathrm{L}\|\xi(\mathrm{t}) - \mathbf{x}_0\|^2$, $\mathrm{t} \geq \mathrm{t}_0$, *and*

(ii) $\|\mathbf{x} - \mathbf{x}_0\|\mathrm{e}^{-\mathrm{L}(\mathrm{t}-\mathrm{t}_0)} \leq \|\xi(\mathrm{t}) - \mathbf{x}_0\| \leq \|\mathbf{x} - \mathbf{x}_0\|\mathrm{e}^{\mathrm{L}(\mathrm{t}-\mathrm{t}_0)}$, $\mathrm{t} \geq \mathrm{t}_0$.

Proof We compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\xi(t) - x_0\|^2 = 2\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\xi(t), \xi(t) - x_0 \right\rangle_{\mathbb{R}^n}.$$

Thus, by the Cauchy–Bunyakovsky–Schwarz inequality,

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}\|\xi(t) - x_0\|^2\right| &\leq 2\left\|\frac{\mathrm{d}}{\mathrm{d}t}\xi(t)\right\| \|\xi(t) - x_0\| \\ &= 2\|\widehat{F}(t, \xi(t))\|\|\xi(t) - x_0\| \\ &\leq 2L\|\xi(t) - x_0\|^2,\end{aligned}$$

giving the first part of the result.

For the second part, we first note that, from the first part of the lemma,

$$-2L\|\xi(t) - x_0\|^2 \leq \frac{\mathrm{d}}{\mathrm{d}t}\|\xi(t) - x_0\|^2 \leq 2L\|\xi(t) - x_0\|^2.$$

The second part of the current lemma follows from Lemma 2. ▼

Let $r_0 < \min\{\delta, \frac{r}{L}\}$ and let $h = \frac{\ln(2k^2)}{2\sigma}$. Define

$$V(t, x) = \int_t^{t+h} \|\Phi^F(\tau, t, x) - x_0\|^2 \,\mathrm{d}\tau.$$

Then we have

$$V(t, x) \leq L^2\|x - x_0\|^2 \int_t^{t+h} \mathrm{e}^{-2\sigma(\tau-t)} \,\mathrm{d}\tau = \frac{L^2\|x - x_0\|^2(1 - \mathrm{e}^{-2\sigma h})}{2\sigma}.$$

By Lemma 3 we have

$$\|\Phi^F(\tau, t, x) - x_0\|^2 \geq \mathrm{e}^{-2L(t-\tau)}\|x - x_0\|^2,$$

from which we conclude that

$$V(t,x) \geq \|x - x_0\|^2 \int_t^{t+h} \mathrm{e}^{-2L(t-\tau)}\,\mathrm{d}\tau = \frac{\|x - x_0\|^2(1 - \mathrm{e}^{-2Lh})}{2L}.$$

Taking

$$C_2 = \min\left\{\frac{L^2(1 - \mathrm{e}^{-2\sigma h})}{2\sigma}, \frac{1 - \mathrm{e}^{-2Lh}}{2L}\right\}$$

gives condition (iii).

By Theorem 5.1.8, solutions depend continuously differentiably on initial condition and time. Therefore, by , we can differentiate V under the integral sign: switch integral and derivative

$$\frac{\partial V}{\partial t}(t,x) = \|\Phi^F(t+h,t,x) - x_0\|^2 - \|\Phi^F(t,t,x)\|^2 + 2\int_t^{t+h} \left\langle \Phi^F(\tau,t,x) - x_0, \frac{\mathrm{d}}{\mathrm{d}t}\Phi^F(\tau,t,x)\right\rangle \mathrm{d}\tau$$

and

$$\frac{\partial V}{\partial x_j}(t,x) = 2\int_t^{t_h} \left\langle \Phi^F(\tau,t,x) - x_0, \frac{\partial}{\partial x_j}\Phi^F(\tau,t,x)\right\rangle \mathrm{d}\tau.$$

By Exercise 3.2.6 and the preceding two equations we then deduce that

$$\begin{aligned}\mathscr{L}_F V(t,x) &= \|\Phi^F(t+h,t,x) - x_0\|^2 - \|x - x_0\|^2\\ &\leq -(1 - L^2\mathrm{e}^{-2\sigma h})\|x - x_0\|^2 \leq -\frac{1}{2}\|x - x_0\|^2,\end{aligned}$$

giving condition (iv).

Now we note, by the Chain Rule, that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \Phi_j^F}{\partial x_k}(\tau,t,x)\right) = \sum_{l=1}^n \frac{\partial \widehat{F}_j}{\partial x_l}(\tau,t,x)\frac{\partial \Phi_l^F}{\partial x_k}(\tau,t,x), \qquad j \in \{1,\dots,n\},$$

and that

$$\frac{\partial \Phi_j^F}{\partial x_k}(t,t,x) = \begin{cases}1, & j = k,\\ 0, & j \neq k.\end{cases}$$

That is to say, the Jacobian matrix of Φ^F satisfies a linear ordinary differential equation with initial condition being the identity matrix. We wish to use Lemma 3 with x_0 being the zero matrix and $x = I_n$. To do so, we need to estimate the right-hand side of the

preceding equation:

$$\begin{aligned}
\left(\sum_{j,k=1}^{n}\left(\sum_{l=1}^{n}\frac{\partial\widehat{F}_j}{\partial x_l}(\tau,t,\boldsymbol{x})\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right)^2\right)^{1/2} &\leq \left(\sum_{j,k=1}^{n}\left(\sum_{l=1}^{n}\left|\frac{\partial\widehat{F}_j}{\partial x_l}(\tau,t,\boldsymbol{x})\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|\right)^2\right)^{1/2} \\
&\leq \left(\sum_{j,k=1}^{n}\left(M\sum_{l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|\right)^2\right)^{1/2} \\
&\leq \left(M^2\sum_{j,k=1}^{n}\left(\sum_{l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|\right)^2\right)^{1/2} \\
&\leq \left(M^2\sum_{j,k=1}^{n}\sum_{l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|^2\right)^{1/2} \\
&\leq \left(M^2 n\sum_{k,l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|^2\right)^{1/2} \\
&\leq M\sqrt{n}\left(\sum_{k,l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|^2\right)^{1/2}.
\end{aligned}$$

Here we have used the hypotheses on $\widehat{\boldsymbol{F}}$. Now we can use Lemma 3 to conclude that

$$\left(\sum_{j,k=1}^{n}\left|\frac{\partial\Phi_j^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|^2\right)^{1/2} \leq \sqrt{n}\mathrm{e}^{M\sqrt{n}(\tau-t)}.$$

Therefore,

$$\left(\sum_{k=1}^{n}\left(\frac{\partial\Phi_k^{\boldsymbol{F}}}{\partial x_j}(\tau,t,\boldsymbol{x})\right)^2\right)^{1/2} \leq \sqrt{n}\mathrm{e}^{M\sqrt{n}(\tau-t)}.$$

Thus, using the Cauchy–Bunyakovsky–Schwarz inequality,

$$\begin{aligned}
\left|\frac{\partial V}{\partial x_j}(t,\boldsymbol{x})\right| &\leq 2\int_t^{t+h}\left|\left\langle\Phi^{\boldsymbol{F}}(\tau,t,\boldsymbol{x})-\boldsymbol{x}_0,\frac{\partial}{\partial x_j}\Phi^{\boldsymbol{F}}(\tau,t,\boldsymbol{x})\right\rangle\right|\,\mathrm{d}\tau \\
&\leq 2L\sqrt{n}\|\boldsymbol{x}-\boldsymbol{x}_0\|\int_t^{t+h}\mathrm{e}^{-\sigma(\tau-t)}\mathrm{e}^{M\sqrt{n}(\tau-t)}\,\mathrm{d}\tau,
\end{aligned}$$

giving condition (ii). ∎

10.10.2 Converse theorems for autonomous equations

We now consider converse theorems for autonomous ordinary differential equations. The results essentially follow from those of the preceding section, but here we state and prove them independently for readers not needing to deal with time-varying equations.

10.10.3 Theorem (A converse theorem for autonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U} &\to \mathbb{R}^n\\ (t,x) &\mapsto \widehat{\mathbf{F}}_0(x),\end{aligned}$$

and let $x_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$*. Assume that* $\sup \mathbb{T} = \infty$, $T_- \triangleq \inf \mathbb{T} > -\infty$, *and that* $\mathbf{F}$ *satisfies Assumption 10.2.1. If* x_0 *is asymptotically stable, then there exists* $\mathrm{V}\colon \mathbb{T}\times \mathrm{U} \to \mathbb{R}$ *such that*

(i) V *is of class* C^1,

(ii) $\mathrm{V} \in \mathrm{LPD}_{s_0}(x_0)$,

(iii) $\mathrm{V} \in \mathrm{LD}_{s_0}(x_0)$,

(iv) $(t,x) \mapsto \frac{\partial \mathrm{V}}{\partial x_j}(t,x)$ *is in* $\mathrm{LD}_{s_0}(x_0)$, *and*

(v) $-\mathscr{L}_{\mathbf{F}}\mathrm{V} \in \mathrm{LPD}_{s_0}(x_0)$.

10.10.4 Theorem (A converse theorem for exponential stability of autonomous ordinary differential equations) *Let* $\mathbf{F}$ *be an ordinary differential equation with right-hand side*

$$\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U} \to \mathbb{R}^n$$

and let $x_0 \in \mathrm{U}$ *be an equilibrium point for* $\mathbf{F}$*. Assume that* $\sup \mathbb{T} = \infty$, $T_- \triangleq \inf \mathbb{T} > -\infty$, *and that there exists* $M, r \in \mathbb{R}_{>0}$ *such that*

$$\left|\frac{\partial \widehat{\mathrm{F}}_{0,j}}{\partial x_k}(x)\right| \le M, \qquad j,k \in \{1,\dots,n\},\ x \in \mathsf{B}(r,x_0).$$

If there exist $L, \delta, \sigma \in \mathbb{R}_{>0}$ *such that, if* $x \in \mathrm{U}$ *satisfies* $\|x - x_0\| < \delta$, *then* $t \mapsto \Phi^{\mathbf{F}}(t,t_0,x_0)$ *is defined on* $[t_0,\infty)$ *and satisfies*

$$\|\Phi^{\mathbf{F}}(t,t_0,x) - x_0\| \le L e^{-\sigma(t-t_0)}\|x - x_0\|,$$

then there exist $\mathrm{V}\colon \mathrm{U} \to \mathbb{R}$ *and* $r_0 \in \mathbb{R}_{>0}$ *such that*

(i) V *is of class* C^1;

(ii) there exists $C_1 \in \mathbb{R}_{>0}$ *such that*

$$\left\|\frac{\partial \mathrm{V}}{\partial x_j}(x)\right\| \le C_1\|x - x_0\|, \qquad j \in \{1,\dots,n\},\ x \in \mathsf{B}(r_0,x_0);$$

(iii) there exists $C_2 \in \mathbb{R}_{>0}$ *such that*

$$C_2\|x - x_0\|^2 \le \mathrm{V}(x) \le C_2^{-1}\|x - x_0\|^2, \qquad x \in \mathsf{B}(r_0,x_0);$$

(iv) there exists $C_3 \in \mathbb{R}_{>0}$ *such that*

$$\mathscr{L}_{\mathbf{F}}\mathrm{V}(x) \le -C_3\|x - x_0\|^2, \qquad x \in \mathsf{B}(r_0,x_0).$$

Proof We start with a few technical lemmata.

1 Lemma *Let* $\mathbf{F}$ *be an autonomous ordinary differential equation with right-hand side*

$$\begin{aligned}\widehat{\mathbf{F}}\colon \mathbb{T}\times \mathrm{U} &\to \mathbb{R}^{\mathrm{n}}\\ (\mathrm{t},\mathbf{x}) &\mapsto \mathbf{F}_0(\mathbf{x})\end{aligned}$$

and let $\mathbf{x}_0 \in \mathrm{U}$. *If there exists* $\mathrm{L} \in \mathbb{R}_{>0}$ *such that*

$$\|\widehat{\mathbf{F}}_0(\mathbf{x})\| \le \mathrm{L}\|\mathbf{x}-\mathbf{x}_0\|, \qquad \mathbf{x}\in \mathrm{U},$$

then, for $\mathrm{t}_0, \mathbf{x}) \in \mathbb{T}\times \mathrm{U}$,

(i) $|\frac{\mathrm{d}}{\mathrm{dt}}\|\Phi^{\mathbf{F}}(\mathrm{t},\mathrm{t}_0,\mathbf{x})-\mathbf{x}_0\|^2| \le 2\mathrm{L}\|\Phi^{\mathbf{F}}(\mathrm{t},\mathrm{t}_0,\mathbf{x})-\mathbf{x}_0\|^2$, $\mathrm{t}\ge \mathrm{t}_0$, *and*

(ii) $\|\mathbf{x}-\mathbf{x}_0\|\mathrm{e}^{-\mathrm{L}(\mathrm{t}-\mathrm{t}_0)} \le \|\Phi^{\mathbf{F}}(\mathrm{t},\mathrm{t}_0,\mathbf{x})-\mathbf{x}_0\| \le \|\mathbf{x}-\mathbf{x}_0\|\mathrm{e}^{\mathrm{L}(\mathrm{t}-\mathrm{t}_0)}$, $\mathrm{t}\ge \mathrm{t}_0$.

Proof We compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\Phi^F(t,t_0,x)-x_0\|^2 = 2\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\Phi^F(t,t_0,x), \Phi^F(t,t_0,x)-x_0\right\rangle_{\mathbb{R}^n}.$$

Thus, by the Cauchy–Bunyakovsky–Schwarz inequality,

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}\|\Phi^F(t,t_0,x)-x_0\|^2\right| &\le 2\left\|\frac{\mathrm{d}}{\mathrm{d}t}\Phi^F(t,t_0,x)\right\|\|\Phi^F(t,t_0,x)-x_0\|\\ &= 2\|\widehat{F}(t,\Phi^F(t,t_0,x))\|\|\Phi^F(t,t_0,x)-x_0\|\\ &\le 2L\|\Phi^F(t,t_0,x)-x_0\|^2,\end{aligned}$$

giving the first part of the result.

For the second part, we first note that, from the first part of the lemma,

$$-2L\|\Phi^F(t,t_0,x)-x_0\|^2 \le \frac{\mathrm{d}}{\mathrm{d}t}\|\Phi^F(t,t_0,x)-x_0\|^2 \le 2L\|\Phi^F(t,t_0,x)-x_0\|^2.$$

The second part of the current lemma follows from Lemma 2. ▼

Let $r_0 < \min\{\delta, \frac{r}{L}\}$ and let $h = \frac{\ln(2k^2)}{2\sigma}$. For some $t\in\mathbb{T}$, define

$$V(x) = \int_t^{t+h} \|\Phi^F(\tau,t,x)-x_0\|^2\,\mathrm{d}\tau.$$

Note that $V(x)$ is independent of t by Exercise 3.1.19. Then we have

$$V(x) \le L^2\|x-x_0\|^2\int_t^{t+h}\mathrm{e}^{-2\sigma(\tau-t)}\,\mathrm{d}\tau = \frac{L^2\|x-x_0\|^2(1-\mathrm{e}^{-2\sigma h})}{2\sigma}.$$

By Lemma 1 we have

$$\|\Phi^F(\tau,t,x)-x_0\|^2 \ge \mathrm{e}^{-2L(t-\tau)}\|x-x_0\|^2,$$

from which we conclude that

$$V(x) \ge \|x-x_0\|^2\int_t^{t+h}\mathrm{e}^{-2L(t-\tau)}\,\mathrm{d}\tau = \frac{\|x-x_0\|^2(1-\mathrm{e}^{-2Lh})}{2L}.$$

Taking

$$C_2 = \min\left\{\frac{L^2(1-\mathrm{e}^{-2\sigma h})}{2\sigma}, \frac{1-\mathrm{e}^{-2Lh}}{2L}\right\}$$

gives condition (iii).

By Theorem 5.1.8, solutions depend continuously differentiably on initial condition and time. Therefore, by , we can differentiate V under the integral sign:

switch integral and derivative

$$\frac{\partial V}{\partial x_j}(\boldsymbol{x}) = 2\int_t^{t+h} \left\langle \Phi^{\boldsymbol{F}}(\tau,t,\boldsymbol{x}) - \boldsymbol{x}_0, \frac{\partial}{\partial x_j}\Phi^{\boldsymbol{F}}(\tau,t,\boldsymbol{x})\right\rangle \mathrm{d}\tau.$$

By Exercise 3.2.6 and the preceding two equations we then deduce that

$$\begin{aligned}\mathscr{L}_{\boldsymbol{F}}V(\boldsymbol{x}) &= \|\Phi^{\boldsymbol{F}}(t+h,t,\boldsymbol{x}) - \boldsymbol{x}_0\|^2 - \|\boldsymbol{x}-\boldsymbol{x}_0\|^2\\ &\le -(1-L^2\mathrm{e}^{-2\sigma h})\|\boldsymbol{x}-\boldsymbol{x}_0\|^2 \le -\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|^2,\end{aligned}$$

giving condition (iv).

Now we note, by the Chain Rule, that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \Phi_j^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right) = \sum_{l=1}^n \frac{\partial \widehat{F}_j}{\partial x_l}(\tau,t,\boldsymbol{x})\frac{\partial \Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x}), \qquad j\in\{1,\dots,n\},$$

and that

$$\frac{\partial \Phi_j^{\boldsymbol{F}}}{\partial x_k}(t,t,\boldsymbol{x}) = \begin{cases}1, & j=k,\\ 0, & j\neq k.\end{cases}$$

That is to say, the Jacobian matrix of $\Phi^{\boldsymbol{F}}$ satisfies a linear ordinary differential equation with initial condition being the identity matrix. We wish to use Lemma 1 with $\boldsymbol{x}_0$ being the zero matrix and $\boldsymbol{x} = \boldsymbol{I}_n$. To do so, we need to estimate the right-hand side of the

preceding equation:

$$\begin{aligned}
\left(\sum_{j,k=1}^{n}\left(\sum_{l=1}^{n}\frac{\partial\widehat{F}_j}{\partial x_l}(\tau,t,\boldsymbol{x})\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right)^2\right)^{1/2} &\leq \left(\sum_{j,k=1}^{n}\left(\sum_{l=1}^{n}\left|\frac{\partial\widehat{F}_j}{\partial x_l}(\tau,t,\boldsymbol{x})\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|\right)^2\right)^{1/2} \\
&\leq \left(\sum_{j,k=1}^{n}\left(M\sum_{l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|\right)^2\right)^{1/2} \\
&\leq \left(M^2\sum_{j,k=1}^{n}\left(\sum_{l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|\right)^2\right)^{1/2} \\
&\leq \left(M^2\sum_{j,k=1}^{n}\sum_{l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|^2\right)^{1/2} \\
&\leq \left(M^2 n\sum_{k,l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|^2\right)^{1/2} \\
&\leq M\sqrt{n}\left(\sum_{k,l=1}^{n}\left|\frac{\partial\Phi_l^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|^2\right)^{1/2}.
\end{aligned}$$

Here we have used the hypotheses on $\widehat{\boldsymbol{F}}$. Now we can use Lemma 1 to conclude that

$$\left(\sum_{j,k=1}^{n}\left|\frac{\partial\Phi_j^{\boldsymbol{F}}}{\partial x_k}(\tau,t,\boldsymbol{x})\right|^2\right)^{1/2} \leq \sqrt{n}\mathrm{e}^{M\sqrt{n}(\tau-t)}.$$

Therefore,

$$\left(\sum_{k=1}^{n}\left(\frac{\partial\Phi_k^{\boldsymbol{F}}}{\partial x_j}(\tau,t,\boldsymbol{x})\right)^2\right)^{1/2} \leq \sqrt{n}\mathrm{e}^{M\sqrt{n}(\tau-t)}.$$

Thus, using the Cauchy–Bunyakovsky–Schwarz inequality,

$$\begin{aligned}
\left|\frac{\partial V}{\partial x_j}(t,\boldsymbol{x})\right| &\leq 2\int_t^{t+h}\left|\left\langle\Phi^{\boldsymbol{F}}(\tau,t,\boldsymbol{x})-\boldsymbol{x}_0,\frac{\partial}{\partial x_j}\Phi^{\boldsymbol{F}}(\tau,t,\boldsymbol{x})\right\rangle\right|\,\mathrm{d}\tau \\
&\leq 2L\sqrt{n}\|\boldsymbol{x}-\boldsymbol{x}_0\|\int_t^{t+h}\mathrm{e}^{-\sigma(\tau-t)}\mathrm{e}^{M\sqrt{n}(\tau-t)}\,\mathrm{d}\tau,
\end{aligned}$$

giving condition (ii). ■

10.10.3 Converse theorem for time-varying linear equations

Next we turn to converse results for linear ordinary differential equations. The first is for time-varying equations.

10.10.5 Theorem (A converse theorem for time-varying linear ordinary differential equations) *Let* F *be a system of linear homogeneous ordinary differential equations in an* n*-dimensional* $\mathbb{R}$*-vector space* V *with constant coefficients and with right-hand side*

$$\begin{aligned}\widehat{F}\colon \mathbb{T}\times V &\to V\\ (t,x) &\mapsto A(t)(x)\end{aligned}$$

for $A\colon \mathbb{T}\to L(V;V)$ *continuous and bounded. Suppose that* $\sup\mathbb{T}=\infty$. *Suppose that* V *has an inner product* $\langle\cdot,\cdot\rangle$. *Let* $Q\colon \mathbb{T}\to L(V;V)$ *have the following properties:*

(i) Q *is continuous;*

(ii) $Q(t)$ *is symmetric for every* $t\in\mathbb{T}$*;*

(iii) Q *is positive-definite;*

(iv) Q *is decrescent.*

Then there exists $P\colon \mathbb{T}\to L(V;V)$ *with the following properties:*

(i) P *is of class* C^1*;*

(ii) $P(t)$ *is symmetric for every* $t\in\mathbb{T}$*;*

(iii) (P,Q) *is a Lyapunov pair for* F*;*

(iv) P *is positive-definite;*

(v) P *is decrescent.*

Proof By Exercise 10.3.2(f), let $C_1,\sigma\in\mathbb{R}_{>0}$ be such that

$$|||\Phi_A(t,t_0)|||\le C_1 e^{-\sigma(t-t_0)},\qquad t\in\mathbb{T},\ t\ge t_0. \tag{10.37}$$

By Lemma 10.6.17, there exists $C_2\in\mathbb{R}_{>0}$ such that

$$C_2\langle x,x\rangle\le f_Q(t,x)\le C_2^{-1}\langle x,x\rangle,\qquad (t,x)\in\mathbb{T}\times V. \tag{10.38}$$

We define

$$P(t)=\int_t^\infty \Phi_A(\tau,t)^T\circ Q(\tau)\circ\Phi_A(\tau,t)\,d\tau.$$

The integral exists by the inequalities (10.37) and (10.38).

For $(t,x)\in\mathbb{T}\times V$ we compute

$$\begin{aligned}f_P(t,x)&=\int_t^\infty f_Q(\tau,\Phi_A(\tau,t)(x))\,d\tau\\ &\le C_2^{-1}\int_t^\infty \|\Phi_A(\tau,t)(x)\|^2\,d\tau\\ &\le C_2^{-1}\|x\|^2\int_t^\infty |||\Phi_A(\tau,t)|||^2\,d\tau\\ &\le \frac{C_1}{C_2}\|x\|^2\int_t^\infty e^{-\sigma(\tau-t)}\,d\tau=\frac{C_1}{C_2\sigma}\|x\|^2.\end{aligned}$$

Since A is bounded, there exists $M \in \mathbb{R}_{>0}$ such that $|||A(t)||| \leq M$ for each $t \in \mathbb{T}$, by Lemma 1 from the proof of Theorem 10.10.2 we have

$$\|\Phi_A(\tau, t)(x)\|^2 \geq \|x\|^2 \mathrm{e}^{-2M(\tau - t)}, \qquad \tau \geq t.$$

Therefore,

$$\begin{aligned} f_P(t, x) &= \int_t^\infty f_Q(\tau, \Phi_A(\tau, t)(x)) \, \mathrm{d}\tau \\ &\geq C_2 \int_t^\infty \|\Phi_A(\tau, t)(x)\|^2 \, \mathrm{d}\tau \\ &\geq C_2 \|x\|^2 \int_t^\infty \mathrm{e}^{-2M(\tau - t)} \, \mathrm{d}\tau = \frac{C_2}{2M} \|x\|^2. \end{aligned}$$

Letting $C = \min\{\frac{C_2}{2M}, \frac{C_2 \sigma}{C_1}\}$, we thus have

$$C\langle x, x\rangle \leq f_P(t, x) \leq C^{-1}\langle x, x\rangle,$$

showing that P is positive-definite and decrescent, by Lemma 10.6.17.

By the Fundamental Theorem of Calculus, P is continuously differentiable. By (5.6) we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \Phi_A(\tau, t) = -\Phi_A(\tau, t) \circ A(t).$$

Thus

$$\begin{aligned} \dot{P}(t) = &- Q(t) + \int_t^\infty \left(\frac{\mathrm{d}}{\mathrm{d}t} \Phi_A(\tau, t)^T \right) \circ Q(\tau) \circ \Phi_A(\tau, t) \, \mathrm{d}\tau \\ &+ \int_t^\infty \Phi_A(\tau, t)^T \circ Q(\tau) \circ \left(\frac{\mathrm{d}}{\mathrm{d}t} \Phi_A(\tau, t) \right) \mathrm{d}\tau \\ = &- Q(t) - \int_t^\infty A(t)^T \circ \Phi_A(\tau, t)^T \circ Q(\tau) \circ \Phi_A(\tau, t) \, \mathrm{d}\tau \\ &- \int_t^\infty \Phi_A(\tau, t)^T \circ Q(\tau) \circ \Phi_A(\tau, t) \circ A(t) \, \mathrm{d}\tau \\ = &- Q(t) - A(t)^T \circ P(t) - P(t) \circ A(t), \end{aligned}$$

which shows that (P, Q) is a Lyapunov pair for F, as desired. ■

10.10.4 Converse theorem for linear equations with constant coefficients

Finally, we give a result for linear ordinary differential equations with constant coefficients. Here the results we give are quite detailed, in keeping with our detailed knowledge of such equations.

10.10.6 Theorem (A converse theorem for linear ordinary differential equations with constant coefficients) *Let* F *be a system of linear homogeneous ordinary differential*

equations in an n*-dimensional* $\mathbb{R}$*-vector space* V *with constant coefficients and with right-hand side*

$$\begin{aligned}\widehat{\mathrm{F}}\colon \mathbb{T}\times\mathsf{V} &\to \mathsf{V}\\ (\mathrm{t},\mathrm{x}) &\mapsto \mathrm{A}(\mathrm{x})\end{aligned}$$

for $\mathrm{A}\in\mathrm{L}(\mathsf{V};\mathsf{V})$. *Suppose that* $\sup\mathbb{T}=\infty$. *Suppose that* V *has an inner product* $\langle\cdot,\cdot\rangle$. *If* A *is Hurwitz, then the following statements hold:*

(i) for any symmetric $\mathrm{Q}\in\mathrm{L}(\mathsf{V};\mathsf{V})$*, there exists a unique symmetric* $\mathrm{P}\in\mathrm{L}(\mathsf{V};\mathsf{V})$ *so that* (P,Q) *is a Lyapunov pair for* F*;*

(ii) if Q *is positive-semidefinite with* P *the unique symmetric linear map for which* (P,Q) *is a Lyapunov pair for* F*, then* P *is positive-semidefinite;*

(iii) if Q *is positive-semidefinite with* P *the unique symmetric linear map for which* (P,Q) *is a Lyapunov pair for* F*, then* P *is positive-definite if and only if* (A,Q) *is observable.*

Proof (i) We claim that, if we define

$$P=\int_0^\infty \mathrm{e}^{A^Tt}\circ Q\circ\mathrm{e}^{At}\,\mathrm{d}t, \tag{10.39}$$

then (P,Q) is a Lyapunov pair for F. First note that since A is Hurwitz, the integral does indeed converge by . We also have ref

$$\begin{aligned}A^T\circ P+P\circ A&=A^T\circ\left(\int_0^\infty \mathrm{e}^{A^Tt}\circ Q\circ\mathrm{e}^{At}\,\mathrm{d}t\right)+\left(\int_0^\infty \mathrm{e}^{A^Tt}\circ Q\circ\mathrm{e}^{At}\,\mathrm{d}t\right)\circ A\\&=\int_0^\infty\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{A^Tt}\circ Q\circ\mathrm{e}^{At}\right)\mathrm{d}t\\&=\mathrm{e}^{A^Tt}\circ Q\circ\mathrm{e}^{At}\Big|_0^\infty=-Q,\end{aligned}$$

as desired. We now show that P as defined is the *only* symmetric linear map for which (P,Q) is a Lyapunov pair for F. Suppose that $\hat{P}$ also has the property that $(\hat{P},Q)$ is a Lyapunov pair for F, and let $\Delta=\hat{P}-P$. Then one sees that

$$A^T\circ\Delta+\Delta\circ A=0.$$

If we let

$$\Lambda(t)=\mathrm{e}^{A^Tt}\circ\Delta\circ\mathrm{e}^{At},$$

then

$$\frac{\mathrm{d}\Lambda}{\mathrm{d}t}(t)=\mathrm{e}^{A^Tt}\circ\left(A^T\circ\Delta+\Delta\circ A\right)\circ\mathrm{e}^{At}=0.$$

Therefore, Λ is constant, and since $\Lambda(0)=\Delta$, it follows that $\Lambda(t)=\Delta$ for all t. However, since A is Hurwitz, it also follows that $\lim_{t\to\infty}\Lambda(t)=0$. Thus $\Delta=0$, so that $\hat{P}=P$.

(ii) If P is defined by (10.39), then we have

$$f_P(x)=\int_0^\infty\langle Q\circ\mathrm{e}^{At}(x),\mathrm{e}^{At}(x)\rangle\,\mathrm{d}t.$$

Therefore, if Q is positive-semidefinite, it follows that P is positive-semidefinite.

(iii) Here we employ a lemma.

1 Lemma *If* Q *is positive-semidefinite then* (A, Q) *is observable if and only if the linear map* P *defined by (10.39) is invertible.*

Proof First suppose that (A, Q) is observable and let $x \in \ker(P)$. Then

$$\int_0^\infty \langle Q \circ \mathrm{e}^{At}(x), \mathrm{e}^{At}(x) \rangle \, \mathrm{d}t = 0.$$

Since Q is positive-semidefinite, this implies that $\mathrm{e}^{At}(x) \in \ker(Q)$ for all t. Differentiating this inclusion k times with respect to t gives $A^k \circ \mathrm{e}^{At}(x) \in \ker(Q)$ for any $k \in \mathbb{Z}_{>0}$. Evaluating at $t = 0$ shows that $x \in \ker(O(A, C))$. Since (A, Q) is observable, this implies that $x = 0$. Thus we have shown that $\ker(P) = \{0\}$, or equivalently that P is invertible.

Now suppose that P is invertible. Then the expression

$$\int_0^\infty \langle Q \circ \mathrm{e}^{At}(x), \mathrm{e}^{At}(x) \rangle \, \mathrm{d}t$$

is zero if and only if $x = 0$. Since Q is positive-semidefinite, this means that the expression

$$\langle Q \circ \mathrm{e}^{At}(x), \mathrm{e}^{At}(x) \rangle$$

is zero if and only if $x = 0$. Since e^{At} is invertible, this implies that Q must be positive-definite, and in particular, invertible. In this case, (A, Q) is clearly observable. ▼

With the lemma at hand, the remainder of the proof is straightforward. Indeed, from part (ii), we know that P is positive-semidefinite. The lemma now says that P is positive-definite if and only if (A, Q) is observable, as desired. ■

Let us resume our example started as Example 10.7.23.

10.10.7 Example (Example 10.7.23 cont'd) We resume looking at the case where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix}.$$

Let us look at a few cases to flesh out some aspects of Theorem 10.10.6.

1. $a > 0$ and $b > 0$: This is exactly the case when $\boldsymbol{A}$ is Hurwitz, so that part (i) of Theorem 10.10.6 implies that, for any symmetric $\boldsymbol{Q}$, there is a unique symmetric $\boldsymbol{P}$ so that $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$. As we saw in the proof of Theorem 10.10.6, one can determine $\boldsymbol{P}$ with the formula

$$\boldsymbol{P} = \int_0^\infty \mathrm{e}^{\boldsymbol{A}^T t} \boldsymbol{Q} \mathrm{e}^{\boldsymbol{A}t} \, \mathrm{d}t. \tag{10.40}$$

However, to do this in this example is a bit tedious since we would have to deal with the various cases of a and b to cover all the various forms taken by $\mathrm{e}^{\boldsymbol{A}t}$. For example, suppose we take

$$\boldsymbol{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and let $a = 2$ and $b = 2$. Then we have

$$e^t = e^{-t}\begin{bmatrix} \cos t + \sin t & \sin t \\ -2\sin t & \cos t - \sin t \end{bmatrix}$$

In this case one can directly apply (10.40) with some effort to get

$$\boldsymbol{P} = \begin{bmatrix} \frac{5}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{8} \end{bmatrix}.$$

If we let $a = 2$ and $b = 1$ then we compute

$$e^{At} = e^{-t}\begin{bmatrix} 1 + t & t \\ -t & 1 - t \end{bmatrix}.$$

Again, a direct computation using (10.40) gives

$$\boldsymbol{P} = \begin{bmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Note that our choice of $\boldsymbol{Q}$ is positive-definite and that $(\boldsymbol{A}, \boldsymbol{Q})$ is, therefore, observable. Therefore, part (iii) of Theorem 10.10.6 implies that $\boldsymbol{P}$ is positive-definite. It may be verified that the $\boldsymbol{P}$'s computed above are indeed positive-definite.

However, it is not necessary to make such hard work of this. After all, the equation

$$\boldsymbol{A}^T\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{Q}$$

is nothing but a linear equation for $\boldsymbol{P}$. That $\boldsymbol{A}$ is Hurwitz merely ensures a unique solution for any symmetric $\boldsymbol{Q}$. If we denote

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}$$

and continue to use

$$\boldsymbol{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

then we must solve the linear equations

$$\begin{bmatrix} 0 & -b \\ 1 & -a \end{bmatrix}\begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$

subject to $a, b > 0$. One can then determine $\boldsymbol{P}$ for general (at least nonzero) a and b to be

$$\boldsymbol{P} = \begin{bmatrix} \frac{a^2+b+b^2}{2ab} & \frac{1}{2b} \\ \frac{1}{2b} & \frac{b+1}{2ab} \end{bmatrix}.$$

In this case, we are guaranteed that this is the unique $\boldsymbol{P}$ that does the job.

2. $a \leq 0$ and $b = 0$: As we have seen, in this case there is not always a solution to the equation
$$\boldsymbol{A}^T\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{Q}. \tag{10.41}$$
Indeed, when $\boldsymbol{Q}$ is positive-semidefinite and $(\boldsymbol{A}, \boldsymbol{Q})$ is observable, this equation is guaranteed to *not* have a solution (see Exercise 10.7.5). This demonstrates that when $\boldsymbol{A}$ is not Hurwitz, part (i) of Theorem 10.10.6 can fail in the matter of existence.
3. $a > 0$ and $b = 0$: In this case we note that, for any $C \in \mathbb{R}$, the matrix
$$\boldsymbol{P}_0 = C\begin{bmatrix} a^2 & a \\ a & 1 \end{bmatrix}$$
satisfies $\boldsymbol{A}^T\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = \boldsymbol{0}$. Thus, if $\boldsymbol{P}$ is any solution to (10.41), then $\boldsymbol{P} + \boldsymbol{P}_0$ is also a solution. If we take
$$\boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 2a \end{bmatrix},$$
then, as we saw in Theorem 10.7.21, if
$$\boldsymbol{P} = \begin{bmatrix} a^2 & a \\ a & 2 \end{bmatrix},$$
then $(\boldsymbol{P}, \boldsymbol{Q})$ is a Lyapunov pair for $\boldsymbol{F}$. What we have shown is that $(\boldsymbol{P} + \boldsymbol{P}_0, \boldsymbol{Q})$ is also a Lyapunov pair for $\boldsymbol{F}$. Thus part (i) of Theorem 10.10.6 can fail in the matter of uniqueness when $\boldsymbol{A}$ is not Hurwitz. •

This version: 2022/03/07

Chapter 11

Input/output stability

Bibliography

Anderson, B. D. O. [1972] *The reduced Hermite criterion with application to proof of the Liénard–Chipart criterion*, Institute of Electrical and Electronics Engineers. Transactions on Automatic Control, **17**(5), pages 669–672, ISSN: 0018-9286, DOI: 10.1109/TAC.1972.1100142.

Anderson, B. D. O., Jury, E. I., and Mansour, M. [1987] *On robust Hurwitz polynomials*, Institute of Electrical and Electronics Engineers. Transactions on Automatic Control, **32**(10), pages 909–913, ISSN: 0018-9286, DOI: 10.1109/TAC.1987.1104459.

Bacciotti, A. and Rosier, L. [2005] *Liapunov Functions and Stability in Control Theory*, Communications and Control Engineering Series, Springer-Verlag: New York/-Heidelberg/Berlin, ISBN: 978-3-540-21332-1.

Barbashin, E. A. and Krasovskiĭ, N. N. [1952] *On global stability of motion*, Rossiĭskaya Akademiya Nauk. Doklady Akademii Nauk, **86**(3), pages 453–456, ISSN: 0869-5652.

Brown, C. M. [2007] *Differential Equations, A Modeling Approach*, number 150 in Quantitative Applications in the Social Sciences, SAGE Publications: Los Angeles/London/New Delhi/Singapore, ISBN: 978-1-4129-4108-2.

Butterworth, S. [1930] *On the theory of filter amplifiers*, Experimental Wireless & the Wireless Engineer, **7**(85), pages 536–541, URL: https://www.americanradiohistory.com/Archive-Experimental%20Wireless/30s/Wireless-Engineer-1930-10.pdf (visited on 02/21/2019).

Chander, P. [1983] *The nonlinear input–output model*, Journal of Economic Theory, **30**(2), pages 219–229, ISSN: 0022-0531, DOI: 10.1016/0022-0531(83)90105-9.

Chapellat, H. and Bhattacharyya, S. P. [1989] *An alternative proof of Kharitonov's theorem*, Institute of Electrical and Electronics Engineers. Transactions on Automatic Control, **34**(4), pages 448–450, ISSN: 0018-9286, DOI: 10.1109/9.28021.

Dasgupta, S. [1988] *Kharitonov's theorem revisited*, Systems & Control Letters, **11**(5), pages 381–384, ISSN: 0167-6911, DOI: 10.1016/0167-6911(88)90096-5.

Duffy, D. G. [2015] *Green's Functions with Applications*, 2nd edition, Advances in Applied Mathematics, CRC Press: Boca Raton, FL, ISBN: 978-1-4822-5103-6.

Franses, P. H. and van Dijk, D. [2003] *Nonlinear Time Series Models in Empirical Finance*, Cambridge University Press: New York/Port Chester/Melbourne/Sydney, ISBN: 978-0-521-77965-4.

Fujiwara, M. [1915] *Über die Wurzeln der algebraischen Gleichungen*, The Tôhoku Mathematical Journal. Second Series, **8**, pages 78–85, ISSN: 0040-8735.

Gantmacher, F. R. [1959] *The Theory of Matrices*, translated by K. A. Hirsch, volume 2, Chelsea: New York, NY, Reprint: [Gantmacher 2000].

— [2000] *The Theory of Matrices*, translated by K. A. Hirsch, volume 2, American Mathematical Society: Providence, RI, ISBN: 978-0-8218-2664-5, Original: [Gantmacher 1959].

Gates, Jr, L. D. [1956] *Linear differential equations in distributions*, Proceedings of the American Mathematical Society, **7**(5), pages 933–939, ISSN: 0002-9939, DOI: 10.2307/2033565.

Hamming, R. W. [1980] *The unreasonable effectiveness of mathematics*, The American Mathematical Monthly, **87**(2), pages 81–90, ISSN: 0002-9890, DOI: 10.2307/2321982.

Henson, M. A. and Seborg, D. E. [1992] *Nonlinear control strategies for continuous fermenters*, Chemical Engineering Science, **47**(4), pages 821–835, ISSN: 0009-2509, DOI: 10.1016/0009-2509(92)80270-M.

Hermite, C. [1854] *Sur le nombre des racines d'une équation algèbrique comprise entre des limites données*, Journal für die Reine und Angewandte Mathematik, **52**, pages 39–51, ISSN: 0075-4102.

Holmes, P. J. [1982] *The dynamics of repeated impacts with a sinusoidally vibrating table*, Journal of Sound and Vibration, **84**(2), ISSN: 0022-460X, DOI: 10.1016/S0022-460X(82)80002-3.

Hopfield, J. J. [1982] *Neural networks and physical systems with emergent collective computational abilities*, Proceedings of the National Academy of Sciences of the United States of America, **79**(8), pages 2554–2558, ISSN: 1091-6490, DOI: 10.1073/pnas.79.8.2554.

Hurwitz, A. [1895] *Über di Bedingungen unter welchen eine Gleichung nur Wurzeln mit negativen reellen Teilen besitz*, Mathematische Annalen, **46**, pages 273–284, ISSN: 0025-5831, URL: https://eudml.org/doc/157760 (visited on 07/11/2014).

Kellett, C. M. [2014] *A compendium of comparison function results*, Mathematics of Control, Signals, and Systems, **26**(3), pages 339–374, ISSN: 0932-4194.

Keshmiri, M., Jahromi, A. F., Mohebbi, A., Amoozgar, M. H., and Xie, W.-F. [2012] *Modeling and control of ball and beam system using model based and non-model based control approaches*, International Journal on Smart Sensing and Intelligent Systems, **5**(1), pages 14–35, ISSN: 1178-5608, URL: http://s2is.org/Issues/v5/n1/papers/paper2.pdf (visited on 02/17/2019).

Kharitonov, V. L. [1978] *Asymptotic stability of an equilibrium position of a family of systems of linear differential equations*, Differentsial'nye Uravneniya, **14**, pages 2086–2088, ISSN: 0374-0641.

LaSalle, J. P. [1968] *Stability theory for ordinary differential equations*, Journal of Differential Equations, **4**(1), pages 57–65, ISSN: 0022-0396, DOI: 10.1016/0022-0396(68)90048-X.

Liapunov, A. M. [1893] *A special case of the problem of stability of motion*, Rossiĭskaya Akademiya Nauk. Matematicheskiĭ Sbornik, **17**, pages 252–333, ISSN: 0368-8666.

Liénard, A. and Chipart, M. [1914] *Sur la signe de la partie réelle des racines d'une équation algébrique*, Journal de Mathématiques Pures et Appliquées. Neuvième Série, **10**(6), pages 291–346, ISSN: 0021-7824.

Mansour, M. and Anderson, B. D. O. [1993] *Kharitonov's theorem and the second method of Lyapunov*, Systems & Control Letters, **20**(3), pages 39–47, ISSN: 0167-6911, DOI: 10.1016/0167-6911(93)90085-K.

Maxwell, J. C. [1868] *On governors*, Proceedings of the Royal Society. London. Series A. Mathematical and Physical Sciences, **16**, pages 270–283, ISSN: 1364-5021, URL: http://www.jstor.org/stable/112510 (visited on 07/10/2014).

Mesarovic, M. D. and Takahara, Y. [1975] *General Systems Theory, Mathematical Foundations*, number 113 in Mathematics in Science and Engineering, Academic Press: New York, NY, ISBN: 978-0-08-095622-0.

— [1989] *Abstract Systems Theory*, number 116 in Lecture Notes in Control and Information Sciences, Springer-Verlag: New York/Heidelberg/Berlin, ISBN: 978-3-540-50529-7.

Minnichelli, R. J., Anagnost, J. J., and Desoer, C. A. [1989] *An elementary proof of Kharitonov's stability theorem with extensions*, Institute of Electrical and Electronics Engineers. Transactions on Automatic Control, **34**(9), pages 995–998, ISSN: 0018-9286, DOI: 10.1109/9.35816.

Newton, I. [1687] *Philosophiae Naturalis Principia Mathematica*, S. Pepys: London, Translation: [Newton 1995].

— [1995] *Principia*, translated by A. Motte, Prometheus Books: Amherst, NY, ISBN: 978-0-87975-980-3, Original: [Newton 1687].

Nishimura, K. and Stachurski, J. [2004] *Discrete Time Models in Economic Theory*, URL: http://johnstachurski.net/_downloads/discrete.pdf (visited on 02/18/2019).

Parks, P. C. [1962] *A new proof of the Routh–Hurwitz stability criterion using the second method of Liapunov*, Proceedings of the Cambridge Philosophical Society, **58**(4), pages 694–702, DOI: 10.1017/S030500410004072X.

Routh, E. J. [1877] *A Treatise on the Stability of a Given State of Motion*, Adam's Prize Essay, Cambrige University.

Wigner, E. P. [1960] *The unreasonable effectiveness of mathematics in the natural sciences*, Communications on Pure and Applied Mathematics, **13**(1), pages 1–14, ISSN: 0010-3640, DOI: 10.1002/cpa.3160130102.