A Mathematical Approach to Classical Control

Single-input, single-output, time-invariant, continuous time, finite-dimensional, deterministic, linear systems

Andrew D. Lewis

January–April 2003
This version: 03/09/2014

Preface

The purpose of this preface is two-fold: (1) to discuss the philosophy of the approach taken, as it is nonstandard for an introductory course; (2) to discuss the content of the book.

The philosophy

Since this book takes an untraditional approach to introductory control, it is worth outlining why I have taken the approach I have.

The goals

Clearly a new text in classical control needs to have some justification for its appearance, as there are already a large number of texts on the market, and these satisfy the demands made by a typical introductory course in feedback control. The approach in this book is not typical. The idea here is to develop control theory, at an introductory classical level, as a rigorous subject. This is different, note, from presenting the mathematics needed to understand control theory. One will often hear things like, "Classical control is merely an application of complex variable theory," or "Linear control is merely an application of linear algebra." While it is true that these parts of control theory do rely on the asserted branches of mathematics, control theory is such an effective blend of many branches of mathematics that to categorise it as a subset of one is a disservice. The subject of control theory, even at an introductory level, has a mathematical life of its own, and it is this life that is being exhibited here.

The main benefit of such an approach is that not just the mathematics behind the subject, but the subject itself can be treated rigorously. The problems of control theory, and these are actual practical problems, often have precise mathematical statements, and the intent in this book is to give these wherever possible. The result is that a student will be able to understand simple problems in a larger context. For some, at least, this is useful. It also makes it possible to consider challenging control problems that cannot really be considered in an exclusively *ad hoc* treatment. It would seem that many classical control texts were written based upon the standard of control practice in, say, the early 1960's. This practice, well laid out in the texts of Truxal [1955] and Horowitz [1963], had reached a point where, for the problems to which it was applicable, it was "finished." This was expressed in one early paper as follows: "The present state of the art is such that it is safe to assume that, for linear single-loop feedback systems, almost no analysis or design problems of any consequence remain." Such statements are seldom prophetic. Indeed, much has been done since the date of publication of the cited paper (1961), even for linear single-loop systems. Now we have means for handling problems that would be almost impossible to treat using the *ad hoc* methods of classical design. And the methods all rely on a firm grasp of not just the mathematics behind control theory, but the mathematics of the subject itself. This is the reason for this book.

The mathematical approach

With the above as backdrop, this book is provided for students who can be relied upon to have a satisfactory background in linear algebra, differential equations (including the matrix exponential), basic complex analysis, and some transform theory. The appendices contain a quick overview of necessary background material, so that an instructor or a student can determine whether the book is useful.

Apart from the above pedagogical concerns, I have also tried to write the book with an eye towards its being a useful reference. In the book, I have tried to prove as many statements as possible; even many that are not commonly proved, but often stated. I do this not because I feel that all of these proofs should be delivered in lectures—I certainly do not do this myself. Rather, my objectives here are scholarly. I do not feel that such lofty goals clash with the rather more pedantic concerns of getting students to come to grips with basic material. Students who find the course challenging may safely omit consideration of the more technical proofs, provided that they understand the concepts behind the results. More curious students, however, are rewarded by having for reference proofs that can be difficult to find in the literature. Moreover, this approach has, in my experience, a pedagogical byproduct. If one teaches an introductory course in a manner not completely "method oriented," natural questions will arise in the presentation. For example, if one even gets around to posing the problem of finding a controller that stabilises a given plant in a unity gain feedback loop, the natural question arises as to whether such controllers exist. The answer is affirmative, but the determination of this answer is nontrivial. A traditional approach to classical control masks the existence of the question, never mind providing the answer. Again, the advantage of the approach taken here, at least for the curious student, is that the answer to this more basic question may be found alongside the more standard ad hoc methods for controller design.

The rôle of control design

A word needs to be said about control design. Greater emphasis is being placed on engineering design in the engineering undergraduate curriculum, and this is a by all means an appropriate tendency. When teaching a control course, one faces a decision relative to design content. Should the design be integrated into the course at every stage, or should it be separated from the analysis parts of the course? In this book, the trend in engineering education is being bucked, and the latter course is taken. Indeed, care has been taken to explicitly separate the book into three parts, with the design part coming last. One can justly argue that this is a mistake, but it is the approach I have decided upon, and it seems to work. My rationale for adopting the approach I do is that in control, there is very simply a lot of analysis to learn before one can do design in a fulfilling way. Thus I get all the tools in place *before* design is undertaken in the latter stages of the book.

How to use the book

It is not possible to cover all of the topics in this book in a single term; at least it is not advisable to attempt this. However, it is quite easy to break the contents of the book into two courses, one at an introductory level, and another dealing with advanced topics. Because this division is not readily made on a chapter-by-chapter basis, it is perhaps worth suggesting two possible courses that can be taught from this book.

An introductory course for students with *no* control background might contain roughly the following material:

1. Chapter 1;
2. Chapter 2, possibly omitting details about zero dynamics (Section 2.3.3), and going lightly on some of the proofs in Section 2.3;
3. Chapter 3, certainly going lightly on the proofs in Section 3.3;
4. Chapter 4, probably omitting Bode's Gain/Phase Theorem (Section 4.4.2) and perhaps material about plant uncertainty models (Section 4.5);
5. Chapter 5, omitting many of the details of signal and system norms in Section 5.3, omitting Liapunov stability (Section 5.4), and omitting the proofs of the Routh/Hurwitz criteria;
6. Chapter 6, going lightly, perhaps, on the detailed account of signal flow graphs in Sections 6.1 and 6.2, and covering as much of the material in Section 6.4 as deemed appropriate; the material in Section 6.5 may form the core of the discussion about feedback in a more traditional course;[1]
7. Chapter 7, probably omitting robust stability (Section 7.3);
8. Chapter 8;
9. maybe some of the material in Chapter 9, if the instructor is so inclined;
10. Chapter 11, although I rarely say much about root-locus in the course I teach;
11. Chapter 12, omitting Section 12.3 if robustness has not been covered in the earlier material;
12. perhaps some of the advanced PID synthesis methods of Chapter 13.

When I teach the introductory course, it is offered with a companion lab class. The lab course follows the lecture course in content, although it is somewhat more "down to earth." Labs start out with the objective of getting students familiar with the ideas introduced in lectures, and by the end of the course, students are putting into practice these ideas to design controllers.

A more advanced course, having as prerequisite the material from the basic course, could be structured as follows:

1. thorough treatment of material in Chapter 2;
2. ditto for Chapter 3;
3. Bode's Gain/Phase Theorem (Section 4.4.2) and uncertainty models (Section 4.5);
4. thorough treatment of signal and system norms from Section 5.3, proofs of Routh/Hurwitz criteria if one is so inclined, and Liapunov methods for stability (Section 5.4);
5. static state feedback, static output feedback, and dynamic output feedback (Section 6.4);
6. robust stability (Section 7.3);
7. design limitations in Chapter 9;
8. robust performance (Section 9.3);

[1] Of course, someone teaching a traditional course is unlikely to be using this book.

9. Chapter 10, maybe omitting Section 10.4 on strong stabilisation;
10. basic loop shaping using robustness criterion (Section 12.3);
11. perhaps the advanced synthesis methods of Chapter 13;
12. Chapter 14;
13. Chapter 15.

The content

In Chapter 1 we engage in a loose discourse on ideas of a control theoretic nature. The value of feedback is introduced via a simple DC servo motor example using proportional feedback. Modelling and linearisation are also discussed in this chapter. From here, the book breaks up into three parts (plus appendices), with the presentation taking a rather less loose form.

Part I. System representations and their properties

Linear systems are typically represented in one of three ways: in the time domain using state space methods (Chapter 2); in the Laplace transform domain using transfer functions (Chapter 3); and in the frequency domain using the frequency response (Chapter 4). These representations are all related in one way or another, and there exist vocal proponents of one or the other representation. I do not get involved in any discussion over which representation is "best," but treat each with equal importance (as near as I can), pointing out the innate similarities shared by the three models.

As is clear from the book's subtitle, the treatment is single-input, single-output (SISO), with a very few exceptions, all of them occurring near the beginning of Chapter 2. The focus on SISO systems allows students to have in mind simple models. MIMO generalisations of the results in the book typically fall into one of two categories, trivial and very difficult. The former will cause no difficulty, and the latter serve to make the treatment more difficult than is feasible in an introductory text. References are given to advanced material.

Specialised topics in this part of the book include a detailed description of zero dynamics in both the state space and the transfer function representations. This material, along with the discussion of the properties of the transfer function in Section 3.3, have a rather technical nature. However, the essential ideas can be easily grasped independent of a comprehension of the proofs. Another specialised topic is a full account of Bode's Gain/Phase Theorem in Section 4.4.2. This is an interesting theorem; however, time does not normally permit me to cover it in an introductory course.

A good understanding of the material in this first part of the book makes the remainder of the book somewhat more easily digestible. It has been my experience that students find this first material the most difficult.

Part II. System analysis

Armed with a thorough understanding of the three representations of a linear system, the student is next guided through methods for analysing such systems. The first concern in such a discussion should be, and here is, stability. A control design cannot be considered in any way successful unless it has certain stability properties. Stability for control systems has an ingredient that separates it from stability for simple dynamical systems. In control, one

is often presented with a system that is nominally unstable, and it is desired to stabilise it using feedback. Thus feedback is another central factor in our discussion of control systems analysis. We are rather more systematic about this than is the norm. The discussion of signal flow graphs in Sections Section 6.1 and 6.2 is quite detailed, and some of this detail can be skimmed. However, the special notion of stability for interconnected systems, here called IBIBO stability, is important, and the notation associated with it appears throughout the remainder of the book. The Nyquist criterion for IBIBO stability is an important part of classical control. Indeed, in Section 7.3 the ideas of the Nyquist plot motivate our discussion of robust stability. A final topic in control systems analysis is performance, and this is covered in two chapters, 8 and 9, the latter being concerned with limitations on performance that arise due to features of the plant.

The latter of the two chapters on performance contains some specialised material concerning limitations on controller design that are covered in the excellent text of Seron, Braslavsky, and Goodwin [1997]. Also in this chapter is presented the "robust performance problem," whose solution comprises Chapter 15. Thus Chapter 9 should certainly be thought of as one of special topics, not likely to be covered in detail in a first course.

Part III. Controller design

The final part of the text proper is a collection of control design schemes. We have tried to present this material in as systematic a manner as possible. This gives some emphasis to the fact that in modern linear control, there are well-developed design methods based on a solid mathematical foundation. That said, an attempt has been made to point out that there will always be an element of "artistry" to a good control design. While an out of the box controller using some of the methods we present may be a good starting point, a good control designer can always improve on such a design using their experience as a guide. This sort of material is difficult to teach, of course. However, an attempt has been made to give sufficient attention to this matter.

This part of the book starts off with a discussion of the stabilisation problem.

Part IV. Background and addenda

There are appendices reviewing relevant material in linear algebra, the matrix exponential, complex variables, and transforms. It is expected that students will have seen all of the material in these appendices, but they can look here to refamiliarise themselves with some basic concepts.

What is not in the book

The major omission of the book is discrete time ideas. These are quite important in our digital age. However, students familiar with the continuous time ideas presented here will have no difficulty understanding their discrete time analogues. That said, it should be understood that an important feature in control is missing with the omission of digital control, and that instructors may wish to insert material of this nature.

This book is in its third go around. The version this year is significantly expanded from previous years, so there are apt to be many errors. If you find an error, no matter how small, *let me know*!

Andrew D. Lewis
Department of Mathematics & Statistics

Queen's University
Kingston, ON K7L 3N6, Canada
andrew@mast.queensu.ca
(613) 533-2395
03/09/2014

Table of Contents

Chapter 1

An introduction to linear control theory

With this book we will introduce you to the basics ideas of control theory, and the setting will be that of single-input, single-output (SISO), finite-dimensional, time-invariant, linear systems. In this section we will begin to explore the meaning of this lingo, and look at some simple physical systems which fit into this category. Traditional introductory texts in control may contain some of this material in more detail [see, for example Dorf and Bishop 2010, Franklin, Powell, and Emani-Naeini 2009]. However, our presentation here is intended to be more motivational than technical. For proper background in physics, one should look to suitable references. A very good summary reference for the various methods of deriving equations for physical systems is [Cannon, Jr. 1967].

Contents

1.1 Some control theoretic terminology

For this book, there should be from the outset a picture you have in mind of what you are trying to accomplish. The picture is essentially given in Figure 1.1. The idea is that you are given a ***plant***, which is the basic system, which has an ***output*** $y(t)$ that you'd like to do something with. For example, you may wish to track a ***reference trajectory*** $r(t)$. One way to do this would be to use an ***open-loop*** control design. In this case, one would omit that part of the diagram in Figure 1.1 which is dashed, and use a ***controller*** to read the reference signal $r(t)$ and use this to specify an ***input*** $u(t)$ to the plant which should give the desired output. This open-loop control design may well work, but it has some inherent problems. If there is a ***disturbance*** $d(t)$ which you do not know about, then this may well cause the output of the plant to deviate significantly from the reference trajectory $r(t)$. Another problem arises with plant ***uncertainties***. One models the plant, typically via differential equations, but these are always an idealisation of the plant's actual behaviour. The reason for the problems is that the open-loop control law has no idea what the output is doing, and it marches on as if everything is working according to an idealised

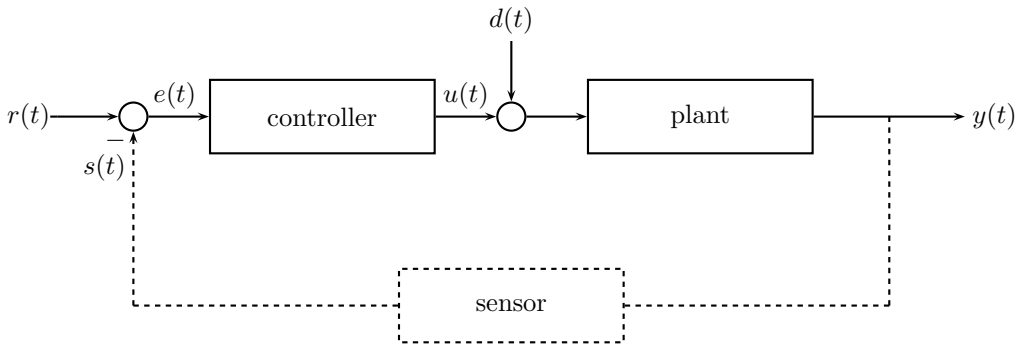


Figure 1.1 A basic control system schematic

model, a model which just might not be realistic. A good way to overcome these difficulties is to use ***feedback***. Here the output is read by ***sensors***, which may themselves be modelled by differential equations, which produce a signal $s(t)$ which is subtracted from the reference trajectory to produce the ***error*** $e(t)$. The controller then make its decisions based on the error signal, rather than just blindly considering the reference signal.

1.2 An introductory example

Let's see how this all plays out in a simple example. Suppose we have a DC servo motor whose output is its angular velocity $\omega(t)$, the input is a voltage $E(t)$, and there is a disturbance torque $T(t)$ resulting from, for example, an unknown external load being applied to the output shaft of the motor. This external torque is something we cannot alter. A little later in this section we will see some justification for the governing differential equations to be given by

$$\frac{\mathrm{d}\omega(t)}{\mathrm{d}t} + \tfrac{1}{\tau}\omega(t) = k_E E(t) + k_T T(t).$$

The schematic for the situation is shown in Figure 1.2. This schematic representation we

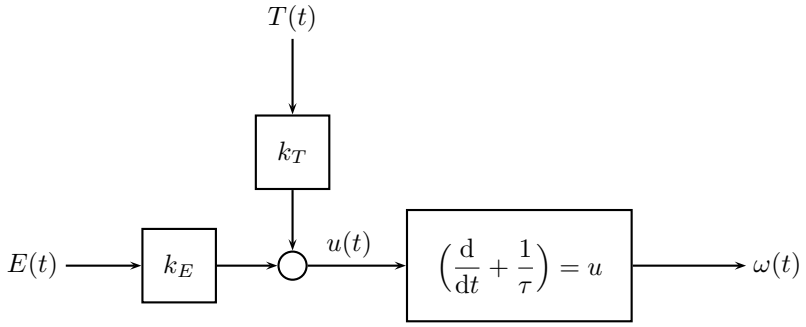


Figure 1.2 DC motor open-loop control schematic

give here is one we shall use frequently, and it is called a ***block diagram***.[1]

[1] When we come to use block diagrams for real, you will see that the thing in the blocks are not differential equations in the time-domain, but in the Laplace transform domain.

Let us just try something naïve and open-loop. The objective is to be able to drive the motor at a specified constant velocity ω_0. This constant desired output is then our reference trajectory. You decide to see what you might do by giving the motor some constant torques to see what happens. Let us provide a constant input torque $E(t) = E_0$ and suppose that the disturbance torque $T(t) = 0$. We then have the differential equation

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} + \tfrac{1}{\tau}\omega = k_E E_0.$$

Supposing, as is reasonable, that the motor starts with zero initial velocity, i.e., $\omega(0) = 0$, the solution to the initial value problem is

$$\omega(t) = k_E E_0 \tau\big(1 - e^{-t/\tau}\big).$$

We give a numerical plot for $k_E = 2$, $E_0 = 3$, and $\frac{1}{\tau} = \frac{1}{2}$ in Figure 1.3. Well, we say, this all

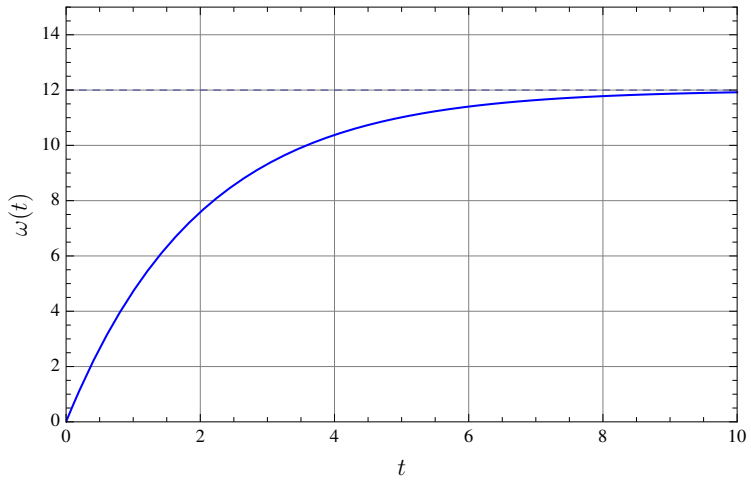


Figure 1.3 Open-loop response of DC motor

looks too easy. To get the desired output velocity ω_0 after a sufficiently long time, we need only provide the input voltage $E_0 = \frac{\omega_0}{\tau k_E}$.

However, there are decidedly problems lurking beneath the surface. For example, what if there *is* a disturbance torque? Let us suppose this to be constant for the moment so $T(t) = -T_0$ for some $T_0 > 0$. The differential equation is then

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} + \tfrac{1}{\tau}\omega = k_E E_0 - k_T T_0,$$

and if we again suppose that $\omega(0) = 0$ the initial value problem has solution

$$\omega(t) = (k_E E_0 - k_T T_0)\tau\big(1 - e^{-t/\tau}\big).$$

If we follow our simple rule of letting the input voltage E_0 be determined by the desired final angular velocity by our rule $E_0 = \frac{\omega_0}{k_E \tau}$, then we will undershoot our desired final velocity by $\omega_{\text{error}} = k_T T_0 \tau$. In this event, the larger is the disturbance torque, the worse we do—in fact, we can do pretty darn bad if the disturbance torque is large. The effect is illustrated in Figure 1.4 with $k_T = 1$ and $T_0 = 2$.

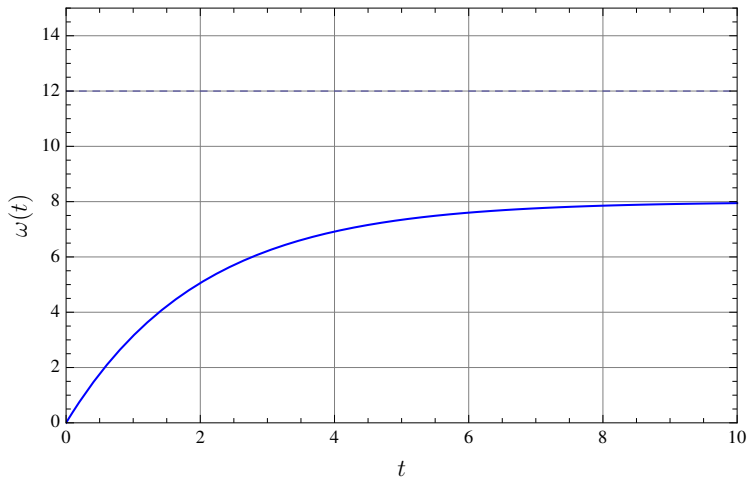


Figure 1.4 Open-loop response of DC motor with disturbance

Another problem arises when we have imperfect knowledge of the motor's physical characteristics. For example, we may not know the time-constant τ as accurately as we'd like. While we estimate it to be τ, it might be some other value $\tilde{\tau}$. In this case, the *actual* differential equation governing behaviour in the absence of disturbances will be

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} + \tfrac{1}{\tilde{\tau}}\omega = k_E E_0,$$

which gives the solution to the initial value problem as

$$\omega(t) = k_E E_0 \tilde{\tau}\big(1 - e^{-t/\tilde{\tau}}\big).$$

The final value will then be in error by the factor $\frac{\tilde{\tau}}{\tau}$. This situation is shown in Figure 1.5 for $\frac{1}{\tilde{\tau}} = \frac{5}{8}$.

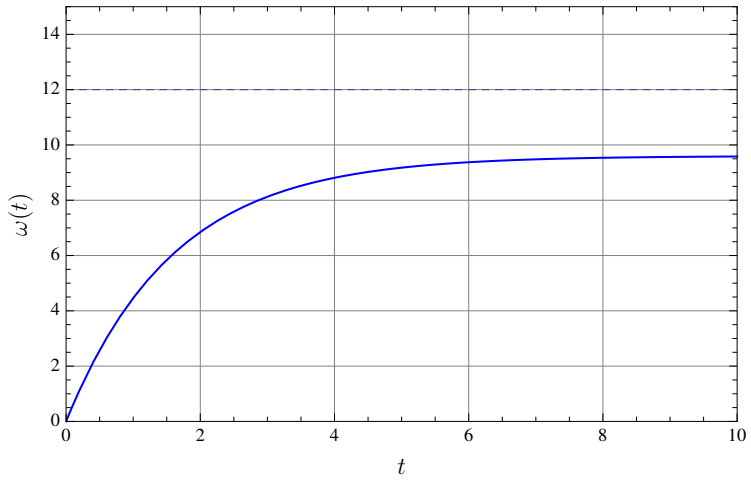


Figure 1.5 Open-loop response of DC motor with "actual" motor time-constant

Okay, I hope now that you can see the problem with our open-loop control strategy. It simply does not account for the inevitable imperfections we will have in our knowledge of

the system and of the environment in which it works. To take all this into account, let us measure the output velocity of the motor's shaft with a tachometer. The tachometer takes the angular velocity and returns a voltage. This voltage, after being appropriately scaled by a factor k_s, is then compared to the voltage needed to generate the desired velocity by feeding it back to our reference signal by subtracting it to get the error. The error we multiply by some constant K, called the ***gain*** for the controller, to get the actual voltage input to the system. The schematic now becomes that depicted in Figure 1.6. The differential equations

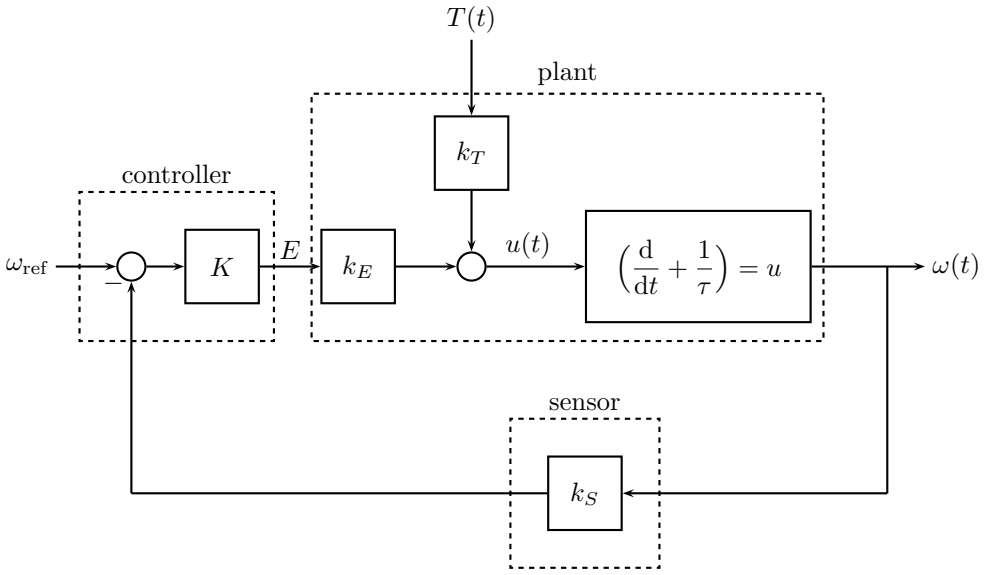


Figure 1.6 DC motor closed-loop control schematic

governing this system are

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} + (\tfrac{1}{\tau} + k_E k_S K)\omega = k_E K \omega_{\text{ref}} + k_T T.$$

We shall see how to systematically obtain equations such as this, so do not worry if you think it in nontrivial to get these equations. Note that the input to the system is, in some sense, no longer the voltage, but the reference signal ω_{ref}. Let us suppose again a constant disturbance torque $T(t) = -T_0$ and a constant reference voltage $\omega_{\text{ref}} = \omega_0$. The solution to the differential equation, again supposing $\omega(0) = 0$, is then

$$\omega(t) = \frac{k_E K \omega_0 - k_T T_0}{\frac{1}{\tau} + k_E k_S K}\big(1 - e^{-(\frac{1}{\tau} + k_E k_S K)t}\big).$$

Let us now investigate this closed-loop control law. As previously, let us first look at the case when $T_0 = 0$ and where we suppose perfect knowledge of our physical constants and our model. In this case, we wish to achieve a final velocity of $\omega_0 = E_0 \tau k_E$ as $t \to \infty$, i.e., the same velocity as we had attained with our open-loop strategy. We see the results of this in Figure 1.7 where we have chosen $k_S = 1$ and $K = 5$. Notice that the motor no longer achieves the desired final speed! However, we have improved the response time for the system significantly from the open-loop controller (cf. Figure 1.3). It is possible to remove the final error by doing something more sophisticated with the error than multiplying it by K, but we will get to that only as the course progresses. Now let's see what happens when we

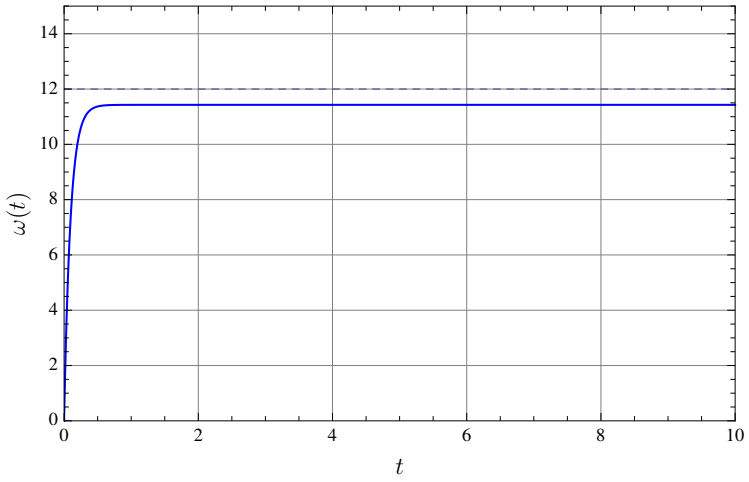


Figure 1.7 Closed-loop response of DC motor

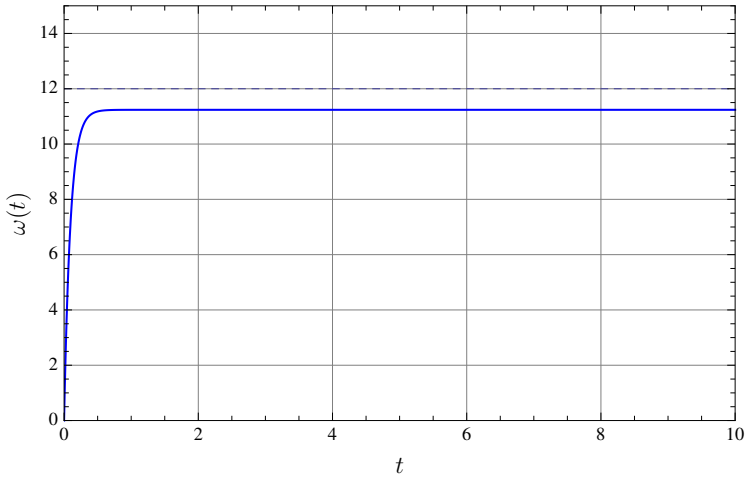


Figure 1.8 Closed-loop response of DC motor with disturbance

add a constant disturbance by setting $T_0 = 2$. The result is displayed in Figure 1.8. We see that the closed-loop controller reacts much better to the disturbance (cf. Figure 1.4), although we still (unsurprisingly) cannot reach the desired final velocity. Finally we look at the situation when we have imperfect knowledge of the physical constants for the plant. We again consider having $\frac{1}{\tilde{\tau}} = \frac{5}{8}$ rather than the guessed value of $\frac{1}{2}$. In this case the closed-loop response is shown in Figure 1.9. Again, the performance is somewhat better than that of the open-loop controller (cf. Figure 1.5), although we have incurred a largish final error in the final velocity.

I hope this helps to convince you that feedback is a good thing! As mentioned above, we shall see that it is possible to design a controller so that the steady-state error is zero, as this is the major deficiency of our very basic controller "designed" above. This simple example, however, does demonstrate that one can achieve improvements in some areas (response time in this case), although sometimes at the expense of deterioration in others (steady-state error in this case).

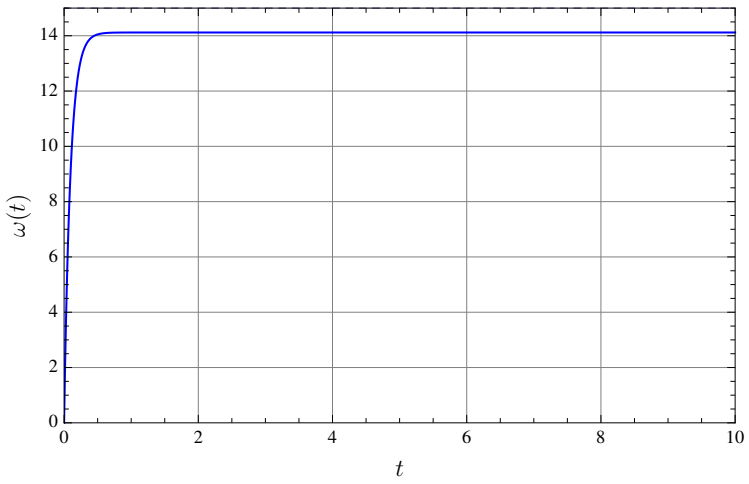


Figure 1.9 Closed-loop response of DC motor with "actual" motor time-constant

1.3 Linear differential equations for physical devices

We will be considering control systems where the plant, the controller, and the sensors are all modelled by linear differential equations. For this reason it makes sense to provide some examples of devices whose behaviour is reasonably well-governed by such equations. The problem of how to assemble such devices to, say, build a controller for a given plant is something we will not be giving terribly much consideration to.

1.3.1 Mechanical gadgets

In Figure 1.10 is a really feeble idealisation of a car suspension system. We suppose that

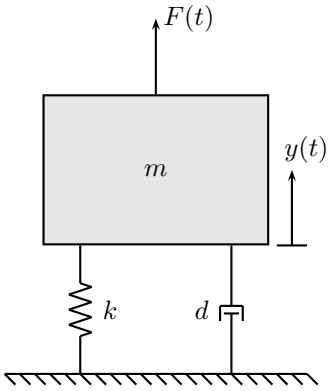


Figure 1.10 Simplified automobile suspension

at $y = 0$ the mass m is in equilibrium. The spring, as we know, then supplies a restoring force $F_k = -ky$ and the dashpot supplies a force $F_d = -d\dot{y}$, where "$\dot{}$" means $\frac{\mathrm{d}}{\mathrm{d}t}$. We also suppose there to be an externally applied force $F(t)$. If we ask Isaac Newton, "Isaac, what are the equations governing the behaviour of this system?" he would reply, "Well, $F = ma$,

now go think on it."[2] After doing so you'd arrive at

$$m\ddot{y}(t) = F(t) - ky(t) - d\dot{y}(t) \quad \Longrightarrow \quad m\ddot{y}(t) + d\dot{y}(t) + ky(t) = F(t).$$

This is a second-order linear differential equation with constant coefficients and with inhomogeneous term $F(t)$.

The same sort of thing happens with rotary devices. In Figure 1.11 is a rotor fixed to a shaft moving with angular velocity ω. Viscous dissipation may be modelled with a force

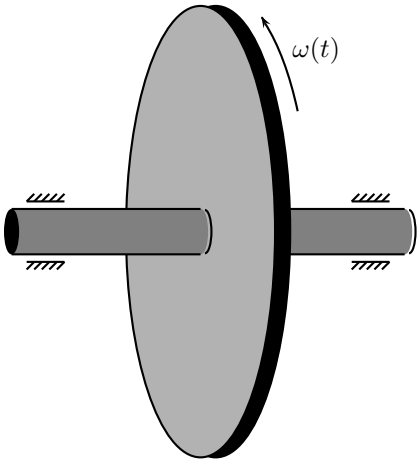


Figure 1.11 Rotor on a shaft

proportional to the angular velocity: $F_d = -d\omega$. In the presence of an external torque $\tau(t)$, the governing differential equation is

$$J\dot{\omega}(t) + d\omega(t) = \tau(t)$$

where J is the moment of inertia of the rotor about its point of rotation. If one wishes to include a rotary spring, then one must consider not the angular velocity $\omega(t)$ as the dependent variable, but rather the angular displacement $\theta(t) = \theta_0 + \omega t$. In either case, the governing equations are linear differential equations.

Let's look at a simple pendulum (see Figure 1.12). If we sum moments about the pivot we get

$$m\ell^2\ddot{\theta} = -mg\ell\sin\theta \quad \Longrightarrow \quad \ddot{\theta} + \tfrac{g}{\ell}\sin\theta = 0.$$

Now this equation, you will notice, is nonlinear. However, we are often interested in the behaviour of the system near the equilibrium points which are $(\theta, \dot{\theta}) = (\theta_0, 0)$ where $\theta_0 \in \{0, \pi\}$. So, let us linearise the equations near these points, and see what we get. We write the solution near the equilibrium as $\theta(t) = \theta_0 + \xi(t)$ where $\xi(t)$ is small. We then have

$$\ddot{\theta} + \tfrac{g}{\ell}\sin\theta = \ddot{\xi} + \tfrac{g}{\ell}\sin(\theta_0 + \xi) = \ddot{\xi} + \tfrac{g}{\ell}\sin\theta_0\cos\xi + \tfrac{g}{\ell}\cos\theta_0\sin\xi.$$

Now note that $\sin\theta_0 = 0$ if $\theta_0 \in \{0, \pi\}$, and $\cos\theta_0 = 1$ if $\theta_0 = 0$ and $\cos\theta_0 = -1$ if $\theta_0 = \pi$. We also use the Taylor expansion for $\sin x$ around $x = 0$: $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \dots$. Keeping

[2] This is in reference to the story, be it true or false, that when Newton was asked how he'd arrived at the inverse square law for gravity, he replied, "I thought on it."

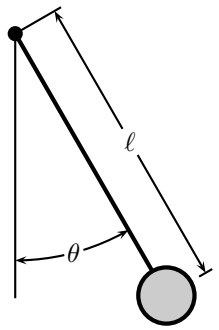


Figure 1.12 A simple pendulum

only the lowest order terms gives the following equations which should approximate the behaviour of the system near the equilibrium $(\theta_0, 0)$:

$$\ddot{\xi} + \tfrac{g}{\ell}\xi = 0, \quad \theta_0 = 0$$
$$\ddot{\xi} - \tfrac{g}{\ell}\xi = 0, \quad \theta_0 = \pi.$$

In each case, we have a linear differential equation which governs the behaviour near the equilibrium. This technique of linearisation is ubiquitous since there really are no linear physical devices, but linear approximations seem to work well, and often very well, particularly in control. We discuss linearisation properly in Section 1.4.

Let us recall the basic rules for deriving the equations of motion for a mechanical system.

1.1 Deriving equations for mechanical systems Given: an interconnection of point masses and rigid bodies.

1. Define a reference frame from which to measure distances.
2. Choose a set of coordinates that determine the configuration of the system.
3. Separate the system into its mechanical components. Thus each component should be either a single point mass or a single rigid body.
4. For each component determine all external forces and moments acting on it.
5. For each component, express the position of the centre of mass in terms of the chosen coordinates.
6. The sum of forces in any direction on a component should equal the mass of the component times the component of acceleration of the component along the direction of the force.
7. For each component, the sum of moments about a point that is either (a) the centre of mass of the component or (b) a point in the component that is stationary should equal the moment of inertia of the component about that point multiplied by the angular acceleration. •

This methodology has been applied to the examples above, although they are too simple to be really representative. We refer to the exercises for examples that are somewhat more interesting. Also, see [Cannon, Jr. 1967] for details on the method we outline, and other methods.

1.3.2 Electrical gadgets

A ***resistor*** is a device across which the voltage drop is proportional to the current through the device. A ***capacitor*** is a device across which the voltage drop is proportional to the charge in the device. An ***inductor*** is a device across which the voltage drop is proportional to the time rate of change of current through the device. The three devices are typically given the symbols as in Figure 1.13. The quantity R is called the ***resistance*** of

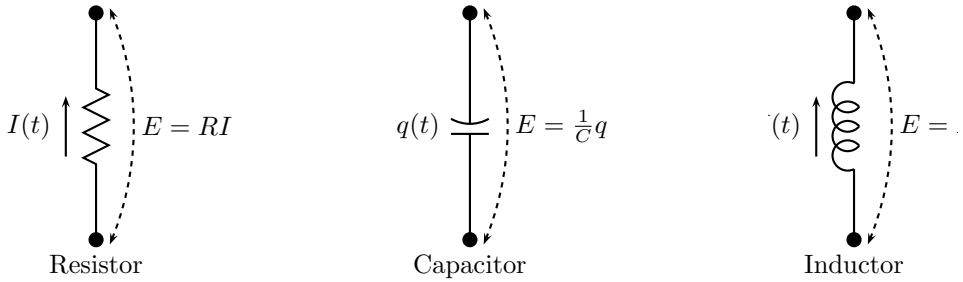


Figure 1.13 Electrical devices

the resistor, the quantity C is called the ***capacitance*** of the capacitor, and the quantity L is called the ***inductance*** of the inductor. Note that the proportionality constant for the capacitor is not C but $\frac{1}{C}$. The current I is related to the charge q by $I = \frac{\mathrm{d}q}{\mathrm{d}t}$. We can then imagine assembling these electrical components in some configuration and using Kirchhoff's laws[3] to derive governing differential equations. In Figure 1.14 we have a particularly simple

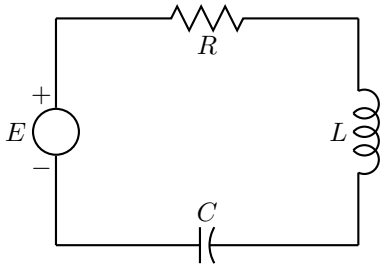


Figure 1.14 A series RLC circuit

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:

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

In either case, we have a linear equation, and again one with constant coefficients.

Let us present a methodology for determining the differential equations for electric circuits. The methodology relies on the notion of a ***tree*** which is a connected collection of

[3] ***Kirchhoff's voltage law*** states that the sum of voltage drops around a closed loop must be zero and ***Kirchhoff's current law*** states that the sum of the currents entering a node must be zero.

branches containing no loops. For a given tree, a ***tree branch*** is a branch in the tree, and a ***link*** is a branch not in the tree.

1.2 Deriving equations for electric circuits Given: an interconnection of ideal resistors, capacitors, and inductors, along with voltage and current sources.

1. Define a tree by collecting together a maximal number of branches to form a tree. Add elements in the following order of preference: voltage sources, capacitors, resistors, inductors, and current sources. That is to say, one adds these elements in sequence until one gets the largest possible tree.
2. The states of the system are taken to be the voltages across capacitors in the tree branches for the tree of part 1 and the currents through inductors in the links for the tree from part 1.
3. Use Kirchhoff's Laws to derive equations for the voltage and current in every tree branch in terms of the state variables.
4. Write the Kirchhoff Voltage Law and the Kirchhoff Current Law for every loop and every node corresponding to a branch assigned a state variable. •

The exercises contain a few examples that can be used to test one's understanding of the above method. We also refer to [Cannon, Jr. 1967] for further discussion of the equations governing electrical networks.

1.3.3 Electro-mechanical gadgets

If you really want to learn how electric motors work, then read a book on the subject. For example, see [Cannon, Jr. 1967].

A DC servo motor works by running current through a rotary toroidal coil which sits in a stationary magnetic field. As current is run through the coil, the induced magnetic field induces the rotor to turn. The torque developed is proportional to the current through the coil: $T = K_t I$ where T is the torque supplied to the shaft, I is the current through the coil, and K_t is the "torque constant." The voltage drop across the motor is proportional to the motor's velocity; $E_m = K_e \dot{\theta}$ where E_m is the voltage drop across the motor, K_e is a constant, and θ is the angular position of the shaft. If one is using a set of consistent units with velocity measured in rads/sec, then apparently $K_e = K_t$.

Now we suppose that the rotor has inertia J and that shaft friction is viscous and so the friction force is given by $-d\dot{\theta}$. Thus the motor will be governed by Newton's equations:

$$J\ddot{\theta} = -d\dot{\theta} + K_t I \quad \Longrightarrow \quad J\ddot{\theta} + d\dot{\theta} = K_t I.$$

To complete the equations, one need to know the relationship between current and θ. This is provided by the relation $E_m = K_e \dot{\theta}$ and the dynamics of the circuit which supplies current to the motor. For example, if the circuit has resistance R and inductance L then we have

$$L\frac{\mathrm{d}I}{\mathrm{d}t} + RI = E - K_e\dot{\theta}$$

with E being the voltage supplied to the circuit. This gives us coupled equations

$$J\ddot{\theta} + d\dot{\theta} = K_t I$$
$$L\frac{\mathrm{d}I}{\mathrm{d}t} + RI = E - K_e\dot{\theta}$$

which we can write in first-order system form as

$$\begin{bmatrix} \dot\theta \\ \dot v_\theta \\ \dot I \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -\frac{d}{J} & \frac{K_t}{J} \\ 0 & -\frac{K_e}{L} & -\frac{R}{L} \end{bmatrix} \begin{bmatrix} \theta \\ v_\theta \\ I \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{L} \end{bmatrix} E$$

where we define the dependent variable $v_\theta = \dot\theta$. If the response of the circuit is much faster than that of the motor, e.g., if the inductance is small, then this gives $E = K_e\dot\theta + RI$ and so the equations reduce to

$$J\ddot\theta + \big(d + \tfrac{K_tK_e}{R}\big)\dot\theta = \tfrac{K_t}{R}E.$$

Thus the dynamics of a DC motor can be roughly described by a first-order linear differential equation in the angular velocity. This is what we saw in our introductory example.

Hopefully this gives you a feeling that there are a large number of physical systems which are modelled by linear differential equations, and it is these to which we will be restricting our attention.

1.4 Linearisation at equilibrium points

When we derived the equations of motion for the pendulum, the equations we obtained were nonlinear. We then decided that if we were only interested in looking at what is going on near an equilibrium point, then we could content ourselves with linearising the equations. We then did this in a sort of hacky way. Let's see how to do this methodically.

We suppose that we have vector differential equations of the form

$$\begin{aligned} \dot x_1 &= f_1(x_1, \dots, x_n) \\ \dot x_2 &= f_2(x_1, \dots, x_n) \\ &\vdots \\ \dot x_n &= f_n(x_1, \dots, x_n). \end{aligned}$$

The n functions $(f_1, \dots, f_n)$ of the n variables $(x_1, \dots, x_n)$ are known smooth functions. Let us denote $\boldsymbol{x} = (x_1, \dots, x_n)$ and $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \dots, f_n(\boldsymbol{x}))$. The differential equation can then be written as

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}). \tag{1.1}$$

It really only makes sense to linearise about an equilibrium point. An ***equilibrium point*** is a point $\boldsymbol{x}_0 \in \mathbb{R}^n$ for which $\boldsymbol{f}(\boldsymbol{x}_0) = \boldsymbol{0}$. Note that the constant function $\boldsymbol{x}(t) = \boldsymbol{x}_0$ is a solution to the differential equation if $\boldsymbol{x}_0$ is an equilibrium point. For an equilibrium point $\boldsymbol{x}_0$ define an $n \times n$ matrix $\boldsymbol{D}\boldsymbol{f}(\boldsymbol{x}_0)$ by

$$\boldsymbol{D}\boldsymbol{f}(\boldsymbol{x}_0) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}_0) & \frac{\partial f_1}{\partial x_2}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}_0) \\ \frac{\partial f_2}{\partial x_1}(\boldsymbol{x}_0) & \frac{\partial f_2}{\partial x_2}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_2}{\partial x_n}(\boldsymbol{x}_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\boldsymbol{x}_0) & \frac{\partial f_n}{\partial x_2}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_n}{\partial x_n}(\boldsymbol{x}_0) \end{bmatrix}.$$

This matrix is often called the ***Jacobian*** of $\boldsymbol{f}$ at $\boldsymbol{x}_0$. The ***linearisation*** of (1.1) about an equilibrium point $\boldsymbol{x}_0$ is then the linear differential equation

$$\dot{\boldsymbol{\xi}} = \boldsymbol{D}\boldsymbol{f}(\boldsymbol{x}_0)\boldsymbol{\xi}.$$

Let's see how this goes with our pendulum example.

1.3 Example The nonlinear differential equation we derived was

$$\ddot{\theta} + \tfrac{g}{\ell} \sin\theta = 0.$$

This is not in the form of (1.1) since it is a second-order equation. But we can put this into first-order form by introducing the variables $x_1 = \theta$ and $x_2 = \dot{\theta}$. The equations can then be written

$$\begin{aligned} \dot{x}_1 &= \dot{\theta} = x_2 \\ \dot{x}_2 &= \ddot{\theta} = -\tfrac{g}{\ell} \sin\theta = -\tfrac{g}{\ell} \sin x_1. \end{aligned}$$

Thus

$$f_1(x_1, x_2) = x_2, \quad f_2(x_1, x_2) = -\frac{g}{\ell} \sin x_1.$$

Note that at an equilibrium point we must have $x_2 = 0$. This makes sense as it means that the pendulum should not be moving. We must also have $\sin x_1 = 0$ which means that $x_1 \in \{0, \pi\}$. This is what we determined previously.

Now let us linearise about each of these equilibrium points. For an arbitrary point $\boldsymbol{x} = (x_1, x_2)$ we compute

$$\boldsymbol{D}\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} 0 & 1 \\ -\frac{g}{\ell} \cos x_1 & 0 \end{bmatrix}.$$

At the equilibrium point $\boldsymbol{x}_1 = (0, 0)$ we thus have

$$\boldsymbol{D}\boldsymbol{f}(\boldsymbol{x}_1) = \begin{bmatrix} 0 & 1 \\ -\frac{g}{\ell} & 0 \end{bmatrix},$$

and at the equilibrium point $\boldsymbol{x}_2 = (0, \pi)$ we thus have

$$\boldsymbol{D}\boldsymbol{f}(\boldsymbol{x}_1) = \begin{bmatrix} 0 & 1 \\ \frac{g}{\ell} & 0 \end{bmatrix}.$$

With these matrices at hand, we may write the linearised equations at each equilibrium point. •

1.5 What you are expected to know

There are five essential areas of background that are assumed of a student using this text. These are

1. linear algebra,
2. ordinary differential equations, including the matrix exponential,
3. basic facts about polynomials,
4. basic complex analysis, and
5. transform theory, especially Fourier and Laplace transforms.

Appendices review each of these in a cursory manner. Students are expected to have seen this material in detail in previous courses, so there should be no need for anything but rapid review in class.

Many of the systems we will look at in the exercises require in their analysis straightforward, but tedious, calculations. It should not be the point of the book to make you go through such tedious calculations. You will be well served by learning to use a computer package for doing such routine calculations, although you should try to make sure you are asking the computer to do something which you in principle understand how to do yourself. I have used Mathematica® to do all the plotting in the book since it is what I am familiar with. Also available are Maple® and Matlab®. Matlab® has a control toolbox, and is the most commonly used tool for control systems.[4]

You are encouraged to use symbolic manipulation packages for doing problems in this book. Just make sure you let us know that you are doing so, and make sure you know what you are doing and that you are not going too far into black box mode.

1.6 Summary

Our objective in this chapter has been to introduce you to some basic control theoretic ideas, especially through the use of feedback in the DC motor example. In the remainder of these notes we look at linear systems, and to motivate such an investigation we presented some physical devices whose behaviour is representable by linear differential equations, perhaps after linearisation about a desired operating point. We wrapped up the chapter with a quick summary of the background required to proceed with reading these notes. Make sure you are familiar with everything discussed here.

[4]Mathematica® and Maple® packages have been made available on the world wide web for doing things such as are done in this book. See `http://mast.queensu.ca/~math332/`.

Exercises

E1.1 Probe your life for occurrences of things which can be described, perhaps roughly, by a schematic like that of Figure 1.1. Identify the components in your system which are the plant, output, input, sensor, controller, etc. Does your system have feedback? Are there disturbances?

E1.2 Consider the DC servo motor example which we worked with in Section 1.2. Determine conditions on the controller gain K so that the voltage E_0 required to obtain a desired steady-state velocity is greater for the closed-loop system than it is for the open-loop system. You may neglect the disturbance torque, and assume that the motor model is accurate. However, do not use the numerical values used in the notes—leave everything general.

E1.3 An amplifier is to be designed with an overall amplification factor of 2500 ± 50. A number of single amplifier stages is available and the gain of any single stage may drift anywhere between 25 and 75. The configuration of the final amplifier is given in Figure E1.1. In each of the blocks we get to insert an amplifier stage with the large

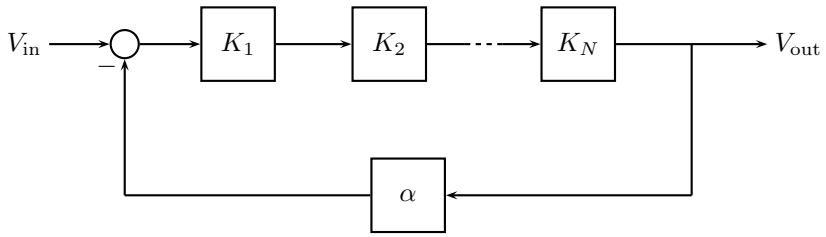


Figure E1.1 A multistage feedback amplifier

and unknown gain variation (in this case the gain variation is at most 50). Thus the gain in the forward path is $K_1K_2\cdots K_N$ where N is the number of amplifier stages and where $K_i \in [25, 75]$. The element in the feedback path is a constant $0 < \alpha < 1$.

(a) For N amplifier stages and a given value for α determine the relationship between V_{in} and V_{out}.

The feedback gain α is known precisely since it is much easier to design a circuit which provides accurate voltage division (as opposed to amplification). Thus, we can assume that α can be exactly specified by the designer.

(b) Based on this information find a value of α in the interval $(0, 1)$ and, for that value of α, the *minimal* required number of amplifier stages, N_{min}, so that the final amplifier design meets the specification noted above.

E1.4 Derive the differential equations governing the behaviour of the coupled masses in Figure E1.2. How do the equations change if viscous dissipation is added between each mass and the ground? (Suppose that both masses are subject to the same dissipative force.)

The following two exercises will recur as exercises in succeeding chapters. Since the computations can be a bit involved—certainly they ought to be done with a symbolic manipulation package—it is advisable to do the computations in an organised manner so that they may be used for future calculations.

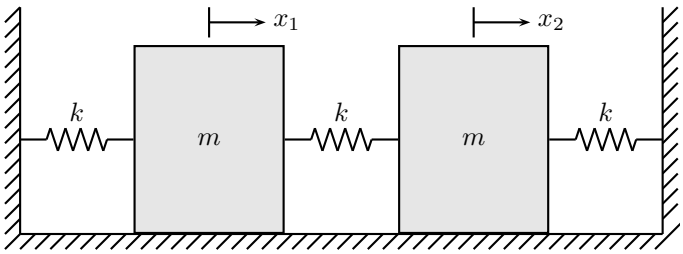


Figure E1.2 Coupled masses

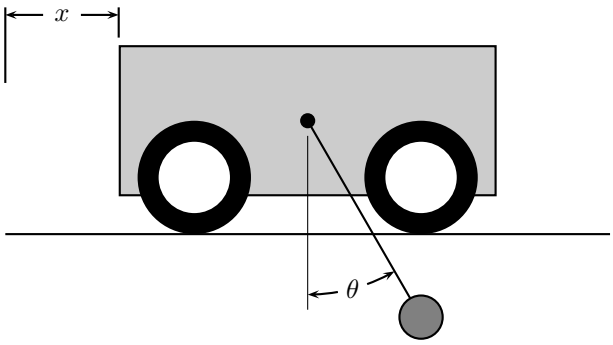


Figure E1.3 Pendulum on a cart

E1.5 Consider the pendulum on a cart pictured in Figure E1.3. Derive the full equations which govern the motion of the system using coordinates (x, θ) as in the figure. Here M is the mass of the cart and m is the mass of the pendulum. The length of the pendulum arm is ℓ. You may assume that there is no friction in the system. Linearise the equations about the points $(x, \theta, \dot{x}, \dot{\theta}) = (x_0, 0, 0, 0)$ and $(x, \theta, \dot{x}, \dot{\theta}) = (x_0, \pi, 0, 0)$, where x_0 is arbitrary.

E1.6 Determine the full equations of motion for the double pendulum depicted in Figure E1.4. The first link (i.e., the one connected to ground) has length ℓ_1 and mass

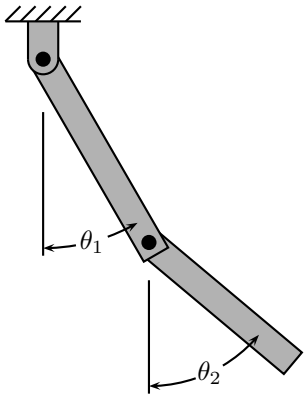


Figure E1.4 Double pendulum

m_1, and the second link has length ℓ_2 and mass m_2. The links have a uniform mass density, so their centres of mass are located at their midpoint. You may assume that

there is no friction in the system. What are the equilibrium points for the double pendulum (there are four)? Linearise the equations about each of the equilibria.

E1.7 Consider the electric circuit of Figure E1.5. To write equations for this system we

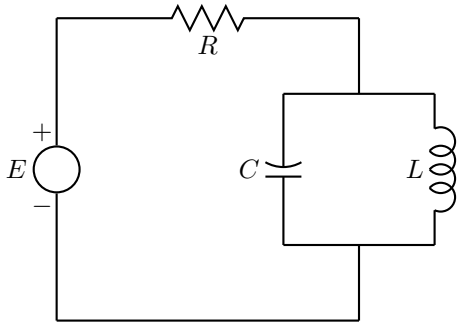


Figure E1.5 Electric circuit

need to select *two* system variables. Using I_C, the current through the capacitor, and I_L, the current through the inductor, derive a first-order system of equations in two variables governing the behaviour of the system.

E1.8 For the circuit of Figure E1.6, determine a differential equation for the current through

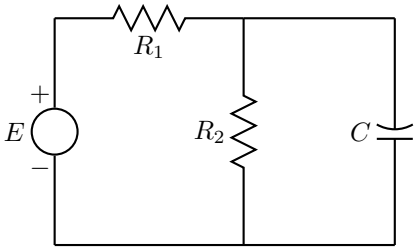


Figure E1.6 Another electric circuit

the resistor R_1 in terms of the voltage $E(t)$.

E1.9 For the circuit of Figure E1.7, As dependent variable for the circuit, use the volt-

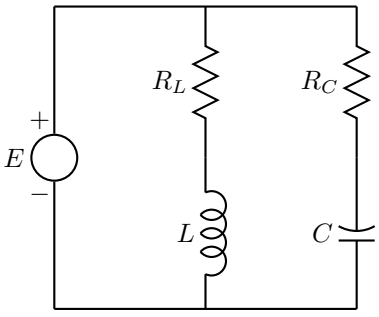


Figure E1.7 Yet another electric circuit

age across the capacitor and the current through the inductor. Derive differential equations for the system as a first-order system with two variables.

E1.10 The mass flow rate from a tank of water with a uniform cross-section can be roughly modelled as being proportional to the height of water in the tank which lies above the exit nozzle. Suppose that two tanks are configured as in Figure E1.8 (the tanks

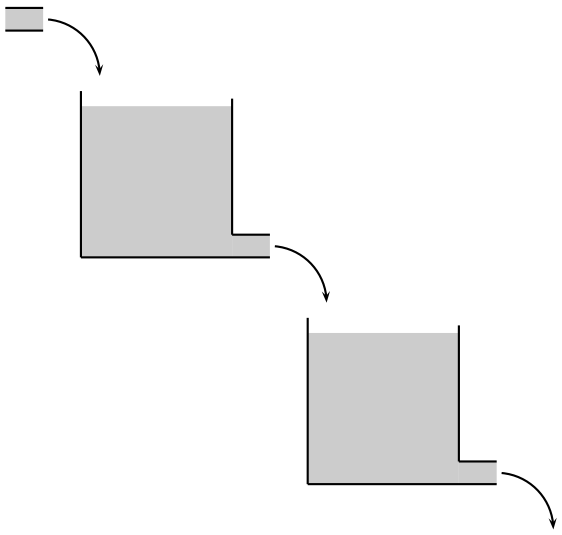


Figure E1.8 Coupled water tanks

are not necessarily identical). Determine the equations of motion which give the mass flow rate from the bottom tank given the mass flow rate into the top tank. In this problem, you must define the necessary variables yourself.

In the next exercise we will consider a more complex and realistic model of flow in coupled tanks. Here we will use the Bernoulli equation for flow from small orifices. This says that if a tank of uniform cross-section has fluid level h, then the velocity flowing from a small nozzle at the bottom of the tank will be given by $v = \sqrt{2gh}$, where g is the gravitational acceleration.

E1.11 Consider the coupled tanks shown in Figure E1.9. In this scenario, the input is the volume flow rate F_{in} which gets divided between the two tanks proportionally to the areas α_1 and α_2 of the two tubes. Let us denote $\alpha = \frac{\alpha_1}{\alpha_1+\alpha_2}$.

(a) Give an expression for the volume flow rates $F_{\text{in},1}$ and $F_{\text{in},2}$ in terms of F_{in} and α.

Now suppose that the areas of the output nozzles for the tanks are a_1 and a_2, and that the cross-sectional areas of the tanks are A_1 and A_2. Denote the water levels in the tanks by h_1 and h_2.

(b) Using the Bernoulli equation above, give an expression for the volume flow rates $F_{\text{out},1}$ and $F_{\text{out},2}$ in terms of a_1, a_2, h_1, and h_2.

(c) Using mass balance (assume that the fluid is incompressible so that mass and volume balance are equivalent), provide two coupled differential equations for the heights h_1 and h_2 in the tanks. The equations should be in terms of F_{in}, α, A_1, A_2, a_1, a_2, and g, as well as the dependent variables h_1 and h_2.

Suppose that the system is in equilibrium (i.e., the heights in the tanks are constant) with the equilibrium height in tank 1 being δ_1.

(d) What is the equilibrium input flow rate ν?

(e) What is the height δ_2 of fluid in tank 2?

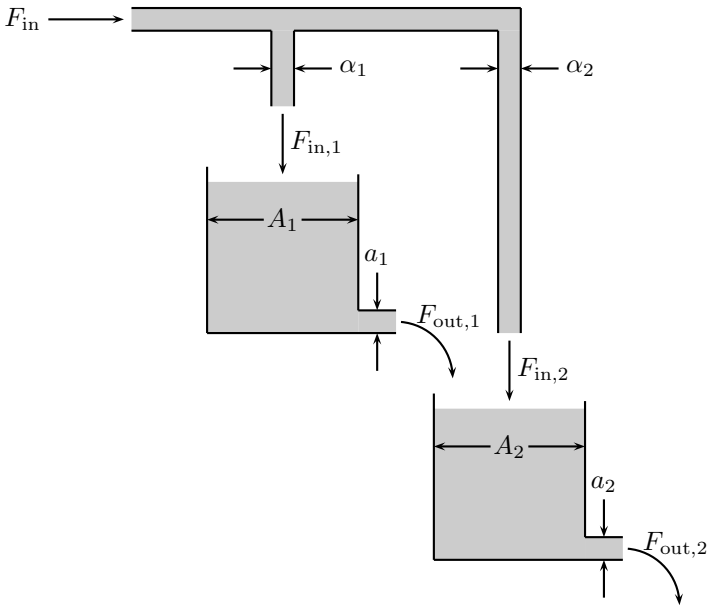


Figure E1.9 Another coupled tank scenario

Part I

System representations and their properties

Chapter 2

State-space representations (the time-domain)

With that little preamble behind us, let us introduce some mathematics into the subject. We will approach the mathematical formulation of our class of control systems from three points of view; (1) time-domain, (2) s-domain or Laplace transform domain, and (3) frequency domain. We will also talk about two kinds of systems; (1) those given to us in "linear system form," and (2) those given to us in "input/output form." Each of these possible formulations has its advantages and disadvantages, and can be best utilised for certain types of analysis or design. In this chapter we concentrate on the time-domain, and we only deal with systems in "linear system form." We will introduce the "input/output form" in Chapter 3.

Some of the material in this chapter, particularly the content of some of the proofs, is pitched at a level that may be a bit high for an introductory control course. However, most students should be able to grasp the content of all results, and understand their implications. A good grasp of basic linear algebra is essential, and we provide some of the necessary material in Appendix A. The material in this chapter is covered in many texts, including [Brockett 1970, Chen 1984, Kailath 1980, Zadeh and Desoer 1979]. The majority of texts deal with this material in multi-input, multi-output (MIMO) form. Our presentation is single-input, single-output (SISO), mainly because this will be the emphasis in the analysis and design portions of the book. Furthermore, MIMO generalisations to the majority of what we say in this chapter are generally trivial. The exception is the canonical forms for controllable and observable systems presented in Sections 2.5.1 and 2.5.2.

Contents

2.1 Properties of finite-dimensional, time-invariant linear control systems

With that little bit of linear algebra behind us, we can have at our time-domain formulation. It is in this setting that many models are handed to us in practice, so it is in my opinion the most basic way to discuss control systems. Here I differ in opinion with most introductory control texts that place our discussion here late in the course, or do not have it at all.

We begin by saying what we look at. Our definition here includes the multi-input, multi-output framework since it is easy to do so. However, we will quickly be specialising to the single-input, single-output situation.

2.1 Definition A ***finite-dimensional, time-invariant linear control system*** is given by a quadruple $\Sigma = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, $\boldsymbol{C} \in \mathbb{R}^{r\times n}$, and $\boldsymbol{D} \in \mathbb{R}^{r\times m}$. The system is ***single-input, single-output*** (***SISO***) if $m = r = 1$ and is ***multi-input, multi-output*** (***MIMO***) otherwise. •

Er... how the heck is this a control system? Like this:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t). \end{aligned} \tag{2.1}$$

Here $\boldsymbol{x} \in \mathbb{R}^n$ is the ***state*** of the system, $\boldsymbol{u} \in \mathbb{R}^m$ is the ***input***, and $\boldsymbol{y} \in \mathbb{R}^r$ is the ***output***. We call the system finite-dimensional because $n < \infty$ and time-invariant because the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ are constant. In the single-input, single-output case note that we may write the equations (2.1) as

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t) \end{aligned} \tag{2.2}$$

for vectors $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$. Here the matrix $\boldsymbol{D}$ is 1×1 and so is essentially a scalar, and $\boldsymbol{c}^t$ denotes the transpose of $\boldsymbol{c}$. We will be coming back again and again to the equations (2.2). They form a large part of what interests us in this book. Note that we will always reserve the symbol n to denote the state dimension of a SISO linear system. Therefore, from now on, if you see a seemingly undefined "n" floating around, it should be the state dimension of whatever system is being discussed at that moment.

2.2 Example Let's look at a system that can be written in the above form. We consider the mass-spring-damper system depicted in Figure 1.10. The differential equation governing the system behaviour is

$$m\ddot{x} + d\dot{x} + kx = u$$

where we denote by $u(t)$ the input force. To convert this into a set of equations of the form (2.1) we define $x_1 = x$ and $x_2 = \dot{x}$. The governing equations are then

$$\begin{aligned} \dot{x}_1 &= \dot{x} = x_2 \\ \dot{x}_2 &= \ddot{x} = -\tfrac{k}{m}x - \tfrac{d}{m}\dot{x} + \tfrac{1}{m}u = -\tfrac{k}{m}x_1 - \tfrac{d}{m}x_2 + \tfrac{1}{m}u. \end{aligned}$$

We can write this in matrix/vector form as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{d}{m} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} u.$$

So if we define

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{d}{m} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix},$$

we have the first of equations (2.1).

We shall look at three ways in which the output equation may appear in this example.

1. Suppose that with a proximeter we measure the displacement of the mass. Thus we have the output $y = x = x_1$ that we can write in matrix form as

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

so that

$$\boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 0 \end{bmatrix}.$$

2. The next scenario supposes that we have a means of measuring the velocity of the mass. Thus we take $y = \dot{x} = x_2$. In this case we have

$$y = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

so that

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 0 \end{bmatrix}.$$

3. The final situation will arise when we mount an accelerometer atop the mass so we have $y = \ddot{x} = -\frac{k}{m}x - \frac{d}{m}\dot{x} + \frac{1}{m}u$. In this case we have

$$y = \begin{bmatrix} -\frac{k}{m} & -\frac{d}{m} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \end{bmatrix} u$$

so that

$$\boldsymbol{c} = \begin{bmatrix} -\frac{k}{m} \\ -\frac{d}{m} \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} \frac{1}{m} \end{bmatrix}. \qquad \bullet$$

In order to be really clear on what we are doing, and in particular to state what we mean by linearity, we should really specify the class of inputs we consider. Let us do this.

2.3 Definition An ***interval*** is a subset I of $\mathbb{R}$ of the form

(i) $I = (-\infty, a)$,
(ii) $I = (-\infty, a]$,
(iii) $I = (a, b)$,
(iv) $I = (a, b]$,
(v) $I = (a, \infty)$,
(vi) $I = [a, b)$,
(vii) $I = [a, b]$,
(viii) $I = [a, \infty)$, or
(ix) $I = \mathbb{R}$.

Let $\mathscr{I}$ denote the set of intervals. If $I \in \mathscr{I}$, a map $f\colon I \to \mathbb{R}^k$ is ***piecewise continuous*** if it is continuous except at a discrete set of points in I,[1] and at points of discontinuity, the

[1]You will recall the notion of a discrete set (or more precisely, it will be recalled for you). For $I \in \mathscr{I}$, a (possibly infinite) collection of distinct points $P \subset I$ is called ***discrete*** if there exists $\epsilon > 0$ so that $|x-y| \geq \epsilon$ for every $x, y \in I$. If I is a bounded set, one verifies that this implies that every discrete set is finite. If I is not bounded, one wants to ensure that the points cannot get closer and closer together, and in so doing one ensures that length of the intervals on which the function is continuous always have a lower bound.

left and right limits of the function exist. An ***admissible input*** for (2.1) is a piecewise continuous map $\boldsymbol{u}\colon I \to \mathbb{R}^m$ where $I \in \mathscr{I}$, and we denote the set of admissible controls by $\mathscr{U}$. •

2.4 Remark All inputs we will encounter in this book will be in fact piecewise infinitely differentiable. However, we will also not be giving the issue too much serious consideration—be advised, however, that when dealing with control theory at any level of seriousness, the specification of the class of inputs is important. Indeed, one might generally ask that the inputs be, in the language of Lebesgue integration, essentially bounded and measurable. •

Often when dealing with time-invariant systems one makes the assumption of ***causality*** which means that inputs are zero for $t < 0$. In this book we will often tacitly make the causality assumption. However, there are brief periods when we will require the opposite of causality. Thus a system is ***anticausal*** when inputs are zero for $t > 0$.

The following result justifies our calling the system (2.1) linear.

2.5 Proposition *Let $I \in \mathscr{I}$ and let $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathscr{U}$ be defined on I with $\boldsymbol{x}_1(t)$ and $\boldsymbol{x}_2(t)$ defined as satisfying*

$$\dot{\boldsymbol{x}}_1 = \boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{B}\boldsymbol{u}_1, \quad \dot{\boldsymbol{x}}_2 = \boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{B}\boldsymbol{u}_2,$$

and $\boldsymbol{y}_1(t)$ and $\boldsymbol{y}_2(t)$ defined by

$$\boldsymbol{y}_1(t) = \boldsymbol{C}\boldsymbol{x}_1(t) + \boldsymbol{D}\boldsymbol{u}_1(t), \quad \boldsymbol{y}_2(t) = \boldsymbol{C}\boldsymbol{x}_1(t) + \boldsymbol{D}\boldsymbol{u}_2(t).$$

For $a_1, a_2 \in \mathbb{R}$, define $\boldsymbol{u}(t) = a_1\boldsymbol{u}_1(t) + a_2\boldsymbol{u}_2(t)$. Then $\boldsymbol{x}(t) \triangleq a_1\boldsymbol{x}_1(t) + a_2\boldsymbol{x}_2(t)$ satisfies

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$$

and $\boldsymbol{y}(t) \triangleq a_1\boldsymbol{y}_1(t) + a_2\boldsymbol{y}_2(t)$ satisfies

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t).$$

Proof We compute

$$\begin{aligned}
\dot{\boldsymbol{x}} &= \frac{\mathrm{d}}{\mathrm{d}t}(a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2) \\
&= a_1\dot{\boldsymbol{x}}_1 + a_2\dot{\boldsymbol{x}}_2 \\
&= a_1(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{B}\boldsymbol{u}_1) + a_2(\boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{B}\boldsymbol{u}_2) \\
&= \boldsymbol{A}(a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2) + \boldsymbol{B}(a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2) \\
&= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}
\end{aligned}$$

as claimed. We also compute

$$\begin{aligned}
\boldsymbol{y} &= a_1\boldsymbol{y}_1 + a_2\boldsymbol{y}_2 \\
&= a_1(\boldsymbol{C}\boldsymbol{x}_1 + \boldsymbol{D}\boldsymbol{u}_1) + a_2(\boldsymbol{C}\boldsymbol{x}_2 + \boldsymbol{D}\boldsymbol{u}_2) \\
&= \boldsymbol{C}(a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2) + \boldsymbol{D}(a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2) \\
&= \boldsymbol{C}\boldsymbol{x} + \boldsymbol{D}\boldsymbol{u},
\end{aligned}$$

again, as claimed. ■

The idea is that if we take as our new input a linear combination of old inputs, the same linear combination of the old states satisfies the control equations, and also the same linear combination of the old outputs satisfies the control equations.

In Proposition 2.5 we tacitly assumed that the solutions $\boldsymbol{x}_1(t)$ and $\boldsymbol{x}_2(t)$ existed for the given inputs $\boldsymbol{u}_1(t)$ and $\boldsymbol{u}_2(t)$. Solutions do in fact exist, and we may represent them in a convenient form.

2.6 Theorem *For $\boldsymbol{u} \in \mathscr{U}$ defined on $I \in \mathscr{I}$, $t_0 \in I$, and $\boldsymbol{x}_0 \in \mathbb{R}^n$, there exists a unique piecewise differentiable curve $\boldsymbol{x}\colon I \to \mathbb{R}^n$ so that*

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t),$$

and $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$.

Proof We demonstrate existence by explicitly constructing a solution. Indeed, we claim that the solution is

$$\boldsymbol{x}(t) = e^{\boldsymbol{A}(t-t_0)}\boldsymbol{x}_0 + \int_{t_0}^{t} e^{\boldsymbol{A}(t-\tau)}\boldsymbol{B}\boldsymbol{u}(\tau)\,\mathrm{d}\tau. \tag{2.3}$$

First, note that the initial conditions are satisfied (just let $t = t_0$). We also compute

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= \boldsymbol{A}e^{\boldsymbol{A}(t-t_0)}\boldsymbol{x}_0 + \boldsymbol{B}\boldsymbol{u}(t) + \int_{t_0}^{t} \boldsymbol{A}e^{\boldsymbol{A}(t-\tau)}\boldsymbol{B}\boldsymbol{u}(\tau)\,\mathrm{d}\tau\\ &= \boldsymbol{A}e^{\boldsymbol{A}(t-t_0)}\boldsymbol{x}_0 + \boldsymbol{A}\int_{t_0}^{t} e^{\boldsymbol{A}(t-\tau)}\boldsymbol{B}\boldsymbol{u}(\tau)\,\mathrm{d}\tau + \boldsymbol{B}\boldsymbol{u}(t)\\ &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t).\end{aligned}$$

Thus $\boldsymbol{x}(t)$ as defined by (2.3) is indeed a solution.

Now we show that $\boldsymbol{x}(t)$ as defined by (2.3) is the only solution with the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$. We suppose that $\tilde{\boldsymbol{x}}(t)$ is a solution to the same initial value problem. Therefore, $\boldsymbol{z}(t) = \tilde{\boldsymbol{x}}(t) - \boldsymbol{x}(t)$ satisfies

$$\dot{\boldsymbol{z}}(t) = \dot{\tilde{\boldsymbol{x}}}(t) - \dot{\boldsymbol{x}}(t) = \boldsymbol{A}\tilde{\boldsymbol{x}}(t) - \boldsymbol{B}\boldsymbol{u}(t) - \boldsymbol{A}\boldsymbol{x}(t) - \boldsymbol{B}\boldsymbol{u}(t) = \boldsymbol{A}\boldsymbol{z}(t).$$

Since $\boldsymbol{x}(t_0) = \tilde{\boldsymbol{x}}(t_0)$ this means that $\boldsymbol{z}(t_0) = 0$. That is, $\boldsymbol{z}(t)$ is a solution to the initial value problem

$$\dot{\boldsymbol{z}}(t) = \boldsymbol{A}\boldsymbol{z}(t), \quad \boldsymbol{z}(t_0) = 0. \tag{2.4}$$

Let us multiply the differential equation on each side by $2\boldsymbol{z}^t(t)$:

$$2\boldsymbol{z}(t)^t\dot{\boldsymbol{z}}(t) = \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{z}^t(t)\boldsymbol{z}(t)) = \frac{\mathrm{d}}{\mathrm{d}t}(\|\boldsymbol{z}(t)\|^2) = 2\boldsymbol{z}^t(t)\boldsymbol{A}\boldsymbol{z}(t).$$

We now note that

$$\begin{aligned}2\boldsymbol{z}^t(t)\boldsymbol{A}\boldsymbol{z}(t) &= 2\sum_{i,j=1}^{n} z_i(t)a_{ij}z_j(t)\\ &\le 2\sum_{i,j=1}^{n}\|\boldsymbol{z}(t)\|\max_{i,j}|a_{ij}|\|\boldsymbol{z}(t)\|\\ &\le 2n^2\max_{i,j}|a_{ij}|\|\boldsymbol{z}(t)\|^2.\end{aligned}$$

Let $\alpha = 2n^2\max_{i,j}|a_{ij}|$ so that we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\|\boldsymbol{z}(t)\|^2) - \alpha\|\boldsymbol{z}(t)\|^2 \le 0.$$

We write

$$\begin{aligned}&e^{-\alpha t}\Big(\frac{\mathrm{d}}{\mathrm{d}t}(\|\boldsymbol{z}(t)\|^2) - \alpha\|\boldsymbol{z}(t)\|^2\Big) \le 0\\ \implies\quad &\frac{\mathrm{d}}{\mathrm{d}t}(e^{-\alpha t}\|\boldsymbol{z}(t)\|^2) \le 0.\end{aligned}$$

This can be integrated to give

$$e^{-\alpha t}\|\boldsymbol{z}(t)\|^2 - e^{-\alpha t_0}\|\boldsymbol{z}(t_0)\|^2 \leq 0$$

for all $t \in I$. Since $\boldsymbol{z}(t_0) = \boldsymbol{0}$ we must have

$$e^{-\alpha t}\|\boldsymbol{z}(t)\|^2 \leq 0, \quad t \in I.$$

Since $e^{-\alpha t} > 0$ this must mean that $\|\boldsymbol{z}(t)\|^2 = 0$ for all $t \in I$ and so $\boldsymbol{z}(t) = \boldsymbol{0}$ for all $t \in I$. But this means that $\tilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t)$, and so solutions are unique. ■

2.7 Remark As per Remark 2.4, if we suppose that $\boldsymbol{u}(t)$ is essentially bounded and measurable, then Theorem 2.6 still holds. •

Of course, once we have the solution for the state variable $\boldsymbol{x}$, it is a simple matter to determine the output $\boldsymbol{y}$:

$$\boldsymbol{y}(t) = \boldsymbol{C}e^{\boldsymbol{A}(t-t_0)}\boldsymbol{x}_0 + \int_{t_0}^{t} \boldsymbol{C}e^{\boldsymbol{A}(t-\tau)}\boldsymbol{B}\boldsymbol{u}(\tau)\,\mathrm{d}\tau + \boldsymbol{D}\boldsymbol{u}(t).$$

Our aim in this book is to study the response of the output $\boldsymbol{y}(t)$ to various inputs $\boldsymbol{u}(t)$, and to devise systematic ways to make the output do things we like.

Let's look at an example.

2.8 Example We return to our mass-spring-damper example. Let us be concrete for simplicity, and choose $m = 1$, $k = 4$, and $d = 0$. The system equations are then

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ -4 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix}u(t).$$

We compute (I use Mathematica®)

$$e^{\boldsymbol{A}t} = \begin{bmatrix}\cos 2t & \frac{1}{2}\sin 2t\\ -2\sin 2t & \cos 2t\end{bmatrix}.$$

Let us suppose that

$$u(t) = \begin{cases}1, & t \geq 0\\ 0, & \text{otherwise.}\end{cases}$$

Thus the input is a step function. Let us suppose we have zero initial condition $\boldsymbol{x}(0) = \boldsymbol{0}$. We then compute

$$\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} = \int_0^t \begin{bmatrix}\cos 2(t-\tau) & \frac{1}{2}\sin 2(t-\tau)\\ -2\sin 2(t-\tau) & \cos 2(t-\tau)\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix}\mathrm{d}\tau = \begin{bmatrix}\frac{1}{4}(1-\cos 2t)\\ \frac{1}{2}\sin 2t\end{bmatrix}.$$

The phase portrait of this curve is shown in Figure 2.1.

As far as outputs are concerned, recall that we had in Example 2.2 considered three cases. With the parameters we have chosen, these are as follows.

1. In the first case we measure displacement and so arrived at

$$\boldsymbol{c} = \begin{bmatrix}1\\ 0\end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix}0\end{bmatrix}.$$

The output is then computed to be

$$y(t) = \tfrac{1}{4}(1 - \cos 2t)$$

which we plot in Figure 2.2.

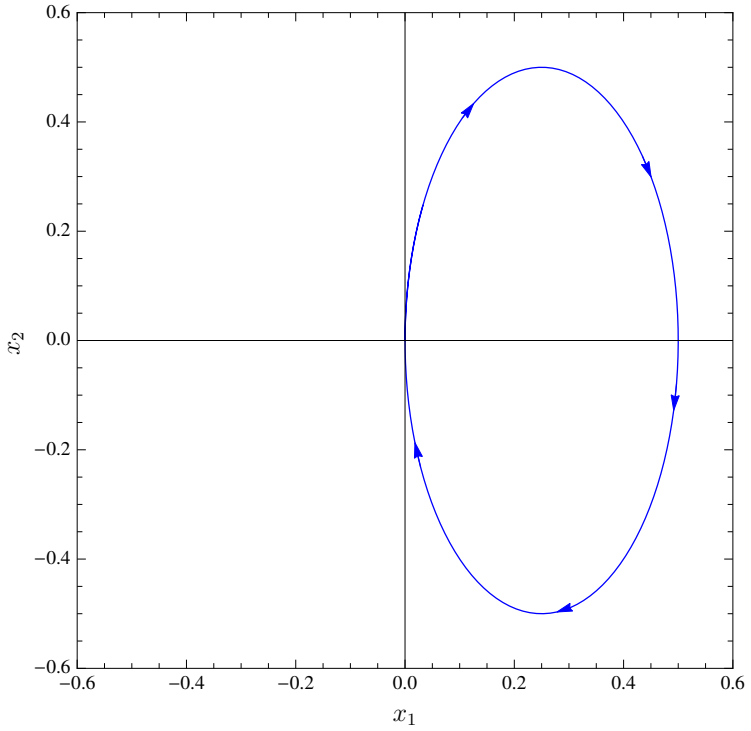


Figure 2.1 Phase curve for step response of mass-spring system

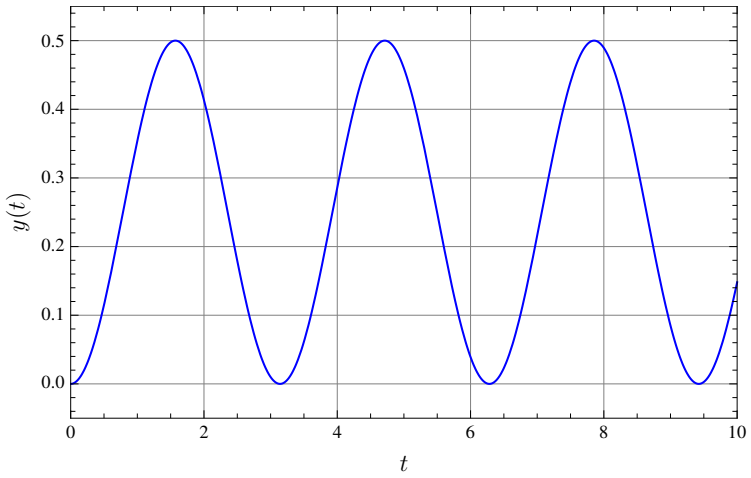


Figure 2.2 Displacement output for mass-spring system

2. If we measure velocity we have

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 0 \end{bmatrix}.$$

The output here is

$$y(t) = \tfrac{1}{2} \sin 2t$$

which we plot in Figure 2.3.

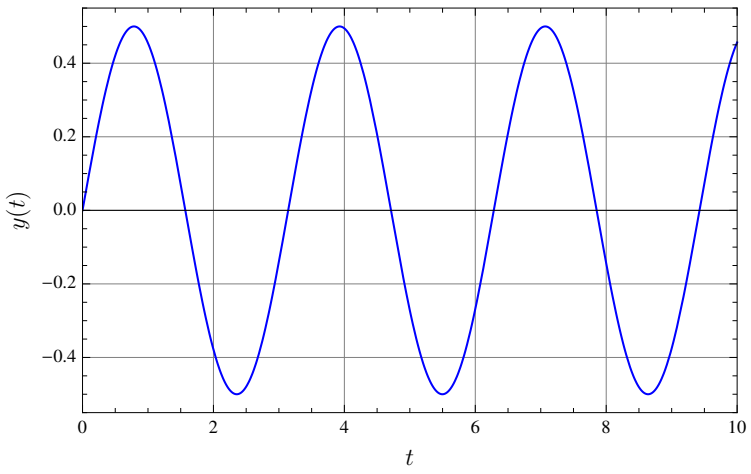


Figure 2.3 Velocity output for mass-spring system

3. The final case was when the output was acceleration, and we then derived

$$\boldsymbol{c} = \begin{bmatrix} -4 \\ 0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 1 \end{bmatrix}.$$

One readily ascertains

$$y(t) = \cos 2t$$

which we plot in Figure 2.4. •

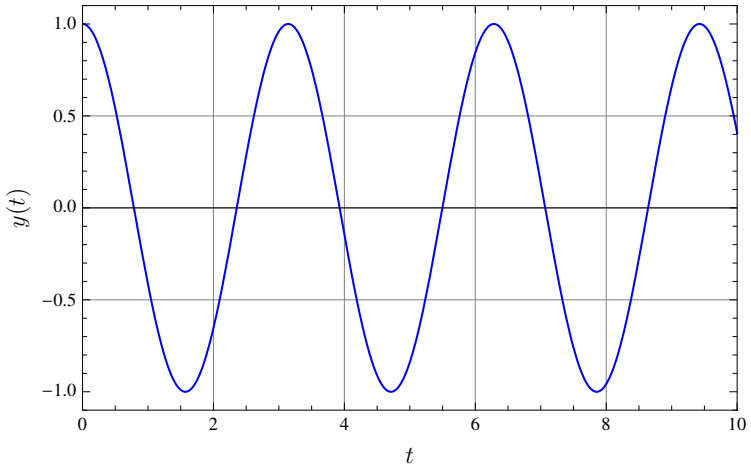


Figure 2.4 Acceleration output for mass-spring system

2.2 Obtaining linearised equations for nonlinear input/output systems

The example of the mass-spring-damper system is easy to put into the form of (2.2) since the equations are already linear. For a nonlinear system, we have seen in Section 1.4 how to linearise nonlinear differential equations about an equilibrium. Now let's see how to linearise a nonlinear input/output system. We first need to say what we mean by a nonlinear input/output system. We shall only consider SISO versions of these.

2.9 Definition A ***SISO nonlinear system*** consists of a pair $(\boldsymbol{f}, h)$ where $\boldsymbol{f}\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and $h\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ are smooth maps. •

What are the control equations here? They are

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}, u) \\ y &= h(\boldsymbol{x}, u). \end{aligned} \tag{2.5}$$

This is a generalisation of the linear equations (2.2). For systems like this, is it no longer obvious that solutions exist or are unique as we asserted in Theorem 2.6 for linear systems. We do not really get into such issues here as they do not comprise an essential part of what we are doing. We are interested in linearising the equations (2.5) about an equilibrium point. Since the system now has controls, we should revise our notion of what an equilibrium point means. To wit, an ***equilibrium point*** for a SISO nonlinear system $(\boldsymbol{f}, h)$ is a pair $(\boldsymbol{x}_0, u_0) \in \mathbb{R}^n \times \mathbb{R}$ so that $\boldsymbol{f}(\boldsymbol{x}_0, u_0) = \boldsymbol{0}$. The idea is the same as the idea for an equilibrium point for a differential equation, except that we now allow the control to enter the picture. To linearise, we linearise with respect to both $\boldsymbol{x}$ and u, evaluating at the equilibrium point. In doing this, we arrive at the following definition.

2.10 Definition Let $(\boldsymbol{f}, h)$ be a SISO nonlinear system and let $(\boldsymbol{x}_0, u_0)$ be an equilibrium point for the system. The ***linearisation*** of (2.5) about $(\boldsymbol{x}_0, u_0)$ is the SISO linear system $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ where

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{D}\boldsymbol{f}(\boldsymbol{x}_0, u_0) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \\ \boldsymbol{b} &= \frac{\partial \boldsymbol{f}}{\partial u}(\boldsymbol{x}_0, u_0) = \begin{bmatrix} \frac{\partial f_1}{\partial u} \\ \frac{\partial f_2}{\partial u} \\ \vdots \\ \frac{\partial f_n}{\partial u} \end{bmatrix} \\ \boldsymbol{c}^t &= \boldsymbol{D}h(\boldsymbol{x}_0, u_0) = \begin{bmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} & \cdots & \frac{\partial h}{\partial x_n} \end{bmatrix} \\ \boldsymbol{D} &= \frac{\partial h}{\partial u}(\boldsymbol{x}_0, u_0), \end{aligned}$$

where all partial derivatives are evaluated at $(\boldsymbol{x}_0, u_0)$. •

2.11 Note Let us suppose for simplicity that all equilibrium points we consider will necessitate that $u_0 = 0$. •

Let us do this for the pendulum.

2.12 Example **(Example 1.3 cont'd)** We consider a torque applied at the pendulum pivot and we take as output the pendulum's angular velocity.

Let us first derive the form for the first of equations (2.5). We need to be careful in deriving the vector function $\boldsymbol{f}$. The forces should be added to the equation at the outset, and *then* the equations put into first-order form and linearised. Recall that the equations for the pendulum, just ascertained by force balance, are

$$m\ell^2\ddot{\theta} + mg\ell \sin\theta = 0.$$

It is to these equations, and not any others, that the external torque should be added since the external torque should obviously appear in the force balance equation. If the external torque is u, the forced equations are simply

$$m\ell^2\ddot{\theta} + mg\ell\sin\theta = u.$$

We next need to put this into the form of the first of equations (2.5). We first divide by $m\ell^2$ and get

$$\ddot{\theta} + \tfrac{g}{\ell}\sin\theta = \frac{u}{m\ell^2}.$$

To put this into first-order form we define, as usual, $(x_1, x_2) = (\theta, \dot{\theta})$. We then have

$$\begin{aligned}\dot{x}_1 &= \dot{\theta} = x_2\\ \dot{x}_2 &= \ddot{\theta} = -\tfrac{g}{\ell}\sin\theta + \tfrac{1}{m\ell^2}u = -\tfrac{g}{\ell}\sin x_1 + \tfrac{1}{m\ell^2}u\end{aligned}$$

so that

$$\boldsymbol{f}(\boldsymbol{x}, u) = \begin{bmatrix} x_2 \\ -\frac{g}{\ell}\sin x_1 + \frac{1}{m\ell^2}u \end{bmatrix}.$$

By a suitable choice for u_0, any point of the form $(x_1, 0)$ can be rendered an equilibrium point. Let us simply look at those for which $u_0 = 0$, as we suggested we do in Note 2.11. We determine that such equilibrium points are of the form $(\boldsymbol{x}_0, u_0) = ((\theta_0, 0), 0)$, $\theta_0 \in \{0, \pi\}$. We then have the linearised state matrix $\boldsymbol{A}$ and the linearised input vector $\boldsymbol{b}$ given by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{\ell}\cos\theta_0 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ \frac{1}{m\ell^2} \end{bmatrix}.$$

The output is easy to handle in this example. We have $h(\boldsymbol{x}, u) = \dot{\theta} = x_2$. Therefore

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1.$$

Putting all this together gives us the linearisation as

$$\begin{aligned}\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ -\frac{g}{\ell}\cos\theta_0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m\ell^2} \end{bmatrix} u\\ y &= \begin{bmatrix} 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.\end{aligned}$$

We then substitute $\theta_0 = 0$ or $\theta_0 = \pi$ depending on whether the system is in the "down" or "up" equilibrium. •

2.3 Input/output response versus state behaviour

In this section we will consider only SISO systems, and we will suppose that the 1×1 *matrix* $\boldsymbol{D}$ *is zero. The first restriction is easy to relax, but the second may not be, depending on what one wishes to do. However, often in applications* $\boldsymbol{D} = \boldsymbol{0}_1$ *in any case.*

We wish to reveal the problems that may be encountered when one focuses on input/output behaviour without thinking about the system states. That is to say, if we restrict our attention to designing inputs $u(t)$ that make the output $y(t)$ behave in a desirable manner, problems may arise. If one has a state-space model like (2.2), then it is possible that while you are making the outputs behave nicely, some of the states in $\boldsymbol{x}(t)$ may be misbehaving badly. This is perhaps best illustrated with a sequence of simple examples. Each of these examples will illustrate an important concept, and after each example, the general idea will be discussed.

2.3.1 Bad behaviour due to lack of observability

We first look at an example that will introduce us to the concept of "observability."

2.13 Example We first consider the system

$$\begin{aligned}\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} &= \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix} u\\ y &= \begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}.\end{aligned} \tag{2.6}$$

We compute

$$e^{\boldsymbol{A}t} = \begin{bmatrix}\frac{1}{2}(e^t + e^{-t}) & \frac{1}{2}(e^t - e^{-t})\\ \frac{1}{2}(e^t - e^{-t}) & \frac{1}{2}(e^t + e^{-t})\end{bmatrix}$$

and so, if we use the initial condition $\boldsymbol{x}(0) = \boldsymbol{0}$, and the input

$$u(t) = \begin{cases}1, & t \geq 0\\ 0, & \text{otherwise,}\end{cases}$$

we get

$$\boldsymbol{x}(t) = \int_0^t \begin{bmatrix}\frac{1}{2}(e^{t-\tau} - e^{-t+\tau}) & \frac{1}{2}(e^{t-\tau} - e^{-t+\tau})\\ \frac{1}{2}(e^{t-\tau} - e^{-t+\tau}) & \frac{1}{2}(e^{t-\tau} + e^{-t+\tau})\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix} \mathrm{d}\tau = \begin{bmatrix}\frac{1}{2}(e^t + e^{-t}) - 1\\ \frac{1}{2}(e^t - e^{-t})\end{bmatrix}.$$

One also readily computes the output as

$$y(t) = e^{-t} - 1$$

which we plot in Figure 2.5. Note that the output is behaving quite nicely, thank you.

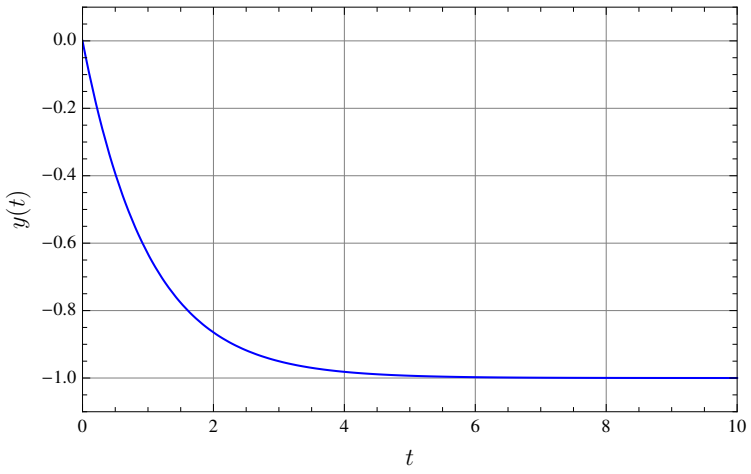


Figure 2.5 Output response of (2.6) to a step input

However, the state is going nuts, blowing up to ∞ as $t \to \infty$ as shown in Figure 2.6. •

What is the problem here? Well, looking at what is going on with the equations reveals the problem. The poor state-space behaviour is obviously present in the equations for the state variable $\boldsymbol{x}(t)$. However, when we compute the output, this bad behaviour gets killed by the output vector $\boldsymbol{c}^t$. There is a mechanism to describe what is going on, and it is called "observability theory." We only talk in broad terms about this here.

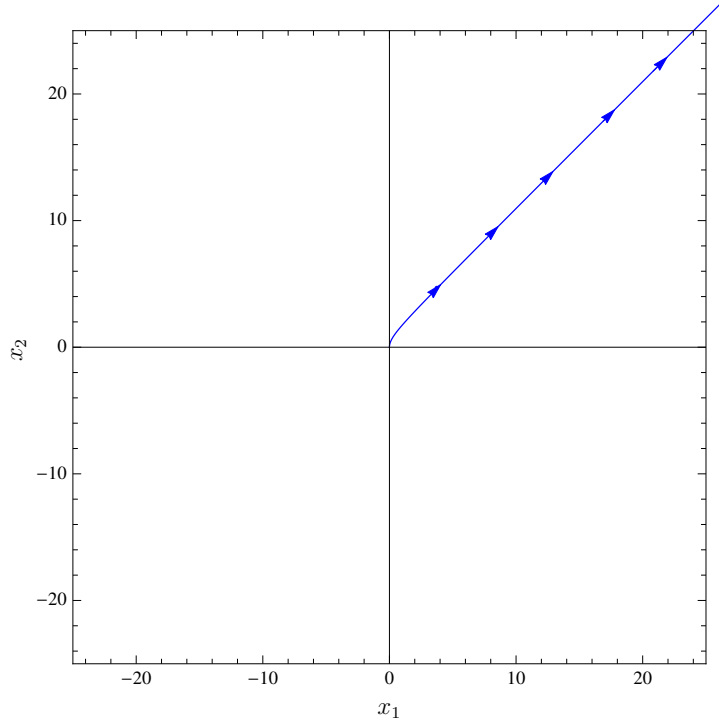


Figure 2.6 State-space behaviour of (2.6) with a step input

2.14 Definition A pair $(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is ***observable*** if the matrix

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} \boldsymbol{c}^t \\ \hline \boldsymbol{c}^t\boldsymbol{A} \\ \hline \vdots \\ \hline \boldsymbol{c}^t\boldsymbol{A}^{n-1} \end{bmatrix}$$

has full rank. If $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, then Σ is ***observable*** if $(\boldsymbol{A}, \boldsymbol{c})$ is observable. The matrix $\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c})$ is called the ***observability matrix*** for $(\boldsymbol{A}, \boldsymbol{c})$. •

The above definition carefully masks the "real" definition of observability. However, the following result provides the necessary connection with things more readily visualised.

2.15 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system, let $u_1, u_2 \in \mathscr{U}$, and let $\boldsymbol{x}_1(t)$, $\boldsymbol{x}_2(t)$, $y_1(t)$, and $y_2(t)$ be defined by*

$$\begin{aligned} \dot{\boldsymbol{x}}_i(t) &= \boldsymbol{A}\boldsymbol{x}_i(t) + \boldsymbol{b}u_i(t) \\ y_i(t) &= \boldsymbol{c}^t\boldsymbol{x}_i(t), \end{aligned}$$

$i = 1, 2$. The following statements are equivalent:

(i) *$(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is observable;*

(ii) *$u_1(t) = u_2(t)$ and $y_1(t) = y_2(t)$ for all t implies that $\boldsymbol{x}_1(t) = \boldsymbol{x}_2(t)$ for all t.*

Proof Let us first show that the second condition is equivalent to saying that the output with zero input is in one-to-one correspondence with the initial condition. Indeed, for arbitrary

inputs $u_1, u_2 \in \mathscr{U}$ with corresponding states $\boldsymbol{x}_1(t)$ and $\boldsymbol{x}_2(t)$, and outputs $y_1(t)$ and $y_2(t)$ we have

$$y_i(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{x}_i(0) + \int_0^t \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}u_i(\tau)\,\mathrm{d}\tau,$$

$i = 1, 2$. If we define $z_i(t)$ by

$$z_i(t) = y_i(t) - \int_0^t \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}u_i(\tau)\,\mathrm{d}\tau,$$

$i = 1, 2$, then we see that $u_1 = u_2$ and $y_1 = y_2$ is equivalent to $z_1 = z_2$. However, since $z_i(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{x}_i(0)$, this means that the second condition of the theorem is equivalent to the statement that equal outputs for zero inputs implies equal initial conditions for the state.

First suppose that $(\boldsymbol{c}, \boldsymbol{A})$ is observable, and let us suppose that $z_1(t) = z_2(t)$, with z_1 and z_2 as above. Therefore we have

$$\begin{bmatrix} z_1(0) \\ z_1^{(1)}(0) \\ \vdots \\ z_1^{(n-1)}(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}^t \\ \boldsymbol{c}^t\boldsymbol{A} \\ \vdots \\ \boldsymbol{c}^t\boldsymbol{A}^{n-1} \end{bmatrix} \boldsymbol{x}_1(0) = \begin{bmatrix} z_2(0) \\ z_2^{(1)}(0) \\ \vdots \\ z_2^{(n-1)}(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}^t \\ \boldsymbol{c}^t\boldsymbol{A} \\ \vdots \\ \boldsymbol{c}^t\boldsymbol{A}^{n-1} \end{bmatrix} \boldsymbol{x}_2(0).$$

However, since $(\boldsymbol{A}, \boldsymbol{c})$ is observable, this gives

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c})\boldsymbol{x}_1(0) = \boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c})\boldsymbol{x}_2(0) \quad \Longrightarrow \quad \boldsymbol{x}_1(0) = \boldsymbol{x}_2(0),$$

which is, as we have seen, equivalent to the assertion that $u_1 = u_2$ and $y_1 = y_2$ implies that $\boldsymbol{x}_1 = \boldsymbol{x}_2$.

Now suppose that $\operatorname{rank}(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c})) \neq n$. Then there exists a nonzero vector $\boldsymbol{x}_0 \in \mathbb{R}^n$ so that $\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c})\boldsymbol{x}_0 = \boldsymbol{0}$. By the Cayley-Hamilton Theorem it follows that $\boldsymbol{c}^t\boldsymbol{A}^k\boldsymbol{x}_1(t) = 0$, $k \geq 1$, if $\boldsymbol{x}_1(t)$ is the solution to the initial value problem

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

Now the series representation for the matrix exponential gives $z_1(t) = 0$ where $z_1(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{x}_0$. However, we also have $z_2(t) = 0$ if $z_2(t) = \boldsymbol{c}^t\boldsymbol{0}$. However, $\boldsymbol{x}_2(t) = \boldsymbol{0}$ is the solution to the initial value problem

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t), \quad \boldsymbol{x}(0) = \boldsymbol{0},$$

from which we infer that we cannot infer the initial conditions from the output with zero input. ■

The idea of observability is, then, that one can infer the initial condition for the state from the input and the output. Let us illustrate this with an example.

2.16 Example **(Example 2.13 cont'd)** We compute the observability matrix for the system in Example 2.13 to be

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

which is not full rank (it has rank 1). Thus the system is not observable.

Now suppose that we start out the system (2.6) with not the zero initial condition, but with the initial condition $\boldsymbol{x}(0) = (1, 1)$. A simple calculation shows that the output is then $y(t) = e^{-t} - 1$, which is just as we saw with zero initial condition. Thus our output was unable to "see" this change in initial condition, and so this justifies our words following Definition 2.14. You might care to notice that the initial condition $(1, 1)$ is in the kernel of the matrix $\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c})$! •

The following property of the observability matrix will be useful for us.

2.17 Theorem *The kernel of the matrix* $\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})$ *is the largest* $\boldsymbol{A}$*-invariant subspace contained in* $\ker(\boldsymbol{c}^t)$.

Proof First let us show that the kernel of $\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})$ is contained in $\ker(\boldsymbol{c}^t)$. If $\boldsymbol{x} \in \ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$ then

$$\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})\boldsymbol{x} = \left[\begin{array}{c} \boldsymbol{c}^t \\ \hline \boldsymbol{c}^t\boldsymbol{A} \\ \hline \vdots \\ \hline \boldsymbol{c}^t\boldsymbol{A}^{n-1} \end{array}\right] \boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and in particular, $\boldsymbol{c}^t\boldsymbol{x} = 0$—that is, $\boldsymbol{x} \in \ker(\boldsymbol{c}^t)$.

Now we show that the kernel of $\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})$ is $\boldsymbol{A}$-invariant. Let $\boldsymbol{x} \in \ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$ and then compute

$$\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})\boldsymbol{A}\boldsymbol{x} = \left[\begin{array}{c} \boldsymbol{c}^t \\ \hline \boldsymbol{c}^t\boldsymbol{A} \\ \hline \vdots \\ \hline \boldsymbol{c}^t\boldsymbol{A}^{n-1} \end{array}\right] \boldsymbol{A}\boldsymbol{x} = \left[\begin{array}{c} \boldsymbol{c}^t\boldsymbol{A} \\ \hline \boldsymbol{c}^t\boldsymbol{A}^{2} \\ \hline \vdots \\ \hline \boldsymbol{c}^t\boldsymbol{A}^{n} \end{array}\right] \boldsymbol{x}.$$

Since $\boldsymbol{x} \in \ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$ we have

$$\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{x} = 0, \ldots, \boldsymbol{c}^t\boldsymbol{A}^{n-1}\boldsymbol{x} = 0.$$

Also, by the Cayley-Hamilton Theorem,

$$\boldsymbol{c}^t\boldsymbol{A}^n\boldsymbol{x} = -p_{n-1}\boldsymbol{c}^t\boldsymbol{A}^{n-1}\boldsymbol{x} - \cdots - p_1\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{x} - p_0\boldsymbol{c}^t\boldsymbol{x} = 0.$$

This shows that

$$\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})\boldsymbol{x} = \boldsymbol{0}$$

or $\boldsymbol{x} \in \ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$.

Finally, we show that if V is an $\boldsymbol{A}$-invariant subspace contained in $\ker(\boldsymbol{c}^t)$, then V is a subspace of $\ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$. Given such a V and $\boldsymbol{x} \in V$, $\boldsymbol{c}^t\boldsymbol{x} = 0$. Since V is $\boldsymbol{A}$-invariant, $\boldsymbol{A}\boldsymbol{x} \in V$, and since V is contained in $\ker(\boldsymbol{c}^t)$, $\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{x} = 0$. Proceeding in this way we see that $\boldsymbol{c}^t\boldsymbol{A}^2\boldsymbol{x} = \cdots = \boldsymbol{c}^t\boldsymbol{A}^{n-1}\boldsymbol{x} = 0$. But this means exactly that $\boldsymbol{x}$ is in $\ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$. ■

The subspace $\ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$ has a simple interpretation in terms of Theorem 2.15. It turns out that if two state initial conditions $\boldsymbol{x}_1(0)$ and $\boldsymbol{x}_2(0)$ differ by a vector in $\ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$, i.e., if $\boldsymbol{x}_2(0) - \boldsymbol{x}_1(0) \in \ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$, then the same input will produce the same output for these different initial conditions. This is exactly what we saw in Example 2.16.

2.18 Remark Although our discussion in this section has been for SISO systems $\Sigma = (\boldsymbol{A},\boldsymbol{b},\boldsymbol{c}^t,\boldsymbol{D})$, it can be easily extended to MIMO systems. Indeed our characterisations of observability in Theorems 2.15 and 2.17 are readily made for MIMO systems. Also, for a MIMO system $\Sigma = (\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ one can certainly define

$$\boldsymbol{O}(\boldsymbol{A},\boldsymbol{C}) = \left[\begin{array}{c} \boldsymbol{C} \\ \hline \boldsymbol{C}\boldsymbol{A} \\ \hline \vdots \\ \hline \boldsymbol{C}\boldsymbol{A}^{n-1} \end{array}\right],$$

and one may indeed verify that the appropriate versions of Theorems 2.15 and 2.17 hold in this case. •

2.3.2 Bad behaviour due to lack of controllability

Okay, so we believe that a lack of observability may be the cause of problems in the state, regardless of the good behaviour of the input/output map. Are there other ways in which things can go awry? Well, yes there are. Let us look at a system that *is* observable, but that does not behave nicely.

2.19 Example Here we look at

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \\ y &= \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \end{aligned} \tag{2.7}$$

This system is observable as the observability matrix is

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}$$

which has rank 2. We compute

$$e^{\boldsymbol{A}t} = \begin{bmatrix} e^t & 0 \\ \frac{1}{2}(e^t - e^{-t}) & e^{-t} \end{bmatrix}$$

from which we ascertain that with zero initial conditions, and a unit step input,

$$\boldsymbol{x}(t) = \begin{bmatrix} 0 \\ 1 - e^{-t} \end{bmatrix}, \quad y(t) = 1 - e^{-t}.$$

Okay, this looks fine. Let's change the initial condition to $\boldsymbol{x}(0) = (1, 0)$. We then compute

$$\boldsymbol{x}(t) = \begin{bmatrix} e^t \\ 1 + \frac{1}{2}(e^t - 3e^{-t}) \end{bmatrix}, \quad y(t) = 1 + \tfrac{1}{2}(e^t - 3e^{-t}).$$

Well, since the system is observable, it can sense this change of initial condition, and how! As we see in Figure 2.7 (where we depict the output response) and Figure 2.8 (where we depict the state behaviour), the system is now blowing up in both state and output. •

It's not so hard to see what is happening here. We do not have the ability to "get at" the unstable dynamics of the system with our input. Motivated by this, we come up with another condition on the linear system, different from observability.

2.20 Definition A pair $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n \times n} \times \mathbb{R}^n$ is ***controllable*** if the matrix

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \left[\begin{array}{c|c|c|c} \boldsymbol{b} & \boldsymbol{A}\boldsymbol{b} & \cdots & \boldsymbol{A}^{n-1}\boldsymbol{b} \end{array} \right]$$

has full rank. If $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, then Σ is ***controllable*** if $(\boldsymbol{A}, \boldsymbol{b})$ is controllable. The matrix $\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})$ is called the ***controllability matrix*** for $(\boldsymbol{A}, \boldsymbol{b})$. •

Let us state the result that gives the intuitive meaning for our definition for controllability.

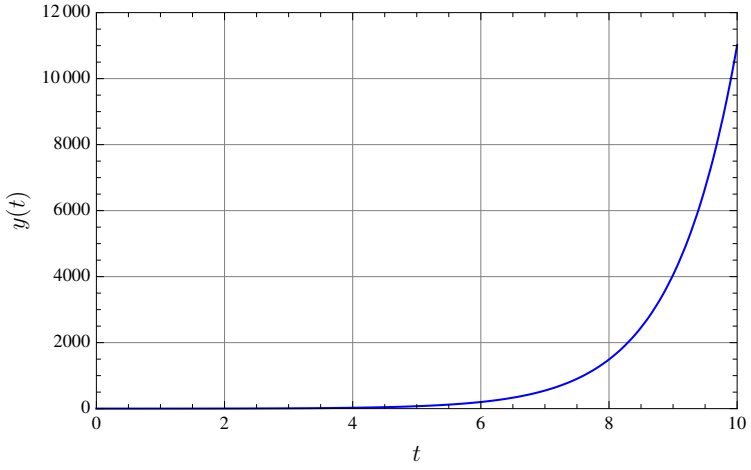


Figure 2.7 The output response of (2.7) with a step input and non-zero initial condition

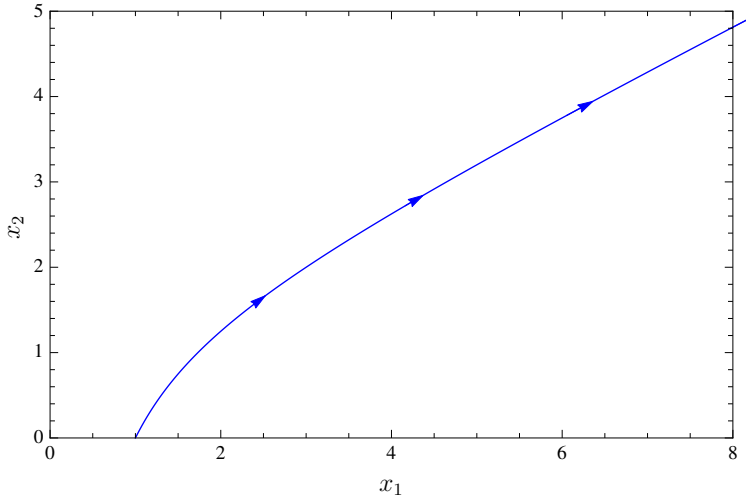


Figure 2.8 The state-space behaviour of (2.7) with a step input and non-zero initial condition

2.21 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system. The pair $(\boldsymbol{A}, \boldsymbol{b})$ is controllable if and only if for each $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ and for each $T > 0$, there exists an admissible input $u\colon [0, T] \to \mathbb{R}$ with the property that if $\boldsymbol{x}(t)$ is the solution to the initial value problem*

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t), \quad \boldsymbol{x}(0) = \boldsymbol{x}_1,$$

then $\boldsymbol{x}(T) = \boldsymbol{x}_2$.

Proof For $t > 0$ define the matrix $\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{b})(t)$ by

$$\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{b})(t) = \int_0^t e^{\boldsymbol{A}\tau} \boldsymbol{b}\boldsymbol{b}^t e^{\boldsymbol{A}^t\tau} \, \mathrm{d}\tau.$$

Let us first show that $\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})$ is invertible if and only if $\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{b})(t)$ is positive-definite for all $t > 0$ (we refer ahead to Section 5.4.1 for notions of definiteness of matrices). Since

$\boldsymbol{P}(\boldsymbol{A},\boldsymbol{b})(t)$ is clearly positive-semidefinite, this means we shall show that $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$ is invertible if and only if $\boldsymbol{P}(\boldsymbol{A},\boldsymbol{b})(t)$ is invertible. Suppose that $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$ is not invertible. Then there exists a nonzero $\boldsymbol{x}_0 \in \mathbb{R}^n$ so that $\boldsymbol{x}_0^t\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})) = \boldsymbol{0}$. By the Cayley-Hamilton Theorem, this implies that $\boldsymbol{x}_0^t\boldsymbol{A}^k\boldsymbol{b} = 0$ for $k \geq 1$. This in turn means that $\boldsymbol{x}_0^t e^{\boldsymbol{A}t}\boldsymbol{b} = 0$ for $t > 0$. Therefore, since $e^{\boldsymbol{A}^t t} = (e^{\boldsymbol{A}t})^t$, it follows that

$$\boldsymbol{x}_0^t e^{\boldsymbol{A}t}\boldsymbol{b}\boldsymbol{b}^t e^{\boldsymbol{A}^t t}\boldsymbol{x}_0 = 0.$$

Thus $\boldsymbol{P}(\boldsymbol{A},\boldsymbol{b})(t)$ is not invertible.

Now suppose that there exists $T > 0$ so that $\boldsymbol{P}(\boldsymbol{A},\boldsymbol{b})(T)$ is not invertible. Therefore there exists a nonzero $\boldsymbol{x}_0 \in \mathbb{R}^n$ so that $\boldsymbol{x}_0^t e^{\boldsymbol{A}t}\boldsymbol{b} = 0$ for $t \in [0,T]$. Differentiating this $n-1$ times with respect to t and evaluating at $t = 0$ gives

$$\boldsymbol{x}_0\boldsymbol{b} = \boldsymbol{x}_0\boldsymbol{A}\boldsymbol{b} = \cdots = \boldsymbol{x}_0\boldsymbol{A}^{n-1}\boldsymbol{b} = 0.$$

This, however, infers the existence of a nonzero vector in $\ker(\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b}))$, giving us our initial claim.

Let us now show how this claim gives the theorem. First suppose that $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$ is invertible so that $\boldsymbol{P}(\boldsymbol{A},\boldsymbol{b})(t)$ is positive-definite for all $t > 0$. One may then directly show, with a slightly tedious computation, that if we define a control $u\colon [0,T] \to \mathbb{R}$ by

$$u(t) = -\boldsymbol{b}^t e^{\boldsymbol{A}^t(T-t)}\boldsymbol{P}(\boldsymbol{A},\boldsymbol{b})^{-1}(T)\big(e^{\boldsymbol{A}T}\boldsymbol{x}_1 - \boldsymbol{x}_2\big),$$

then the solution to the initial value problem

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t), \quad \boldsymbol{x}(0) = \boldsymbol{x}_1$$

has the property that $\boldsymbol{x}(T) = \boldsymbol{x}_2$.

Now suppose that $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$ is not invertible so that there exists $T > 0$ so that $\boldsymbol{P}(\boldsymbol{A},\boldsymbol{b})(T)$ is not invertible. Thus there exists a nonzero $\boldsymbol{x}_0 \in \mathbb{R}^n$ so that

$$\boldsymbol{x}_0^t e^{\boldsymbol{A}t}\boldsymbol{b} = 0, \quad t \in [0,T]. \tag{2.8}$$

Let $\boldsymbol{x}_1 = e^{-\boldsymbol{A}T}\boldsymbol{x}_0$ and let u be an admissible control. If the resulting state vector is $\boldsymbol{x}(t)$, we then compute

$$\boldsymbol{x}(T) = e^{\boldsymbol{A}T}e^{-\boldsymbol{A}T}\boldsymbol{x}_0 + \int_0^T e^{\boldsymbol{A}(T-\tau)}\boldsymbol{b}u(\tau)\,\mathrm{d}\tau.$$

Using (2.8), we have

$$\boldsymbol{x}_0^t\boldsymbol{x}(T) = \boldsymbol{x}_0^t\boldsymbol{x}_0.$$

Therefore, it is not possible to find a control for which $\boldsymbol{x}(T) = \boldsymbol{0}$. ■

This test of controllability for linear systems was obtained by Kalman, Ho, and Narendra [1963]. The idea is quite simple to comprehend: controllability reflects that we can reach any state from any other state. We can easily see how this comes up in an example.

2.22 Example **(Example 2.19 cont'd)** We compute the controllability matrix for Example 2.19 to be

$$\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b}) = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}$$

which has rank $1 < 2$ and so the system is not controllable.

Let's see how this meshes with what we said following Definition 2.20. Suppose we start at $\boldsymbol{x}(0) = (0,0)$. Since any solution to

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$
$$y = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

which has initial condition $x_1(0) = 0$ will have the property that $\boldsymbol{x}_1(t) = 0$ for all t, the control system is essentially governed by the x_2 equation:

$$\dot{x}_2 = -x_2 + u.$$

Therefore we can only move in the x_2-direction and all points with $x_1 \neq 0$ will not be accessible to us. This is what we mean by controllability. You might note that the set of points reachable are those in the columnspace of the matrix $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$. •

Based on the above discussion, we say that a triple $(\boldsymbol{A},\boldsymbol{b},\boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n \times \mathbb{R}^n$ is ***complete*** if $(\boldsymbol{A},\boldsymbol{b})$ is controllable and if $(\boldsymbol{A},\boldsymbol{c})$ is observable.

We have a property of the controllability matrix that is sort of like that for the observability matrix in Theorem 2.17.

2.23 Theorem *The columnspace of the matrix $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$ is the smallest $\boldsymbol{A}$-invariant subspace containing $\boldsymbol{b}$.*

Proof Obviously $\boldsymbol{b}$ is in the columnspace of $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$. We will show that this columnspace is $\boldsymbol{A}$-invariant. Let $\boldsymbol{x}$ be in the columnspace of $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$, i.e., in the range of the linear map $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$. Then there is a vector $\boldsymbol{y} \in \mathbb{R}^n$ with the property that

$$\boldsymbol{x} = \boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})\boldsymbol{y} = \left[\begin{array}{c|c|c|c} \boldsymbol{b} & \boldsymbol{A}\boldsymbol{b} & \cdots & \boldsymbol{A}^{n-1}\boldsymbol{b} \end{array}\right] \boldsymbol{y}.$$

We then have

$$\boldsymbol{A}\boldsymbol{x} = \left[\begin{array}{c|c|c|c} \boldsymbol{A}\boldsymbol{b} & \boldsymbol{A}^2\boldsymbol{b} & \cdots & \boldsymbol{A}^{n}\boldsymbol{b} \end{array}\right] \boldsymbol{y}.$$

The result will follow if we can show that each of the vectors

$$\boldsymbol{A}\boldsymbol{b}, \ldots, \boldsymbol{A}^n\boldsymbol{b}$$

is in the columnspace of $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$. It is clear that

$$\boldsymbol{A}\boldsymbol{b}, \ldots, \boldsymbol{A}^{n-1}\boldsymbol{b}$$

are in the columnspace of $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$. By the Cayley-Hamilton Theorem we have

$$\boldsymbol{A}^n\boldsymbol{b} = -p_{n-1}\boldsymbol{A}^{n-1}\boldsymbol{b} - \cdots - p_1\boldsymbol{A}\boldsymbol{b} - p_0\boldsymbol{b},$$

which shows that $\boldsymbol{A}^n\boldsymbol{b}$ is also in the columnspace of $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$.

Now we show that if V is an $\boldsymbol{A}$-invariant subspace with $\boldsymbol{b} \in V$ then V contains the columnspace of $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$. If V is such a subspace then $\boldsymbol{b} \in V$. Since V is $\boldsymbol{A}$-invariant, $\boldsymbol{A}\boldsymbol{b} \in V$. Proceeding in this way we ascertain that $\boldsymbol{A}^2\boldsymbol{b}, \ldots, \boldsymbol{A}^{n-1}\boldsymbol{b} \in V$, and therefore the columnspace of $\boldsymbol{C}(\boldsymbol{A},\boldsymbol{b})$ is contained in V. ■

There is a somewhat subtle thing happening here that should be understood. If a pair $(\boldsymbol{A}, \boldsymbol{b})$ is controllable, this implies that one can steer between any two points in $\mathbb{R}^n$ with a suitable control. It does *not* mean that one can follow *any* curve in $\mathbb{R}^n$ that connects the two points. This then raises the question, "What curves in $\mathbb{R}^n$ *can* be followed by solutions of the differential equation

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t)?"$$

Let us explore the answer to this question, following Basile and Marro [1968]. Because we will deal only with the single-input case, things are somewhat degenerate. Let $T(\boldsymbol{A}, \boldsymbol{b}) \subset \mathbb{R}^n$ be the subspace

$$T(\boldsymbol{A}, \boldsymbol{b}) = \begin{cases} \text{span}(\boldsymbol{b}), & \boldsymbol{b} \text{ is an eigenvector for } \boldsymbol{A} \\ \{\boldsymbol{0}\}, & \text{otherwise.} \end{cases}$$

The following lemma asserts the essential property of the subspace $T(\boldsymbol{A}, \boldsymbol{b})$.

2.24 Lemma *$T(\boldsymbol{A}, \boldsymbol{b})$ is the largest subspace of $\mathbb{R}^n$ with the property that*

$$\boldsymbol{A}(T(\boldsymbol{A}, \boldsymbol{b})) + T(\boldsymbol{A}, \boldsymbol{b}) \subset \text{span}(\boldsymbol{b}).$$

Proof First we note that $T(\boldsymbol{A}, \boldsymbol{b})$ does indeed have the property that $\boldsymbol{A}(T(\boldsymbol{A}, \boldsymbol{b}))+T(\boldsymbol{A}, \boldsymbol{b}) \subset \text{span}(\boldsymbol{b})$. This is clear if $T(\boldsymbol{A}, \boldsymbol{b}) = \{\boldsymbol{0}\}$. If $T(\boldsymbol{A}, \boldsymbol{b}) = \text{span}(\boldsymbol{b})$ then it is the case that $\boldsymbol{A}\boldsymbol{b} = \lambda \boldsymbol{b}$ for some $\lambda \in \mathbb{R}$. It then follows that if $\boldsymbol{x}_1 = a_1\boldsymbol{b}$ and $\boldsymbol{x}_2 = a_2\boldsymbol{b}$ for $a_1, a_2 \in \mathbb{R}$ then

$$\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{x}_2 = (a_1\lambda + a_2)\boldsymbol{b} \in \text{span}(\boldsymbol{b}).$$

Now we need to show that $T(\boldsymbol{A}, \boldsymbol{b})$ is the largest subspace with this property. Suppose that V is a subspace of $\mathbb{R}^n$ with the property that $\boldsymbol{A}(V) + V \subset \text{span}(\boldsymbol{b})$. Thus for each $\boldsymbol{x}_1, \boldsymbol{x}_2 \in V$ we have

$$\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{x}_2 \in \text{span}(\boldsymbol{b}).$$

In particular, if we choose $\boldsymbol{x}_1 = \boldsymbol{0}$ we see that if $\boldsymbol{x}_2 \in V$ then $\boldsymbol{x}_2 \in \text{span}(\boldsymbol{b})$. Similarly, if $\boldsymbol{x}_2 = \boldsymbol{0}$ we see that if $\boldsymbol{x}_1 \in V$ then $\boldsymbol{A}\boldsymbol{x}_1 \in \text{span}(\boldsymbol{b})$. Thus we have shown that if V is a subspace with the property that $\boldsymbol{A}(V) + V \subset \text{span}(\boldsymbol{b})$, this implies that

$$V = \text{span}(\boldsymbol{b}) \cap \boldsymbol{A}^{-1}(\text{span}(\boldsymbol{b}))$$

where

$$\boldsymbol{A}^{-1}(\text{span}(\boldsymbol{b})) = \{\boldsymbol{v} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{v} \in \text{span}(\boldsymbol{b})\}$$

(note that we are not saying that $\boldsymbol{A}$ is invertible!). It now remains to show that $T(\boldsymbol{A}, \boldsymbol{b}) = \text{span}(\boldsymbol{b}) \cap \boldsymbol{A}^{-1}(\text{span}(\boldsymbol{b}))$. We consider the two cases where (1) $\boldsymbol{b}$ is an eigenvector for $\boldsymbol{A}$ and (2) $\boldsymbol{b}$ is not an eigenvector for $\boldsymbol{A}$. In the first case, $\boldsymbol{b} \in \boldsymbol{A}^{-1}(\text{span}(\boldsymbol{b}))$ so we clearly have

$$\text{span}(\boldsymbol{b}) \cap \boldsymbol{A}^{-1}(\text{span}(\boldsymbol{b})) = \text{span}(\boldsymbol{b}).$$

In the second case, $\boldsymbol{b} \notin \boldsymbol{A}^{-1}(\text{span}(\boldsymbol{b}))$ so that

$$\text{span}(\boldsymbol{b}) \cap \boldsymbol{A}^{-1}(\text{span}(\boldsymbol{b})) = \{\boldsymbol{0}\}.$$

But this is our result. ■

Now we can use this lemma to describe the set of curves in $\mathbb{R}^n$ that can be followed exactly by our control system.

2.25 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, 0_1)$ be a SISO linear system and let $T(\boldsymbol{A}, \boldsymbol{b}) \subset \mathbb{R}^n$ be the subspace defined above. If $I \subset \mathbb{R}$ is an interval and if $\boldsymbol{r}\colon I \to T(\boldsymbol{A}, \boldsymbol{b})$ is continuously differentiable, then there exists a continuous function $u\colon I \to \mathbb{R}$ with the property that*

$$\dot{\boldsymbol{r}}(t) = \boldsymbol{A}\boldsymbol{r}(t) + \boldsymbol{b}u(t).$$

Furthermore, $T(\boldsymbol{A}, \boldsymbol{b})$ is the largest subspace of $\mathbb{R}^n$ with this property.

Proof For the first part of the proposition, we note that if $\boldsymbol{r}\colon I \to T(\boldsymbol{A}, \boldsymbol{b})$ then

$$\dot{\boldsymbol{r}}(t) = \lim_{\tau \to 0} \frac{\boldsymbol{r}(t+\tau) - \boldsymbol{r}(t)}{\tau} \in T(\boldsymbol{A}, \boldsymbol{b})$$

since $\boldsymbol{r}(t+\tau), \boldsymbol{r}(t) \in T(\boldsymbol{A}, \boldsymbol{b})$. Therefore, by Lemma 2.24,

$$\dot{\boldsymbol{r}}(t) - \boldsymbol{A}\boldsymbol{r}(t) \in T(\boldsymbol{A}, \boldsymbol{b}), \quad t \in I.$$

Therefore, for each $t \in I$ there exists $u(t) \in \mathbb{R}$ so that

$$\dot{\boldsymbol{r}}(t) - \boldsymbol{A}\boldsymbol{r}(t) = u(t)\boldsymbol{b}.$$

The first part of the proposition now follows since the $T(\boldsymbol{A}, \boldsymbol{b})$-valued function of t, $\dot{\boldsymbol{r}}(t) - \boldsymbol{A}\boldsymbol{r}(t)$ is continuous.

Now suppose that V is a subspace of $\mathbb{R}^n$ with the property that for every continuously differentiable $\boldsymbol{r}\colon I \to V$, there exists a continuous function $u\colon I \to \mathbb{R}$ with the property that

$$\dot{\boldsymbol{r}}(t) = \boldsymbol{A}\boldsymbol{r}(t) + \boldsymbol{b}u(t).$$

Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in V$ and define $\boldsymbol{r}\colon \mathbb{R} \to V$ by $-\boldsymbol{x}_1 + t\boldsymbol{x}_2$. Then we have $\boldsymbol{r}(0) = -\boldsymbol{x}_1$ and $\dot{\boldsymbol{r}}(0) = \boldsymbol{x}_2$. Therefore there must exist a continuous $u\colon \mathbb{R} \to \mathbb{R}$ so that

$$\boldsymbol{x}_2 = -\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}u(0).$$

Since this construction can be made for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in V$, we must have $\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{x}_2 \in \operatorname{span}(\boldsymbol{b})$ for every $\boldsymbol{x}_1, \boldsymbol{x}_2 \in V$. By Lemma 2.24 this means that $V = T(\boldsymbol{A}, \boldsymbol{b})$. ■

Thus we see for single-input systems, the state trajectories we may *exactly* follow are actually quite limited. Nevertheless, even though one cannot follow all state trajectories, it is possible for a system to be controllable.

2.26 Remark As was the case for observability in Remark 2.18, it is easy to talk about controllability in the MIMO setting. Indeed, if for a MIMO system $\Sigma = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ we define

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{B}) = \left[\begin{array}{c|c|c|c} \boldsymbol{B} & \boldsymbol{A}\boldsymbol{B} & \cdots & \boldsymbol{A}^{n-1}\boldsymbol{B} \end{array} \right],$$

then the appropriate versions of Theorems 2.21 and 2.23 hold. •

2.3.3 Bad behaviour due to unstable zero dynamics

Now you are doubtless thinking that we must have ourselves covered. Surely if a system is complete then our state-space behaviour will be nice if the output is nice. But this is in fact not true, as the following example shows.

2.27 Example We take as our system

$$\begin{aligned}\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} &= \begin{bmatrix}0 & 1\\ -2 & -3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix}u\\ y &= \begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}.\end{aligned} \tag{2.9}$$

First, let's see that the system is observable and controllable. The respective matrices are

$$\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}) = \begin{bmatrix}1 & -1\\ 2 & 4\end{bmatrix}, \quad \boldsymbol{C}(\boldsymbol{A},\boldsymbol{b}) = \begin{bmatrix}0 & 1\\ 1 & -3\end{bmatrix}$$

which both have rank 2. We compute

$$e^{\boldsymbol{A}t} = \begin{bmatrix}2e^{-t} - e^{-2t} & e^{-t} - e^{-2t}\\ 2(e^{-2t} - e^{-t}) & 2e^{-2t} - e^{-t}\end{bmatrix}.$$

In this example, we do not use a step input, but rather a violent input:

$$u(t) = \begin{cases} e^t, & t \geq 0\\ 0, & \text{otherwise.}\end{cases}$$

Thus our input blows up as time increases. The usual calculations, using zero initial conditions, give

$$\boldsymbol{x}(t) = \begin{bmatrix}\frac{1}{6}e^t + \frac{1}{3}e^{-2t} - \frac{1}{2}e^{-t}\\ \frac{1}{6}e^t - \frac{2}{3}e^{-2t} + \frac{1}{2}e^{-t}\end{bmatrix}, \quad y(t) = e^{-2t} - e^{-t}.$$

Thus the output is behaving nicely (see Figure 2.9) while the state is blowing up to infinity

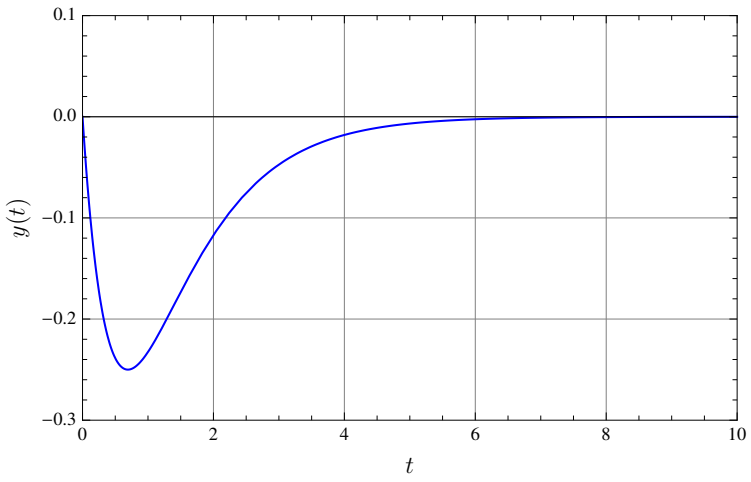


Figure 2.9 The output response of (2.9) to an exponential input

(see Figure 2.10). •

Things are a bit more subtle with this example. The problem is that the large input is not being transmitted to the output. Describing the general scenario here is not altogether easy, but we work through it so that you may know what is going on.

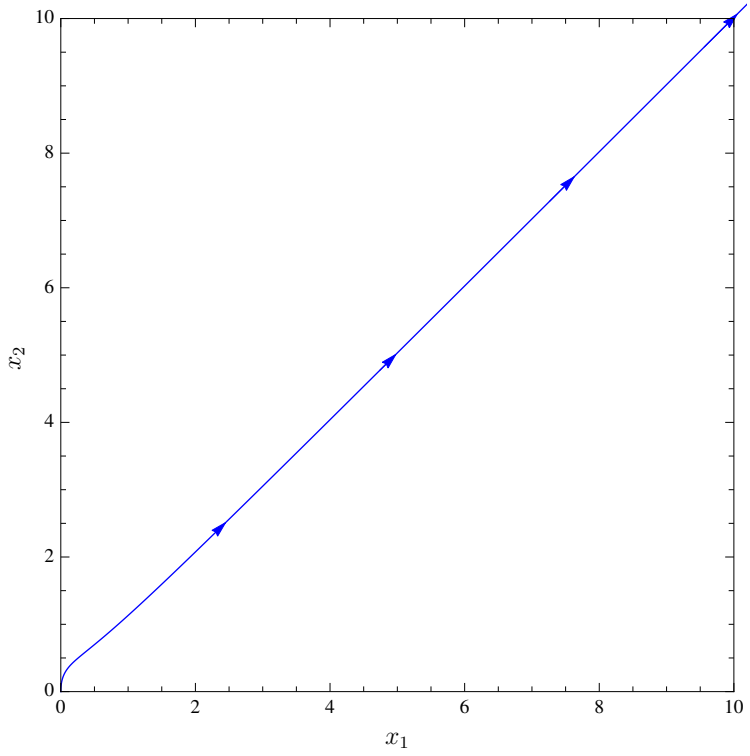


Figure 2.10 The state-space behaviour of (2.9) with an exponential input

2.28 Algorithm for determining zero dynamics We start with a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$. We do not assume that $(\boldsymbol{A}, \boldsymbol{b})$ is controllable or that $(\boldsymbol{A}, \boldsymbol{c})$ is observable.

1. Define $Z_0 = \mathbb{R}^n$.
2. Inductively define a sequence of subspaces of $\mathbb{R}^n$ by
$$Z_{k+1} = \ker(\boldsymbol{c}^t) \cap \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \in Z_k + \operatorname{span}(\boldsymbol{b})\}.$$
3. This sequence will eventually stop, i.e., there will be a least K so that $Z_{K+1} = Z_K$. Denote $Z_\Sigma = Z_K$, and suppose that $\dim(Z_\Sigma) = \ell$.
4. It turns out that is it possible to find $\boldsymbol{f} \in \mathbb{R}^n$ with the property that $\boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}} \triangleq \boldsymbol{A} + \boldsymbol{b}\boldsymbol{f}^t \in \mathbb{R}^{n\times n}$ has Z_Σ as an invariant subspace. Choose such an $\boldsymbol{f}$.
5. Choose a basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_\ell\}$ for Z_Σ, and extend this to a basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ for $\mathbb{R}^n$.
6. Write
$$\begin{aligned}
\boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}}\boldsymbol{v}_1 &= b_{11}\boldsymbol{v}_1 + \cdots + b_{\ell 1}\boldsymbol{v}_\ell \\
&\vdots \\
\boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}}\boldsymbol{v}_\ell &= b_{1\ell}\boldsymbol{v}_1 + \cdots + b_{\ell\ell}\boldsymbol{v}_\ell \\
\boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}}\boldsymbol{v}_{\ell+1} &= b_{1,\ell+1}\boldsymbol{v}_1 + \cdots + b_{\ell,\ell+1}\boldsymbol{v}_\ell + b_{\ell+1,\ell+1}\boldsymbol{v}_{\ell+1} + \cdots + b_{n,\ell+1}\boldsymbol{v}_n \\
&\vdots \\
\boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}}\boldsymbol{v}_n &= b_{1n}\boldsymbol{v}_1 + \cdots + b_{\ell n}\boldsymbol{v}_\ell + b_{\ell+1,n}\boldsymbol{v}_{\ell+1} + \cdots + b_{nn}\boldsymbol{v}_n.
\end{aligned}$$

7. Define an $\ell \times \ell$ matrix by

$$\boldsymbol{N}_\Sigma = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1\ell} \\ b_{21} & b_{22} & \cdots & b_{2\ell} \\ \vdots & \vdots & \ddots & \vdots \\ b_{\ell 1} & b_{\ell 2} & \cdots & b_{\ell\ell} \end{bmatrix}.$$

8. The linear differential equation

$$\dot{\boldsymbol{w}} = \boldsymbol{N}_\Sigma \boldsymbol{w}$$

is called the ***zero dynamics*** for Σ. •

This is plainly nontrivial! Let's illustrate what is going on with our example.

2.29 Example **(Example 2.27 cont'd)** We shall go through the algorithm step by step.

1. We take $V_0 = \mathbb{R}^2$ as directed.
2. As per the instructions, we need to compute $\ker(\boldsymbol{c}^t)$ and we easily see that

$$\ker(\boldsymbol{c}^t) = \operatorname{span}((1,1)).$$

Now we compute

$$\{\boldsymbol{x} \in \mathbb{R}^2 \mid \boldsymbol{A}\boldsymbol{x} \in Z_0 + \operatorname{span}(\boldsymbol{b})\} = \mathbb{R}^2$$

since $Z_0 = \mathbb{R}^2$. Therefore $Z_1 = \ker(\boldsymbol{c}^t)$. To compute Z_2 we compute

$$\{\boldsymbol{x} \in \mathbb{R}^2 \mid \boldsymbol{A}\boldsymbol{x} \in Z_1 + \operatorname{span}(\boldsymbol{b})\} = \mathbb{R}^2$$

since $\ker(\boldsymbol{c}^t)$ and $\operatorname{span}(\boldsymbol{b})$ are complementary subspaces. Therefore $Z_2 = \ker(\boldsymbol{c}^t)$ and so our sequence terminates at Z_1.

3. We have

$$Z_\Sigma = \ker(\boldsymbol{c}^t) = \operatorname{span}((1,1)).$$

4. Let $\boldsymbol{f} = (f_1, f_2)$. We compute

$$\boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} f_1 & f_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 + f_1 & -3 + f_2 \end{bmatrix}.$$

In order that this matrix leave Z_Σ invariant, it must map the basis vector $(1,1)$ for Z_Σ to a multiple of itself. We compute

$$\boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 + f_1 & -3 + f_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ f_1 + f_2 - 5 \end{bmatrix}.$$

In order that this vector be a multiple of $(1,1)$ we must have $f_1 + f_2 - 5 = 1$ or $f_1 + f_2 = 6$. Let us choose $f_1 = f_2 = 3$.

5. We choose the basis $\{\boldsymbol{v}_1 = (1,1), \boldsymbol{v}_2 = (1,-1)\}$ for $\mathbb{R}^2$, noting that $\boldsymbol{v}_1$ is a basis for Z_Σ.
6. We compute

$$\begin{aligned} \boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}}(\boldsymbol{v}_1) &= (1,1) = 1\boldsymbol{v}_1 + 0\boldsymbol{v}_2 \\ \boldsymbol{A}_{\boldsymbol{b},\boldsymbol{f}}(\boldsymbol{v}_2) &= (-1,1) = 0\boldsymbol{v}_1 - 1\boldsymbol{v}_2. \end{aligned}$$

7. The matrix $\boldsymbol{N}_\Sigma$ is 1×1 and is given by
$$\boldsymbol{N}_\Sigma = \begin{bmatrix} 1 \end{bmatrix}.$$
8. The zero dynamics are then
$$\begin{bmatrix} \dot{w}_1 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} w_1 \end{bmatrix}$$
which is a scalar system.

Okay, so how is our bad behaviour reflected here? Well, note that the zero dynamics are unstable! This, it turns out, is the problem. •

2.30 Remarks

1. Systems with *stable* zero dynamics (i.e., all eigenvalues for the matrix $\boldsymbol{N}_\Sigma$ have nonpositive real part) are sometimes called ***minimum phase*** systems. Note that the response Figure 2.9 shows an output that initially does something opposite from what it ends up eventually doing—the output decreases before it finally increases to its final value. This, it turns out, is behaviour typical of a system that is not minimum phase. We shall be investigating the properties of ***nonminimum phase*** systems as we go along (see Theorem 3.15).
2. The zero dynamics as we construct them are not obviously independent of the choices made in the algorithm. That is to say, it is not clear that, had we chosen a different vector $\boldsymbol{f}$, or a different basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$, that we would not arrive at an utterly different matrix $\boldsymbol{N}_\Sigma$. Nonetheless, it is true that if we were to have chosen a different vector $\boldsymbol{f}$ and the same basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ that the resulting matrix $\boldsymbol{N}_\Sigma$ would be unchanged. A choice of a different basis would only change the matrix $\boldsymbol{N}_\Sigma$ by a similarity transformation, and so, in particular, its eigenvalues would be unchanged. •

Let us complete this section by giving a description of the subspace Z_Σ.

2.31 Theorem *Let $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO system and let $\mathscr{Z}$ be the set of all subspaces V of $\mathbb{R}^n$ with the properties*

(i) $V \subset \ker(\boldsymbol{c}^t)$ *and*

(ii) $\boldsymbol{A}(V) \subset V + \operatorname{span}(\boldsymbol{b})$.

The subspace Z_Σ constructed in Algorithm 2.28 is the largest subspace in $\mathscr{Z}$.

Proof By the inductive procedure of Algorithm 2.28 it is clear that $Z_\Sigma \in \mathscr{Z}$. We then need only show that Z_Σ is the largest subspace in $\mathscr{Z}$. Let $V \in \mathscr{Z}$ and let $\boldsymbol{x} \in V$. This means that $\boldsymbol{x} \in \ker(\boldsymbol{c}^t)$ and so $\boldsymbol{x} \in Z_1$ (since in Algorithm 2.28 we always have $Z_1 = \ker(\boldsymbol{c}^t)$). We also have
$$\boldsymbol{A}\boldsymbol{x} \in V + \operatorname{span}(\boldsymbol{b}) \subset Z_1 + \operatorname{span}(\boldsymbol{b}).$$
Therefore $\boldsymbol{x} \in Z_2$. Proceeding in this way we see that $\boldsymbol{x} \in Z_i$ for $i = 1, \dots, K$, and so $\boldsymbol{x} \in Z_\Sigma$. This concludes the proof. ■

2.3.4 A summary of what we have said in this section

We have covered a lot of ground here with a few simple examples, and some general definitions. The material in this section has touched upon some fairly deep concepts in linear control theory, and a recap is probably a good idea. Let us outline the three things we have found that can go wrong, and just how they go wrong.

1. *Unseen unstable dynamics due to lack of observability:* This was illustrated in Example 2.13. The idea was that any input we gave the system leads to a nice output. However, some inputs cause the states to blow up. The problem here is that lack of observability causes the output to not recognise the nasty state-space behaviour.
2. *Lurking unstable dynamics caused by lack of controllability:* It is possible, as we saw in Example 2.19, for the dynamics to be unstable, even though they are fine for some initial conditions. And these unstable dynamics are not something we can get a handle on with our inputs; this being the case because of the lack of controllability.
3. *Very large inputs can cause no output due to the existence of unstable zero dynamics:* This is the situation illustrated in Example 2.27. The problem here is that all the input energy can be soaked by the unstable modes of the zero dynamics, provided the input is of the right type.

It is important to note that if we have any of the badness of the type listed above, there ain't nothing we can do about it. It is a limitation of the physical system, and so one has to be aware of it, and cope as best one can.

We shall see these ideas arise in various ways when we discuss transfer functions in Chapter 3. As we say, the connection here is a little deep, so if you really want to see what is going on here, be prepared to invest some effort—it is really a very neat story, however.

2.4 The impulse response

In this section we will only consider SISO systems, and we will suppose that the 1×1 *matrix* $\boldsymbol{D}$ *is zero. Generalisations to cases where the first condition does not hold are straightforward. Generalisation to the case where* $\boldsymbol{D}$ *is non-zero is essentially carried out in Exercise E3.1.*

2.4.1 The impulse response for causal systems

Typically, we will use the impulse response in situations where we are interested in positive times. Thus we consider everything before $t = 0$ to be zero, and then at $t = 0$ the action starts. It is this "standard" situation we deal with in this section.

Recall from Theorem 2.6 that the solution to the initial value problem

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

is

$$\boldsymbol{x}(t) = e^{\boldsymbol{A}t}\boldsymbol{x}_0 + \int_0^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}u(\tau)\,\mathrm{d}\tau.$$

Therefore the output $y(t)$ behaves like

$$y(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{x}_0 + \int_0^t \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}u(\tau)\,\mathrm{d}\tau. \tag{2.10}$$

We wish to determine the output when we start with zero initial condition, and at $t_0 = 0$ give the system a sharp "jolt." Let us argue intuitively for a moment. Our input will be zero except for a short time near $t = 0$ where it will be large. One expects, therefore, that the integration over τ in (2.10) will only take place over a very small interval near zero. Outside this interval, u will vanish. With this feeble intuition in mind, we define the ***causal***

impulse response, or simply the ***impulse response***, of (2.1) to be

$$h_\Sigma^+(t) = \begin{cases} \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}, & t \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

More succinctly, we may write $h_\Sigma^+(t) = 1(t)\boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}$, where $1(t)$ is the unit step function. In the next section we will define h_Σ^-. However, since we shall almost always be using h_Σ^+, let us agree to simply write h_Σ for h_Σ^+, resorting to the more precise notation only in those special circumstances where we need to be clear on which version of the impulse response we need. The idea is that the only contribution from u in the integral is at $\tau = 0$. A good question is "Does there exist $u \in \mathscr{U}$ so that the resulting output is $h_\Sigma(t)$ with zero state initial condition?" The answer is, "No there is not." So you will never see the impulse response if you only allow yourself piecewise continuous inputs. In fact, you can allow inputs that are a whole lot more general than piecewise continuous, and you will *still* not ever see the impulse response. However, the impulse response is still an important ingredient in looking at the input/output behaviour of the system. The following trivial result hints at why this is so.

2.32 Proposition *For any $u \in \mathscr{U}$ the output of the system*

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\ \boldsymbol{y}(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) \end{aligned}$$

with the initial condition $\boldsymbol{x} = \boldsymbol{x}_0$ is

$$y(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{x}_0 + \int_0^t h_\Sigma(t-\tau)u(\tau)\,\mathrm{d}\tau.$$

That is to say, from the impulse response one can construct the solution associated with *any* input by performing a ***convolution*** of the input with the impulse response. This despite the fact that no input in $\mathscr{U}$ will ever produce the impulse response itself!

We compute the impulse response for the mass-spring-damper system.

2.33 Examples For this example we have

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{d}{m} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Since the nature of $e^{\boldsymbol{A}t}$ changes character depending on the choice of m, d, and k, let's choose specific numbers to compute the impulse response. In all cases we take $m = 1$. We also have the two cases of output to consider (we do not in this section consider the case when $\boldsymbol{D} \neq \boldsymbol{0}_1$).

1. We first take $d = 3$ and $k = 2$. The matrix exponential is then

$$e^{\boldsymbol{A}t} = \begin{bmatrix} 2e^{-t}2 - e^{-2t} & e^{-t} - e^{-2t} \\ 2(e^{-2t} - e^{-t}) & 2e^{-2t} - e^{-t} \end{bmatrix}.$$

   (a) We first consider the case when $\boldsymbol{c} = (1, 0)$, i.e., when the output is displacement. The impulse response is then

$$h_\Sigma(t) = 1(t)(e^{-t} - e^{-2t})$$

   which we show in the left plot in Figure 2.11.

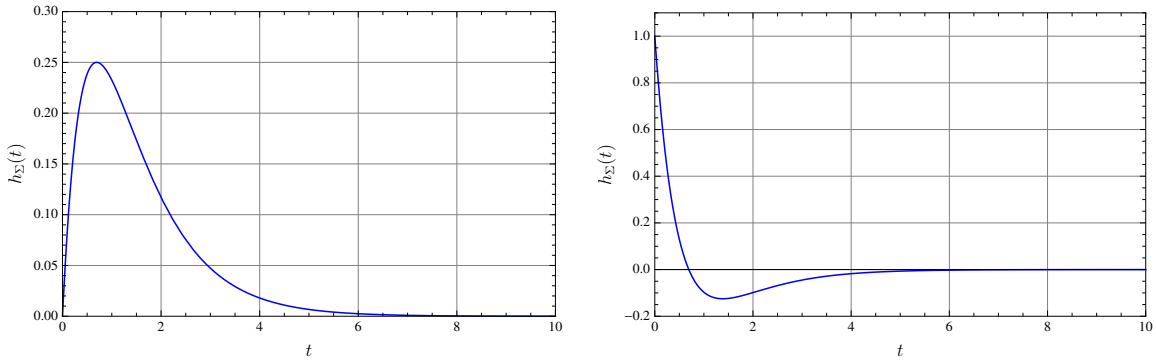


Figure 2.11 The displacement and velocity impulse response for a mass-spring-damper system with $m = 1$, $d = 3$, and $k = 2$

(b) We next consider the case when $\boldsymbol{c} = (0, 1)$ so that the output is velocity. The impulse response is then

$$h_\Sigma(t) = 1(t)(2e^{-2t} - e^{-t}),$$

which we show in the right plot in Figure 2.11. This is the "overdamped case" when there are distinct real eigenvalues.

2. Next we take $d = 2$ and $k = 1$. We compute

$$e^{\boldsymbol{A}t} = \begin{bmatrix} e^{-t}(1+t) & te^{-t} \\ -te^{-t} & e^{-t}(1-t) \end{bmatrix}.$$

(a) Taking $\boldsymbol{c} = (1, 0)$ we compute

$$h_\Sigma(t) = 1(t)(te^{-t})$$

which is the left plot in Figure 2.12.

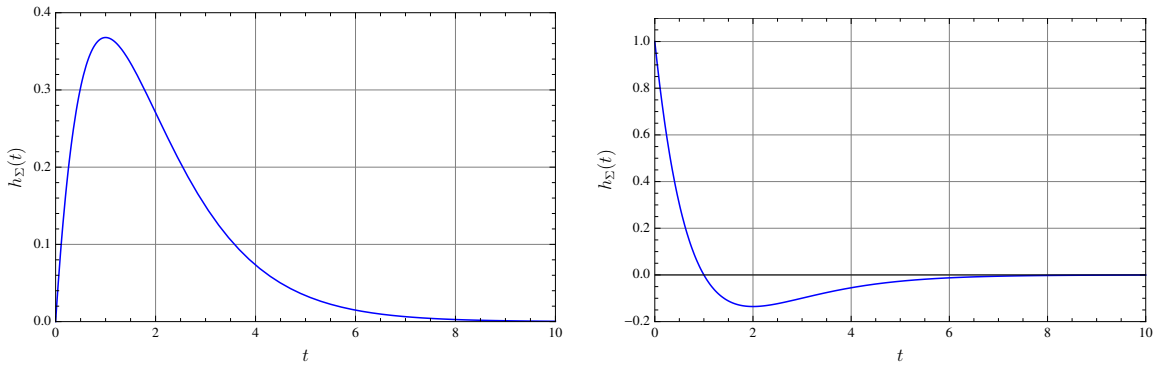


Figure 2.12 The displacement and velocity impulse response for a mass-spring-damper system with $m = 1$, $d = 2$, and $k = 1$

(b) If we let $\boldsymbol{c} = (0, 1)$ then we compute

$$h_\Sigma(t) = 1(t)(e^{-t}(1-t))$$

which is the right plot in Figure 2.12. This is the "critically damped" case when the eigenvalue is repeated.

3. The next case we look at is the "underdamped" one when we have complex roots with negative real part. We take $d = 2$ and $k = 10$ and compute

$$e^{\boldsymbol{A}t} = \begin{bmatrix} e^{-t}(\cos 3t + \frac{1}{3}\sin 3t) & \frac{1}{3}e^{-t}\sin 3t \\ -\frac{10}{3}e^{-t}\sin 3t & e^{-t}(\cos 3t - \frac{1}{3}\sin 3t) \end{bmatrix}.$$

(a) Taking $\boldsymbol{c} = (1, 0)$ we compute

$$h_\Sigma(t) = 1(t)(\tfrac{1}{3}e^{-t}\sin 3t)$$

which is the left plot in Figure 2.13.

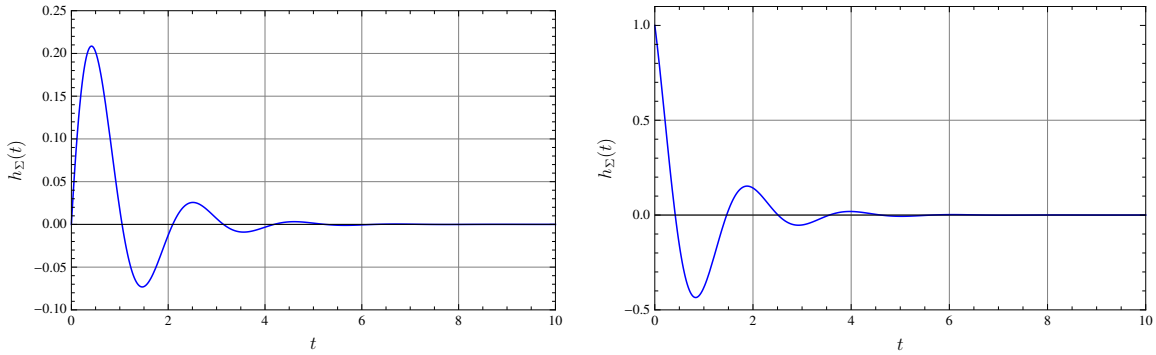


Figure 2.13 The displacement and velocity impulse response for a mass-spring-damper system with $m = 1$, $d = 2$, and $k = 10$

(b) If we let $\boldsymbol{c} = (0, 1)$ then we compute

$$h_\Sigma(t) = 1(t)\big(e^{-t}(\cos 3t - \tfrac{1}{3}\sin 3t)\big)$$

which is the right plot in Figure 2.13.

4. The final case we take is that when there is no damping: $d = 0$ and $k = 1$. Then we have

$$e^{\boldsymbol{A}t} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

(a) Taking $\boldsymbol{c} = (1, 0)$ we compute

$$h_\Sigma(t) = 1(t)\sin t$$

which is the left plot in Figure 2.14.

(b) If we let $\boldsymbol{c} = (0, 1)$ then we compute

$$h_\Sigma(t) = 1(t)\cos t$$

which is the right plot in Figure 2.14. •

Let us see if we can give some justification to the formula for the impulse response. For $\epsilon > 0$ define $u_\epsilon \in \mathscr{U}$ by

$$u_\epsilon(t) = \begin{cases} \frac{1}{\epsilon}, & t \in [0, \epsilon] \\ 0, & \text{otherwise.} \end{cases}$$

The behaviour of these inputs as ϵ shrinks is shown in Figure 2.15. It turns out that these inputs in the limit $\epsilon \to 0$ give the impulse response. Note, however, that in the limit we do not get an input in $\mathscr{U}$!

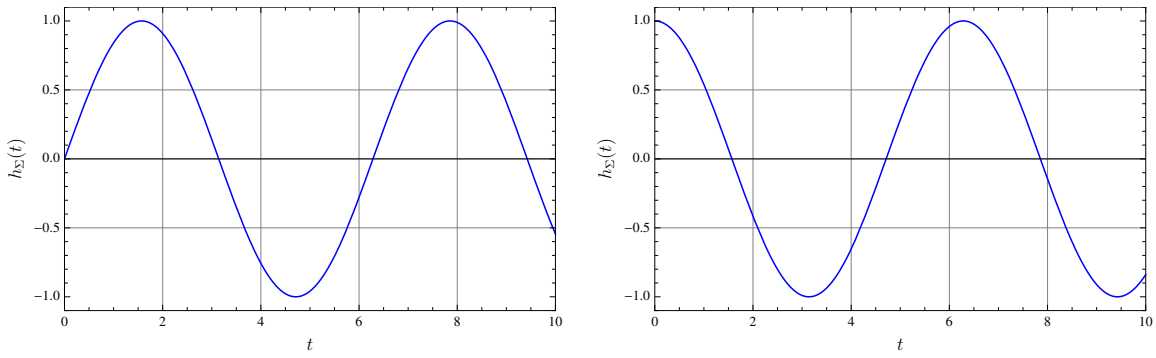


Figure 2.14 The displacement and velocity impulse response for a mass-spring-damper system with $m = 1$, $d = 0$, and $k = 1$

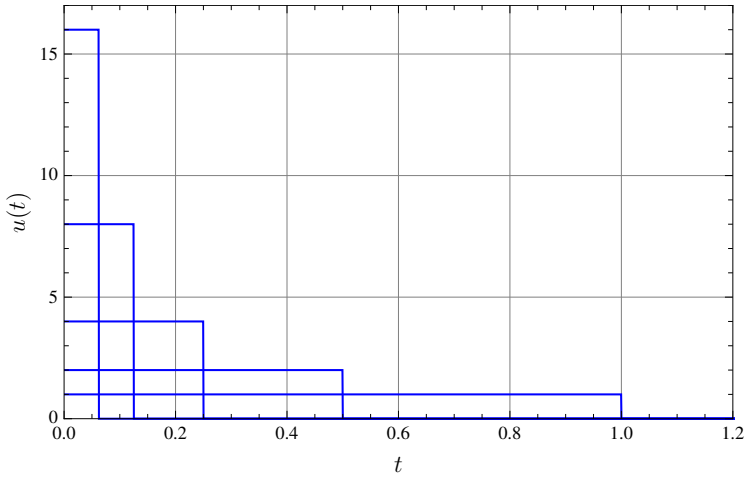


Figure 2.15 A sequence of inputs giving the impulse response in the limit

2.34 Theorem *If*

$$y_\epsilon(t) = \int_0^t \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)} \boldsymbol{b} u_\epsilon(\tau)\,\mathrm{d}\tau$$

then

$$\lim_{\epsilon \to 0} y_\epsilon(t) = h_\Sigma(t).$$

Proof We use the definition of the matrix exponential:

$$\begin{aligned} y_\epsilon(t) &= \int_0^t \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)} \boldsymbol{b} u_\epsilon(\tau)\,\mathrm{d}\tau \\ &= \frac{1}{\epsilon} \int_0^\epsilon \boldsymbol{c}^t e^{\boldsymbol{A}t} \Big(\boldsymbol{I}_n - \boldsymbol{A}\tau + \frac{\boldsymbol{A}^2\tau^2}{2!} + \dots \Big) \boldsymbol{b}\,\mathrm{d}\tau. \end{aligned}$$

Since the sum for the matrix exponential converges uniformly and absolutely on $[0, \epsilon]$ we

may distribute the integral over the sum:

$$\begin{aligned} y_\epsilon(t) &= \frac{1}{\epsilon}\boldsymbol{c}^t e^{\boldsymbol{A}t}\Big(\boldsymbol{I}_n\epsilon - \frac{\boldsymbol{A}\epsilon^2}{2!} + \frac{\boldsymbol{A}^2\epsilon^3}{3!} + \dots\Big)\boldsymbol{b} \\ &= \boldsymbol{c}^t e^{\boldsymbol{A}t}\Big(\boldsymbol{I}_n - \frac{\boldsymbol{A}\epsilon}{2!} + \frac{\boldsymbol{A}^2\epsilon^2}{3!} + \dots\Big)\boldsymbol{b}. \end{aligned}$$

Clearly the result holds when we take the limit $\epsilon \to 0$. ■

2.4.2 The impulse response for anticausal systems

Now we turn our attention to a situation that we will only have need to resort to in Section 15.3; the situation is one where we deal with functions of time that end at $t = 0$. Thus functions are defined on the interval $(-\infty, 0]$. The definition of the impulse response in these cases has the same motivation as in the causal case. We shall use Theorem 2.34 for our motivation. For $\epsilon > 0$, let us define

$$u_\epsilon(t) = \begin{cases} \frac{1}{\epsilon}, & t \in [-\epsilon, 0] \\ 0, & \text{otherwise.} \end{cases}$$

We then define

$$y_\epsilon(t) = \int_t^0 \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}u_\epsilon(\tau)\,\mathrm{d}\tau, \quad t \le 0,$$

and then $h_\Sigma^- = \lim_{\epsilon\to 0} y_\epsilon$. Let us determine the expression for h_Σ^- be performing the computations carried out in the proof of Theorem 2.34, but now for $t \le 0$:

$$\begin{aligned} y_\epsilon(t) &= \int_t^0 \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}u_\epsilon(\tau)\,\mathrm{d}\tau \\ &= \frac{1}{\epsilon}\int_{-\epsilon}^0 \boldsymbol{c}^t e^{\boldsymbol{A}t}\Big(\boldsymbol{I}_n - \boldsymbol{A}\tau + \frac{\boldsymbol{A}^2\tau^2}{2!} + \cdots\Big)\boldsymbol{b}\,\mathrm{d}\tau \\ &= \frac{1}{\epsilon}\boldsymbol{c}^t e^{\boldsymbol{A}t}\Big(-\boldsymbol{I}_n\epsilon + \frac{\boldsymbol{A}\epsilon^2}{2!} - \frac{\boldsymbol{A}^2\epsilon^3}{3!} + \cdots\Big)\boldsymbol{b} \\ &= -\boldsymbol{c}^t e^{\boldsymbol{A}t}\Big(\boldsymbol{I}_n - \frac{\boldsymbol{A}\epsilon}{2!} + \frac{\boldsymbol{A}^2\epsilon^2}{3!} + \cdots\Big)\boldsymbol{b}. \end{aligned}$$

Therefore, we conclude that

$$h_\Sigma^-(t) = \begin{cases} -\boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}, & t \le 0 \\ 0, & \text{otherwise.} \end{cases}$$

This may be written as $h_\Sigma^-(t) = -1(-t)\boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}$, which we call the ***anticausal impulse response***.

The anticausal impulse response is useful for solving a final value problem, as the following result states.

2.35 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system. If $u\colon (-\infty, 0] \to \mathbb{R}$ is admissible then the output of the final value problem*

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \\ \boldsymbol{y}(t) &= \boldsymbol{c}^t\boldsymbol{x}(t), \end{aligned}$$

is given by

$$y(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{x}_0 + \int_t^0 h_\Sigma^-(t-\tau)u(\tau)\,\mathrm{d}\tau, \quad t \leq 0.$$

Proof The result will follow if we can show that the solution to the final value problem

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}\boldsymbol{u}(t) \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

is given by

$$\boldsymbol{x}(t) = e^{\boldsymbol{A}t}\boldsymbol{x}_0 - \int_t^0 e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}u(\tau)\,\mathrm{d}\tau, \quad t \leq 0.$$

Clearly the final condition $\boldsymbol{x}(0) = \boldsymbol{x}_0$ is satisfied. With $\boldsymbol{x}(t)$ so defined, we compute

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= \boldsymbol{A}e^{\boldsymbol{A}t}\boldsymbol{x}_0 + \boldsymbol{b}\boldsymbol{u}(t) - \int_t^0 \boldsymbol{A}e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}\boldsymbol{u}(\tau)\,\mathrm{d}\tau \\ &= \boldsymbol{A}e^{\boldsymbol{A}t}\boldsymbol{x}_0 - \boldsymbol{A}\int_t^0 e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}\boldsymbol{u}(\tau)\,\mathrm{d}\tau + \boldsymbol{b}\boldsymbol{u}(t) \\ &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}\boldsymbol{u}(t).\end{aligned}$$

Thus $\boldsymbol{x}(t)$ also satisfies the differential equation. ■

Thus the anticausal impulse response acts for anticausal inputs in much the same way as the causal impulse response acts for causal inputs.

Note again that we shall only rarely require h_Σ^-, so, again, whenever you see h_Σ, it implicitly refers to h_Σ^+.

2.5 Canonical forms for SISO systems

In this section we look at the appearance of a "typical" SISO linear system of the form (2.2). To do so, we shall take an arbitrary system of that form and make a linear change of coordinates. So let us first make sure we understand what is a linear change of coordinates, and how it manifests itself in the multi-input, multi-output system equations (2.1). We take as our state coordinates $\boldsymbol{x}$, and define new state coordinates $\boldsymbol{\xi} = \boldsymbol{T}\boldsymbol{x}$ where $\boldsymbol{T}$ is an invertible $n \times n$ matrix.[2] We can easily derive the equations that govern the behaviour of the state variables $\boldsymbol{\xi}$. The following result holds in the MIMO case.

2.36 Proposition *If* $\boldsymbol{u}(t) \in \mathbb{R}^m$, $\boldsymbol{x}(t) \in \mathbb{R}^n$, *and* $\boldsymbol{y}(t) \in \mathbb{R}^r$ *satisfy*

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t),\end{aligned}$$

and if $\boldsymbol{\xi} = \boldsymbol{T}\boldsymbol{x}$, *then* $\boldsymbol{u}(t) \in \mathbb{R}^m$, $\boldsymbol{\xi}(t) \in \mathbb{R}^n$, *and* $\boldsymbol{y}(t) \in \mathbb{R}^r$ *satisfy*

$$\begin{aligned}\dot{\boldsymbol{\xi}}(t) &= \boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}\boldsymbol{\xi}(t) + \boldsymbol{T}\boldsymbol{B}\boldsymbol{u}(t) \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{T}^{-1}\boldsymbol{\xi}(t) + \boldsymbol{D}\boldsymbol{u}(t).\end{aligned}$$

[2] Often $\boldsymbol{T}$ is arrived at as follows. One has n linearly independent vectors $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$ in $\mathbb{R}^n$ which therefore form a basis. If we assemble into the columns of a matrix $\boldsymbol{T}^{-1}$ the components of the vectors $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$—that is we take

$$\boldsymbol{T}^{-1} = \left[\begin{array}{c|c|c|c} \boldsymbol{f}_1 & \cdots & \boldsymbol{f}_n \end{array}\right]—$$

then $\boldsymbol{\xi} = \boldsymbol{T}\boldsymbol{x}$ are exactly the components of $\boldsymbol{x} \in \mathbb{R}^n$ in the basis $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$.

Proof We compute

$$\dot{\boldsymbol{\xi}}(t) = \boldsymbol{T}\dot{\boldsymbol{x}}(t) = \boldsymbol{T}\boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{T}\boldsymbol{B}\boldsymbol{u}(t) = \boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}\boldsymbol{\xi}(t) + \boldsymbol{T}\boldsymbol{B}\boldsymbol{u}(t),$$

and $\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) = \boldsymbol{C}\boldsymbol{T}^{-1}\boldsymbol{\xi}(t) + \boldsymbol{D}\boldsymbol{u}(t)$. ■

One may consider more general changes of variable where one defines $\boldsymbol{\eta} = Q^{-1}\boldsymbol{y}$ and $\boldsymbol{\mu} = R^{-1}\boldsymbol{u}$, but since our interest is mainly in the SISO case, such transformations simply boil down to scaling of the variables, and so constitute nothing of profound interest.

2.5.1 Controller canonical form

We now revert to the SISO setting, and prove a "normal form" result for controllable SISO linear systems. Recall that a pair $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is controllable if the vectors $\{\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \boldsymbol{A}^2\boldsymbol{b}, \dots, \boldsymbol{A}^{n-1}\boldsymbol{b}\}$ form a basis for $\mathbb{R}^n$.

2.37 Theorem *If $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is controllable then there exists an invertible $n \times n$ matrix $\boldsymbol{T}$ with the property that*

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Proof We begin with some seemingly unrelated polynomial constructions. Let the characteristic polynomial of $\boldsymbol{A}$ be

$$P_{\boldsymbol{A}}(\lambda) = \lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0.$$

Define $n+1$ polynomials in indeterminant λ by

$$P_i(\lambda) = \sum_{k=0}^{n-i} p_{k+i}\lambda^k, \quad i = 0, \dots, n.$$

Note that $P_0 = P_{\boldsymbol{A}}$ and $P_n(\lambda) = 1$ if we declare that $p_n = 1$. These polynomials satisfy the relation

$$\lambda P_i(\lambda) = P_{i-1}(\lambda) - p_{i-1}P_n(\lambda). \tag{2.11}$$

Indeed, we compute

$$\begin{aligned} \lambda P_i(\lambda) &= \sum_{k=0}^{n-i} p_{k+i}\lambda^{k+1} \\ &= \sum_{k=0}^{n-i} p_{k+i}\lambda^{k+1} + p_{i-1} - p_{i-1} \\ &= \sum_{k=-1}^{n-i} p_{k+i}\lambda^{k+1} - p_{i-1}P_n(\lambda) \\ &= \sum_{k'=0}^{n-(i-1)} p_{k'+(i-1)}\lambda^{k'} - p_{i-1}P_n(\lambda) \\ &= P_{i-1}(\lambda) - p_{i-1}P_n(\lambda), \end{aligned}$$

as asserted.

Now we define $n+1$ vectors by

$$\boldsymbol{f}_i = P_i(\boldsymbol{A})\boldsymbol{b}, \quad i = 0, \ldots, n.$$

Note that $P_i(\boldsymbol{A})$ is simply given by

$$P_i(\boldsymbol{A}) = \sum_{k=0}^{n-i} p_{k+i}\boldsymbol{A}^k, \quad i = 0, \ldots, n.$$

By the Cayley-Hamilton Theorem, $\boldsymbol{f}_0 = \boldsymbol{0}$, and we claim that the vectors $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$ are linearly independent. To see this latter assertion, note that since $(\boldsymbol{A}, \boldsymbol{b})$ is controllable the vectors $\{\boldsymbol{g}_1 = \boldsymbol{A}^{n-1}\boldsymbol{b}, \boldsymbol{g}_2 = \boldsymbol{A}^{n-2}\boldsymbol{b}, \ldots, \boldsymbol{g}_n = \boldsymbol{b}\}$ are linearly independent and so form a basis for $\mathbb{R}^n$. We also have

$$\boldsymbol{f}_i = \sum_{j=1}^{n} T_{ji}\boldsymbol{g}_j$$

where $\boldsymbol{T}$ is the $n \times n$ matrix

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ p_{n-1} & 1 & 0 & \cdots & 0 \\ p_{n-2} & p_{n-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_1 & p_2 & p_3 & \cdots & 1 \end{bmatrix}$$

which is clearly invertible. Therefore $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$ are themselves linearly independent and so form a basis.

We define the matrix $\boldsymbol{T}$ by asking that its inverse be given by

$$\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{f}_1 \,|\, \cdots \,|\, \boldsymbol{f}_n \end{bmatrix}$$

so that $\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}$ is simply the representation of the linear map $\boldsymbol{A}$ in the basis $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$. The relation (2.11) gives

$$\boldsymbol{A}\boldsymbol{f}_i = \boldsymbol{f}_{i-1} - p_{i-1}\boldsymbol{f}_n,$$

from which we get the representation of $\boldsymbol{A}$ in the basis $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$ as in the theorem statement. It is trivially true that the coordinates of $\boldsymbol{b}$ in this basis are $(0, 0, \ldots, 0, 1)$ since $\boldsymbol{f}_n = \boldsymbol{b}$. ■

The pair $(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}\boldsymbol{b})$ of the theorem are sometimes called the ***controller canonical form*** for the pair $(\boldsymbol{A}, \boldsymbol{b})$. This is also sometimes known as the ***second Luenberger-Brunovsky canonical form*** for $(\boldsymbol{A}, \boldsymbol{b})$.

What is the import of this theorem? Well, let us suppose that we are handed an $n \times n$ matrix $\boldsymbol{A}$ and an n-vector $\boldsymbol{b}$ in the form of that in the theorem statement:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

What does the system look like, really? Well, define a scalar variable x by $x = x_1$. We then note that if $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t)$ then

$$\begin{aligned}
\dot{x} &= \dot{x}_1 = x_2 \\
\ddot{x} &= \ddot{x}_1 = \dot{x}_2 = x_3 \\
&\vdots \\
x^{(n)} &= \dot{x}_n = -p_0 x_1 - p_1 x_2 - \cdots - p_{n-1} x_n \\
&= -p_0 x - p_1 \dot{x} - \cdots - p_{n-1} x^{(n-1)} + u.
\end{aligned}$$

Thus the vector equation $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t)$ reduces to the scalar equation

$$x^{(n)}(t) + p_{n-1}x^{(n-1)}(t) + \cdots + p_1 x^{(1)}(t) + p_0 x(t) = u(t).$$

Therefore, when we study the controllable SISO linear system

$$\begin{aligned}
\dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\
\boldsymbol{y}(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t),
\end{aligned} \tag{2.12}$$

we can always make a change of coordinates that will render the system an nth order one whose state variable is a scalar. *This is important*. It is also clear that if one conversely starts with a scalar system

$$\begin{aligned}
x^{(n)}(t) + p_{n-1}x^{(n-1)}(t) + \cdots + p_1 x^{(1)}(t) + p_0 x(t) &= bu(t) \\
y(t) = c_{n-1}x^{(n-1)}(t) + c_{n-2}x^{(n-2)}(t) + \cdots + c_1 x^{(1)}(t) + c_0 x(t) + du(t),&
\end{aligned}$$

one may place it in the form of (2.12) where

$$\boldsymbol{A} = \begin{bmatrix}
0 & 1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1 \\
-p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1}
\end{bmatrix},$$

$$\boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 1 \end{bmatrix}$$

by rescaling variables (for example, if $d = 0$ we may choose $\tilde{u}(t) = bu(t)$).

We shall see that for controllable SISO systems, we may move easily from the linear systems formulation, i.e., equation (2.2), to the scalar equation formulation in all settings we examine, and here we have provided the time-domain setting for making this change. We look at alternative canonical forms for controllable pairs $(\boldsymbol{A}, \boldsymbol{b})$ in Exercises E2.31 E2.32, and E2.33.

2.5.2 Observer canonical form

Now let us focus on the situation when the system is observable. The proof here is simpler than for the controllable case, since we use a "duality" between controllability and observability.

2.38 Theorem *If* $(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ *is observable then there exists an invertible* $n \times n$ *matrix* $\boldsymbol{T}$ *so that*

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -p_0 \\ 1 & 0 & 0 & \cdots & 0 & -p_1 \\ 0 & 1 & 0 & \cdots & 0 & -p_2 \\ 0 & 0 & 1 & \cdots & 0 & -p_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -p_{n-2} \\ 0 & 0 & 0 & \cdots & 1 & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Proof We make the simple observation that $(\boldsymbol{A}, \boldsymbol{c})$ is observable if and only if $(\boldsymbol{A}^t, \boldsymbol{c})$ is controllable. This follows from the easily seen fact that $\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{c}) = \boldsymbol{O}(\boldsymbol{A}^t, \boldsymbol{c})^t$. Therefore, by Theorem 2.37 there exists an invertible $n \times n$ matrix $\tilde{\boldsymbol{T}}$ so that

$$\tilde{\boldsymbol{T}}\boldsymbol{A}^t\tilde{\boldsymbol{T}}^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}, \quad \tilde{\boldsymbol{T}}\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Thus

$$\tilde{\boldsymbol{T}}^{-t}\boldsymbol{A}\tilde{\boldsymbol{T}}^{t} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -p_0 \\ 1 & 0 & 0 & \cdots & 0 & -p_1 \\ 0 & 1 & 0 & \cdots & 0 & -p_2 \\ 0 & 0 & 1 & \cdots & 0 & -p_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -p_{n-2} \\ 0 & 0 & 0 & \cdots & 1 & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{c}^t\tilde{\boldsymbol{T}}^{t} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

The result now follows by letting $\boldsymbol{T} = \tilde{\boldsymbol{T}}^{-t}$. ■

The pair $(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}^{-t}\boldsymbol{c})$ in the theorem statement are said to be in ***observer canonical form*** or in ***second Luenberger-Brunovsky canonical form***

Let us look at the value of this canonical form by expression the system equations for a system that has this form. The differential equation $\dot{x}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t)$ reads

$$\begin{aligned} \dot{x}_1 &= -p_0x_n = b_0u \\ \dot{x}_2 &= x_1 - p_1x_n = -p_0x_- p_1x_n = b_1u \\ &\vdots \\ \dot{x}_n &= x_{n-1} - p_{n-1}x_n = b_{n-1}u, \end{aligned}$$

where we write $\boldsymbol{b} = (b_0, b_1, \ldots, b_{n-1})$. Now we differentiate the expression for $\dot{x}_n$ with respect to t $n-1$ times, and use the equations for $\dot{x}_1, \ldots, \dot{x}_n$ to get the equation

$$x_n^{(n)} + p_{n-1}x_n^{(n-1)} + \cdots + p_1x_n^{(1)} + p_0x_n = b_{n-1}u^{(n-1)} + \cdots + b_1u^{(1)} + b_0u.$$

The equation $y = \boldsymbol{c}^t\boldsymbol{x} + \boldsymbol{D}u$ simply reads $y = x_n + \boldsymbol{D}u$. The upshot, therefore, is that for an observable system one can always make a change of coordinates so that the system is effectively described by the equation

$$y^{(n)} + p_{n-1}y^{(n-1)} + \cdots + p_1y^{(1)} + p_0y = b_nu^{(n)} + b_{n-1}u^{(n-1)} + \cdots + b_1u^{(1)} + b_0u,$$

where b_n is defined by $\boldsymbol{D} = [b_n]$. Thus an observable system can be immediately put into the form of a differential equation for the output in terms of the output and its derivatives. This is an essential observation for the discussion of input/output systems that is initiated in Section 3.4. As with controllable pairs, in the exercises (See E2.34, E2.35, and E2.36) we provide alternate canonical forms for observable pairs.

2.5.3 Canonical forms for uncontrollable and/or unobservable systems

As we have seen, for systems that are either controllable or observable, it is possible to find a set of coordinates in which the system looks simple, in some sense. Now let us address the situation when we know that the system is not both controllable and observable.

First we consider the situation when $(\boldsymbol{A}, \boldsymbol{b})$ is not controllable. The following result expresses the "simplest" form such a pair may take.

2.39 Theorem *If $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is not controllable then there exist an invertible matrix $\boldsymbol{T}$ and a positive integer $\ell < n$ with the property that*

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{n-\ell,\ell} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0}_{n-\ell} \end{bmatrix}. \tag{2.13}$$

Furthermore, $\boldsymbol{T}$ may be chosen so that $(\boldsymbol{A}_{11}, \boldsymbol{b}_1) \in \mathbb{R}^{\ell\times\ell} \times \mathbb{R}^\ell$ are in controller canonical form.

Proof Let V be the smallest $\boldsymbol{A}$-invariant subspace containing $\boldsymbol{b}$, and suppose that $\dim(V) = \ell$. Since $(\boldsymbol{A}, \boldsymbol{b})$ is not controllable, by Theorem 2.23, $\ell < n$. Choose a basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ for $\mathbb{R}^n$ with the property that $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_\ell\}$ is a basis for V. We define $\boldsymbol{T}$ so that

$$\boldsymbol{T}^{-1} = \left[\begin{array}{c|c|c|c} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{array}\right].$$

Since V is $\boldsymbol{A}$-invariant and since $\boldsymbol{b} \in V$, the relations in (2.13) must hold. Note that $\boldsymbol{A}_{11}$ is simply the representation in the basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_\ell\}$ of the restriction of $\boldsymbol{A}$ to V, and that $\boldsymbol{b}_1$ is the representation of $\boldsymbol{b} \in V$ in this same basis. Now we look at the final assertion. By the very definition of V, the pair $(\boldsymbol{A}_{11}, \boldsymbol{b}_1)$ is controllable. Therefore, by Theorem 2.37 there exists an $\ell \times \ell$ invertible matrix $\boldsymbol{T}^t$ so that

$$\boldsymbol{T}^t\boldsymbol{A}_{11}\boldsymbol{T}^{-t}, \quad \boldsymbol{T}^t\boldsymbol{b}_1$$

is in controller canonical form. ■

Now let us do the same thing, except now we look at the situation when $(\boldsymbol{A}, \boldsymbol{c})$ is not observable.

2.40 Theorem *If $(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is not observable then there exist an invertible matrix $\boldsymbol{T}$ and a positive integer $k < n$ with the property that*

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0}_{k,n-k} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{c}_1^t & \boldsymbol{0}_{n-k}^t \end{bmatrix}. \tag{2.14}$$

Furthermore, $\boldsymbol{T}$ may be chosen so that $(\boldsymbol{A}_{11}, \boldsymbol{c}_1) \in \mathbb{R}^{k\times k} \times \mathbb{R}^k$ are in observer canonical form.

Proof Since $(\boldsymbol{A}, \boldsymbol{c})$ is not observable, $(\boldsymbol{A}^t, \boldsymbol{c})$ is not controllable. Therefore, by Theorem 2.39, there exists an invertible matrix $\tilde{\boldsymbol{T}}$ so that

$$\tilde{\boldsymbol{T}}\boldsymbol{A}^t\tilde{\boldsymbol{T}}^{-1} = \begin{bmatrix} \tilde{\boldsymbol{A}}_{11} & \tilde{\boldsymbol{A}}_{12} \\ \boldsymbol{0}_{n-k,k} & \tilde{\boldsymbol{A}}_{22} \end{bmatrix}, \quad \tilde{\boldsymbol{T}}\boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{0}_{n-k} \end{bmatrix},$$

with $(\tilde{\boldsymbol{A}}_{11}, \boldsymbol{c}_1)$ in controller canonical form. Therefore,

$$\tilde{\boldsymbol{T}}^{-t}\boldsymbol{A}\tilde{\boldsymbol{T}}^t, \quad \boldsymbol{c}^t\tilde{\boldsymbol{T}}^t$$

will have the form stated in the theorem, and thus the result follows by taking $\boldsymbol{T} = \tilde{\boldsymbol{T}}^{-1}$. ■

Finally, we look at the case where $(\boldsymbol{A}, \boldsymbol{b})$ is not controllable and where $(\boldsymbol{A}, \boldsymbol{c})$ is not observable.

2.41 Theorem *Suppose that $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}', \boldsymbol{D})$ is neither controllable nor observable. Then there exists integers $j, k, \ell > 0$ and an invertible matrix $\boldsymbol{T}$ so that*

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \boldsymbol{A}_{13} & \boldsymbol{A}_{14} \\ \boldsymbol{0}_{k,j} & \boldsymbol{A}_{22} & \boldsymbol{0}_{j,\ell} & \boldsymbol{A}_{24} \\ \boldsymbol{0}_{\ell,j} & \boldsymbol{0}_{\ell,k} & \boldsymbol{A}_{33} & \boldsymbol{A}_{34} \\ \boldsymbol{0}_{m,j} & \boldsymbol{0}_{m,k} & \boldsymbol{0}_{m,\ell} & \boldsymbol{A}_{44} \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \boldsymbol{0}_\ell \\ \boldsymbol{0}_m \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{0}_j & \boldsymbol{c}_2 & \boldsymbol{0}_\ell & \boldsymbol{c}_4 \end{bmatrix},$$

where $m = n - j - k - \ell$, and where the pair

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{k,j} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}$$

is controllable and the pair

$$\begin{bmatrix} \boldsymbol{A}_{22} & \boldsymbol{A}_{24} \\ \boldsymbol{0}_{m,k} & \boldsymbol{A}_{44} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{c}_2 \\ \boldsymbol{c}_4 \end{bmatrix}$$

is observable.

Proof Choose a basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ for $\mathbb{R}^n$ with the property that

1. $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_j\}$ is a basis for $\operatorname{image}(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})) \cap \ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$,
2. $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_j, \boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_{j+k}\}$ is a basis for $\operatorname{image}(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))$, and
3. $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_j, \boldsymbol{v}_{j+k+1}, \dots, \boldsymbol{v}_{j+k+\ell}\}$ is a basis for $\ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$.

Now define $\boldsymbol{T}$ by

$$\boldsymbol{T}^{-1} = \left[\begin{array}{c|c|c|c|c|c|c|c|c} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_j & \boldsymbol{v}_{j+1} & \cdots & \boldsymbol{v}_{j+k} & \boldsymbol{v}_{j+k+1} & \cdots & \boldsymbol{v}_{j+k+\ell} \end{array} \middle| \begin{array}{c|c|c} \boldsymbol{v}_{j+k+\ell+1} & \cdots & \boldsymbol{v}_n \end{array}\right].$$

From the properties of the basis vectors it follows that

$$\boldsymbol{b} \in \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_j, \boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_{j+k})$$

and that

$$\boldsymbol{c} \in \operatorname{span}(\boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_{j+k}, \boldsymbol{v}_{j+k+\ell+1}, \dots, \boldsymbol{v}_n).$$

From these two observations follow the form of $\boldsymbol{T}\boldsymbol{b}$ and $\boldsymbol{c}^t\boldsymbol{T}^{-1}$ in the theorem statement. Furthermore, since $\operatorname{image}(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))$ and $\ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$ are $\boldsymbol{A}$-invariant (Theorems 2.17 and 2.23), it follows that $\operatorname{image}(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})) \cap \ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$ is $\boldsymbol{A}$-invariant and that $\operatorname{image}(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})) + \ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$ is $\boldsymbol{A}$-invariant. From these observations we conclude the following:

1. $\boldsymbol{A}\boldsymbol{v}_i \in \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{j+k})$ for $i \in \{1, \dots, j+k\}$;
2. $\boldsymbol{A}\boldsymbol{v}_i \in \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_j, \boldsymbol{v}_{j+k+1}, \dots, \boldsymbol{v}_{j+k+\ell})$ for $i \in \{1, \dots, j, j+k+1, \dots, j+k+\ell\}$;
3. $\boldsymbol{A}\boldsymbol{v}_i \in \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_j)$ for $i \in \{1, \dots, j\}$;
4. $\boldsymbol{A}\boldsymbol{v}_i \in \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_j, \boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_{j+k}, \boldsymbol{v}_{j+k+1}, \dots, \boldsymbol{v}_{j+k+\ell})$, $i \in \{1, \dots, j+k+\ell\}$.

From these observations follow the form of $\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}$ in the theorem statement.

Now let us show that the pair

$$\tilde{\boldsymbol{A}}_1 = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{k,j} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \tilde{\boldsymbol{b}}_1 = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}$$

is controllable. First, by direct calculation, we have

$$\boldsymbol{C}(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}\boldsymbol{b}) = \begin{bmatrix} \tilde{\boldsymbol{b}}_1 & \tilde{\boldsymbol{A}}_1\tilde{\boldsymbol{b}}_1 & \cdots & \tilde{\boldsymbol{A}}^{n-1}\tilde{\boldsymbol{b}}_1 \\ \boldsymbol{0}_{\ell+m} & \boldsymbol{0}_{\ell+m} & \cdots & \boldsymbol{0}_{\ell+m} \end{bmatrix}.$$

Now, by our choice of basis vectors we also have

$$\operatorname{image}(\boldsymbol{C}(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}\boldsymbol{b})) = \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_j, \boldsymbol{v}_{j+1}, \dots, \boldsymbol{v}_{j+k}).$$

Thus the matrix

$$\begin{bmatrix} \tilde{\boldsymbol{b}}_1 & \tilde{\boldsymbol{A}}_1\tilde{\boldsymbol{b}}_1 & \cdots & \tilde{\boldsymbol{A}}^{n-1}\tilde{\boldsymbol{b}}_1 \end{bmatrix}$$

must have maximal rank. However, by the Cayley-Hamilton Theorem it follows that the matrix

$$\begin{bmatrix} \tilde{\boldsymbol{b}}_1 & \tilde{\boldsymbol{A}}_1\tilde{\boldsymbol{b}}_1 & \cdots & \tilde{\boldsymbol{A}}^{j+k-1}\tilde{\boldsymbol{b}}_1 \end{bmatrix}$$

also has full rank, showing that $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}})$ is controllable.

That the pair

$$\begin{bmatrix} \boldsymbol{A}_{22} & \boldsymbol{A}_{24} \\ \boldsymbol{0}_{m,k} & \boldsymbol{A}_{44} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{c}_2 \\ \boldsymbol{c}_4 \end{bmatrix}$$

is observable follows in the same manner as the previous step, noting that

$$\ker(\boldsymbol{O}(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}^{-t}\boldsymbol{c})) = \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_j, \boldsymbol{v}_{j+k+1}, \dots, \boldsymbol{v}_{j+k+\ell}).$$ ■

If we write the state vector as $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$ in the decomposition given by the theorem then we may roughly say that

1. $\boldsymbol{x}_1$ represents the states that are controllable but not observable,
2. $\boldsymbol{x}_2$ represents the states that are controllable and observable,
3. $\boldsymbol{x}_3$ represents the states that are neither controllable nor observable,
4. $\boldsymbol{x}_4$ represents the states that are observable but not controllable.

2.6 Summary

This is, as we mentioned in the introduction to the chapter, a difficult bit of material. Here's what you should take away with you, and make sure you are clear on before proceeding.

1. You should know *exactly* what we mean when we say "SISO linear system." This terminology will be used constantly in the remainder of the book.

2. You should be able to take a physical system and put it into the form of an SISO linear system if requested to do so. To do this, linearisation may be required.
3. Given a SISO linear system with a specified input $u(t)$, you should know how to determine, both on paper and with the computer, the output $y(t)$ given an initial value $\boldsymbol{x}(0)$ for the state.
4. You should be able to determine whether a SISO linear system is observable or controllable, and know how the lack of observability or controllability affects a system.
5. You should know roughly the import of $\ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$ and of the columnspace of $\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})$.
6. You should know that there is a thing called "zero dynamics," and you should convince yourself that you can work through Algorithm 2.28 to determine this, at least if you had some time to work it out. We will revisit zero dynamics in Section 3.3, and there you will be given an easy way to determine whether the zero dynamics are stable or unstable.
7. You should be able to determine, by hand and with the computer, the impulse response of a SISO linear system. You should also understand that the impulse response is somehow basic in describing the behaviour of the system—this will be amply borne out as we progress through the book.
8. You should know that a controllable pair $(\boldsymbol{A}, \boldsymbol{b})$ has associated to it a canonical form, and you should be able to write down this canonical form given the characteristic polynomial for $\boldsymbol{A}$.

Exercises

The next three exercises are concerned with interconnections of SISO linear systems. We shall be discussing system interconnections briefly at the beginning of Chapter 3, and thoroughly in Chapter 6. Indeed system interconnections are essential to the notion of feedback.

E2.1 Consider two SISO linear systems governed by the differential equations

$$\text{System 1 equations} \begin{cases} \dot{\boldsymbol{x}}_1(t) = \boldsymbol{A}_1\boldsymbol{x}_1(t) + \boldsymbol{b}_1 u_1(t) \\ y_1(t) = \boldsymbol{c}_1^t \boldsymbol{x}_1(t) \end{cases}$$

$$\text{System 2 equations} \begin{cases} \dot{\boldsymbol{x}}_2(t) = \boldsymbol{A}_2\boldsymbol{x}_2(t) + \boldsymbol{b}_2 u_2(t) \\ y_2(t) = \boldsymbol{c}_2^t \boldsymbol{x}_2(t), \end{cases}$$

where $\boldsymbol{x}_1 \in \mathbb{R}^{n_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{n_2}$. The input and output signals of System 1, denoted $u_1(t)$ and $y_1(t)$, respectively, are both scalar. The input and output signals of System 2, denoted $u_2(t)$ and $y_2(t)$, respectively, are also both scalar. The matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ and the vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{c}_1$, and $\boldsymbol{c}_2$ are of appropriate dimension.

Since each system is single-input, single-output we may "connect" them as shown in Figure E2.1. The output of System 1 is fed into the input of System 2 and so the

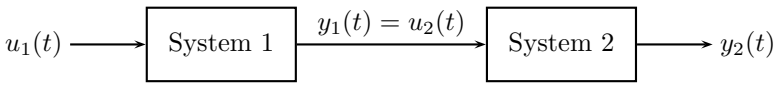


Figure E2.1 SISO linear systems connected in series

interconnected system becomes a single-input, single-output system with input $u_1(t)$ and output $y_2(t)$.

(a) Write the state-space equations for the combined system in the form

$$\begin{bmatrix} \dot{\boldsymbol{x}}_1 \\ \dot{\boldsymbol{x}}_2 \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} + \boldsymbol{b} u_1$$

$$y_2 = \boldsymbol{c}^t \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} + \boldsymbol{D} u_1,$$

where you must determine the expressions for $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$. Note that the combined state vector is in $\mathbb{R}^{n_1+n_2}$.

(b) What is the characteristic polynomial of the interconnected system $\boldsymbol{A}$ matrix? Does the interconnected system share any eigenvalues with either of the two component systems?

E2.2 Consider again two SISO linear systems governed by the differential equations

$$\text{System 1 equations} \begin{cases} \dot{\boldsymbol{x}}_1(t) = \boldsymbol{A}_1\boldsymbol{x}_1(t) + \boldsymbol{b}_1 u_1(t) \\ y_1(t) = \boldsymbol{c}_1^t \boldsymbol{x}_1(t) \end{cases}$$

$$\text{System 2 equations} \begin{cases} \dot{\boldsymbol{x}}_2(t) = \boldsymbol{A}_2\boldsymbol{x}_2(t) + \boldsymbol{b}_2 u_2(t) \\ y_2(t) = \boldsymbol{c}_2^t \boldsymbol{x}_2(t), \end{cases}$$

where $\boldsymbol{x}_1 \in \mathbb{R}^{n_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{n_2}$. The input and output signals of System 1, denoted $u_1(t)$ and $y_1(t)$, respectively, are both scalar. The input and output signals of System 2, denoted $u_2(t)$ and $y_2(t)$, respectively, are also both scalar. The matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ and the vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{c}_1$, and $\boldsymbol{c}_2$ are of appropriate dimension.

Since each system is single-input, single-output we may "connect" them as shown in Figure E2.2. The input to both systems is the same, and their outputs are added

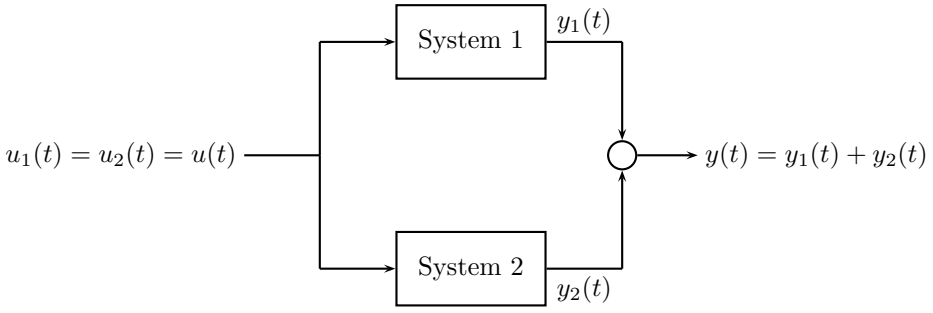


Figure E2.2 SISO linear systems connected in parallel

to get the new output.

(a) Write the state-space equations for the combined system in the form

$$\begin{bmatrix} \dot{\boldsymbol{x}}_1 \\ \dot{\boldsymbol{x}}_2 \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} + \boldsymbol{b}u$$
$$y = \boldsymbol{c}^t \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} + \boldsymbol{D}u,$$

where you must determine the expressions for $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$. Note that the combined state vector is in $\mathbb{R}^{n_1+n_2}$.

(b) What is the characteristic polynomial of the interconnected system $\boldsymbol{A}$ matrix? Does the interconnected system share any eigenvalues with either of the two component systems?

E2.3 Consider yet again two SISO linear systems governed by the differential equations

$$\text{System 1 equations} \begin{cases} \dot{\boldsymbol{x}}_1(t) = \boldsymbol{A}_1\boldsymbol{x}_1(t) + \boldsymbol{b}_1u_1(t) \\ y_1(t) = \boldsymbol{c}_1^t\boldsymbol{x}_1(t) \end{cases}$$
$$\text{System 2 equations} \begin{cases} \dot{\boldsymbol{x}}_2(t) = \boldsymbol{A}_2\boldsymbol{x}_2(t) + \boldsymbol{b}_2u_2(t) \\ y_2(t) = \boldsymbol{c}_2^t\boldsymbol{x}_2(t), \end{cases}$$

where $\boldsymbol{x}_1 \in \mathbb{R}^{n_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{n_2}$. The input and output signals of System 1, denoted $u_1(t)$ and $y_1(t)$, respectively, are both scalar. The input and output signals of System 2, denoted $u_2(t)$ and $y_2(t)$, respectively, are also both scalar. The matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ and the vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{c}_1$, and $\boldsymbol{c}_2$ are of appropriate dimension.

Since each system is single-input, single-output we may "connect" them as shown in Figure E2.3. Thus the input to System 1 is the actual system input u, minus the output from System 2. The input to System 2 is the output from System 1, and the actual system output is the output of System 2.

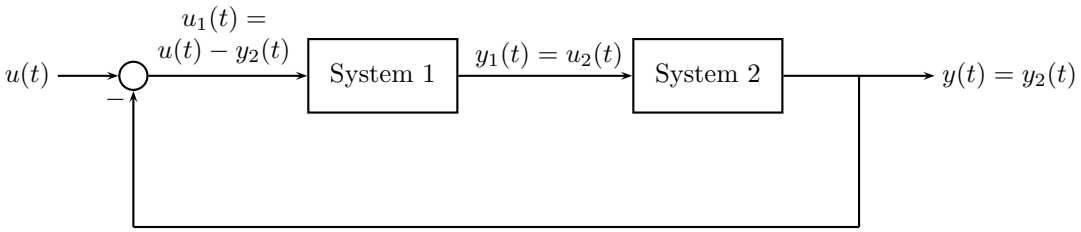


Figure E2.3 SISO linear systems connected in a negative feedback loop

(a) Write the state-space equations for the combined system in the form

$$\begin{bmatrix} \dot{\boldsymbol{x}}_1 \\ \dot{\boldsymbol{x}}_2 \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} + \boldsymbol{b}u$$
$$y = \boldsymbol{c}^t \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} + \boldsymbol{D}u,$$

where you must determine the expressions for $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$. Note that the combined state vector is in $\mathbb{R}^{n_1+n_2}$.

(b) What is the characteristic polynomial of the interconnected system $\boldsymbol{A}$ matrix? Does the interconnected system share any eigenvalues with either of the two component systems?
Hint: *See Exercise E3.7.*

E2.4 Consider the pendulum/cart system of Exercise E1.5. If one adds a force that is applied horizontally to the cart, this leads to a natural input for the system. As output, there are (at least) four natural possibilities: the position of the cart, the velocity of the cart, the pendulum angle, and the pendulum angular velocity. For each of the following eight cases, determine the linearised equations of the form (2.2) for the linearisations:

(a) the equilibrium point $(0, 0)$ with cart position as output;
(b) the equilibrium point $(0, 0)$ with cart velocity as output;
(c) the equilibrium point $(0, 0)$ with pendulum angle as output;
(d) the equilibrium point $(0, 0)$ with pendulum angular velocity as output;
(e) the equilibrium point $(0, \pi)$ with cart position as output;
(f) the equilibrium point $(0, \pi)$ with cart velocity as output;
(g) the equilibrium point $(0, \pi)$ with pendulum angle as output;
(h) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output.

In this problem you first need to determine the nonlinear equations of the form (2.5), and then linearise.

E2.5 Consider the double pendulum of Exercise E1.6. There are at least two ways in which one can provide a single input to the system. The two we consider are

1. a torque at the base of the bottom link relative to the ground (we call this the "pendubot" configuration), and
2. a torque applied to top link from the bottom link (we call this the "acrobot" configuration).

There are various outputs we can consider, but let us choose the angle θ_2 of the "top" pendulum arm.

For each of the following cases, determine the equations in the form (2.2) for the linearisations:

(a) the equilibrium point $(0, 0, 0, 0)$ with the pendubot input;
(b) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;
(c) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;
(d) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;
(e) the equilibrium point $(0, 0, 0, 0)$ with the acrobot input;
(f) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;
(g) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;
(h) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input.

The equilibrium points are written using coordinates $(\theta_1, \theta_2, \dot\theta_1, \dot\theta_2)$. In this problem you first need to determine the nonlinear equations of the form (2.5), and then linearise.

E2.6 Consider the coupled tanks of Exercise E1.11. Take the input u to be F_{in}. Suppose the system is at equilibrium with the height in tank 1 denoted δ_1, and the input flow and height in tank 2 as determined in parts (d) and (e) of Exercise E1.11. Obtain the linearised equations for the system at this equilibrium.

E2.7 Obtain the output $y(t)$ for the SISO linear system

$$\begin{bmatrix}\dot x_1\\ \dot x_2\end{bmatrix} = \begin{bmatrix}\sigma & \omega\\ -\omega & \sigma\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix} u$$
$$y = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix},$$

for $\sigma \in \mathbb{R}$ and $\omega > 0$ when $u(t) = \cos t$ and when $\boldsymbol{x}(0) = \boldsymbol{0}$.

E2.8 Use a computer package to determine the output response of the SISO linear system $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ to a unit step input when

$$\boldsymbol{A} = \begin{bmatrix}-2 & 3 & 1 & 0\\ -3 & -2 & 0 & 1\\ 0 & 0 & -2 & 3\\ 0 & 0 & -3 & -2\end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1.$$

E2.9 (a) Come up with $(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{3\times 3} \times \mathbb{R}^3$ so that $(\boldsymbol{A}, \boldsymbol{c})$ is observable.
(b) Come up with $(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{3\times 3} \times \mathbb{R}^3$ so that $(\boldsymbol{A}, \boldsymbol{c})$ is not observable. Choosing either $\boldsymbol{A}$ or $\boldsymbol{c}$ to be zero is not acceptable.

E2.10 (a) Come up with $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{4\times 4} \times \mathbb{R}^4$ so that $(\boldsymbol{A}, \boldsymbol{b})$ is controllable.
(b) Come up with $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{4\times 4} \times \mathbb{R}^4$ so that $(\boldsymbol{A}, \boldsymbol{b})$ is not controllable. Choosing either $\boldsymbol{A}$ or $\boldsymbol{b}$ to be zero is not acceptable.

E2.11 Define $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ by

$$\boldsymbol{A} = \begin{bmatrix}0 & 1 & 0 & 0 & \cdots & 0\\ 0 & 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & 0 & \cdots & 1\\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1}\end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix}0\\0\\0\\\vdots\\0\\1\end{bmatrix}$$

for $p_0, p_1, \ldots, p_{n-1} \in \mathbb{R}$. Show that $(\boldsymbol{A}, \boldsymbol{b})$ is controllable by verifying that the controllability matrix has full rank.

E2.12 Define $(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -p_0 \\ 1 & 0 & 0 & \cdots & 0 & -p_1 \\ 0 & 1 & 0 & \cdots & 0 & -p_2 \\ 0 & 0 & 1 & \cdots & 0 & -p_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -p_{n-2} \\ 0 & 0 & 0 & \cdots & 1 & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{c}^t = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

for $p_0, p_1, \ldots, p_{n-1} \in \mathbb{R}$. Show that $(\boldsymbol{A}, \boldsymbol{c})$ is observable by verifying that the observability matrix has full rank.

The next two exercises give conditions for controllability and observability called the ***Popov-Belevitch-Hautus*** conditions [see Hautus 1969].

E2.13 Show that $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is controllable if and only if the matrix

$$\left[\begin{array}{c|c} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \end{array}\right]$$

has rank n for all $s \in \mathbb{C}$.

E2.14 Show that $(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is observable if and only if the matrix

$$\begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} \\ \boldsymbol{c}^t \end{bmatrix}$$

has rank n for all $s \in \mathbb{C}$.

E2.15 Show that the definitions of controllability and observability are invariant under linear changes of state variable.

E2.16 Consider the circuit of Exercise E1.7. Take as output the current through the resistor.

(a) Give conditions under which the system is observable.

(b) Give conditions under which the system is controllable.

E2.17 Consider the circuit of Exercise E1.8. Take as output the current through the resistor R_1.

(a) Give conditions under which the system is observable.

(b) Give conditions under which the system is controllable.

E2.18 Consider the circuit of Exercise E1.9. Take as output the current emanating from the voltage source (by the Kirchhoff current law, this is also the sum of the currents through the two resistors).

(a) Give conditions under which the system is observable.

(b) Give conditions under which the system is controllable.

E2.19 For the coupled masses of Exercise E1.4 (assume no damping), suppose you apply a force to the leftmost mass of magnitude $F_1 = u(t)$. You also apply a force to the rightmost mass that is proportional to F_1; thus you take $F_2 = \alpha u(t)$ for some $\alpha \in \mathbb{R}$. The system is still single-input since the two forces are essentially determined by $u(t)$. As an output for the system take the displacement of the rightmost mass.

(a) Determine for which values of α the system is observable. For those cases when the system is *not* observable, can you give a physical interpretation of why it is not by looking at $\ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$?

(b) Determine for which values of α the system is controllable. For those cases when the system is *not* controllable, can you give a physical interpretation of why it is not by looking at $\text{image}(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))$?

E2.20 For the pendulum/cart system of Exercises E1.5 and E2.4, determine whether the linearisations in the following cases are observable and/or controllable:

(a) the equilibrium point $(0, 0)$ with cart position as output;
(b) the equilibrium point $(0, 0)$ with cart velocity as output;
(c) the equilibrium point $(0, 0)$ with pendulum angle as output;
(d) the equilibrium point $(0, 0)$ with pendulum angular velocity as output;
(e) the equilibrium point $(0, \pi)$ with cart position as output;
(f) the equilibrium point $(0, \pi)$ with cart velocity as output;
(g) the equilibrium point $(0, \pi)$ with pendulum angle as output;
(h) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output.

Make sense of your answers by examining $\ker(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))$ and $\text{image}(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$.

E2.21 Consider the the double pendulum of Exercises E1.6 and E2.5. For each of the following cases, determine whether the linearisation is controllable and/or observable:

(a) the equilibrium point $(0, 0, 0, 0)$ with the pendubot input;
(b) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;
(c) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;
(d) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;
(e) the equilibrium point $(0, 0, 0, 0)$ with the acrobot input;
(f) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;
(g) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;
(h) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input.

In each case, the output is the angle of the second link.

E2.22 Consider the coupled tank system of Exercises E1.11 and E2.6. Determine whether the system is controllable and/or observable for the following outputs:

(a) the output is the level in tank 1;
(b) the output is the level in tank 2;
(c) the output is the difference in the levels.

E2.23 Determine the zero dynamics for the pendulum/cart system of Exercises E1.5 and E2.4 for each of the following linearisations:

(a) the equilibrium point $(0, 0)$ with cart position as output;
(b) the equilibrium point $(0, 0)$ with cart velocity as output;
(c) the equilibrium point $(0, 0)$ with pendulum angle as output;
(d) the equilibrium point $(0, 0)$ with pendulum angular velocity as output;
(e) the equilibrium point $(0, \pi)$ with cart position as output;
(f) the equilibrium point $(0, \pi)$ with cart velocity as output;
(g) the equilibrium point $(0, \pi)$ with pendulum angle as output;
(h) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output.

E2.24 For the double pendulum of Exercises E1.6 and E2.5, and for each of the following cases, determine the zero dynamics:

(a) the equilibrium point $(0, 0, 0, 0)$ with the pendubot input;
(b) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;
(c) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;
(d) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;
(e) the equilibrium point $(0, 0, 0, 0)$ with the acrobot input;
(f) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;
(g) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;
(h) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input.

In each case, use the angle of the second link as output.

E2.25 Determine the linearised zero dynamics of for the coupled tank system of Exercises E1.11 and E2.6 for the following outputs:

(a) the height in tank 1;
(b) the height in tank 2;
(c) the difference of the heights in the tanks.

E2.26 Define a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$ with

$$\boldsymbol{A} = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

for $\sigma \in \mathbb{R}$ and $\omega > 0$. Determine the impulse response $h_\Sigma(t)$. Plot the impulse response for various values of σ.

E2.27 Consider the pendulum/cart system of Exercises E1.5 and E2.4, and determine the impulse response of the system for the following linearisations:

(a) the equilibrium point $(0, 0)$ with cart position as output;
(b) the equilibrium point $(0, 0)$ with cart velocity as output;
(c) the equilibrium point $(0, 0)$ with pendulum angle as output;
(d) the equilibrium point $(0, 0)$ with pendulum angular velocity as output;
(e) the equilibrium point $(0, \pi)$ with cart position as output;
(f) the equilibrium point $(0, \pi)$ with cart velocity as output;
(g) the equilibrium point $(0, \pi)$ with pendulum angle as output;
(h) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output.

E2.28 Select values for the parameters of the double pendulum system of Exercises E1.6 and E2.5. For each of the following cases, determine the impulse response for the linearisation:

(a) the equilibrium point $(0, 0, 0, 0)$ with the pendubot input;
(b) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;
(c) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;
(d) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;
(e) the equilibrium point $(0, 0, 0, 0)$ with the acrobot input;
(f) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;
(g) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;
(h) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input.

In each case, use the angle of the second link as output.

E2.29 Determine the linearised impulse response for the coupled tank system of Exercises E1.11 and E2.6 for the following outputs:

(a) the height in tank 1;
(b) the height in tank 2;
(c) the difference of the heights in the tanks.

E2.30 Let

$$\boldsymbol{A} = \begin{bmatrix} -2 & 1 & 3 \\ 0 & -2 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

(a) Verify that $(\boldsymbol{A}, \boldsymbol{b})$ is controllable.
(b) Find the controller canonical form for $(\boldsymbol{A}, \boldsymbol{b})$.

The next few exercises deal with alternative canonical forms for controllable pairs $(\boldsymbol{A}, \boldsymbol{b})$ and for observable pairs $(\boldsymbol{A}, \boldsymbol{c})$.

E2.31 Let $(\boldsymbol{A}, \boldsymbol{b})$ be a controllable pair. Show that the representations of $\boldsymbol{A}$ and $\boldsymbol{b}$ in the basis $\{\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \ldots, \boldsymbol{A}^{n-1}\boldsymbol{b}\}$ are given by

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & -p_0 \\ 1 & 0 & \cdots & 0 & -p_1 \\ 0 & 1 & \cdots & 0 & -p_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -p_{n-2} \\ 0 & 0 & \cdots & 1 & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix},$$

where

$$\boldsymbol{T}^{-1} = \left[\begin{array}{c|c|c|c} \boldsymbol{b} & \boldsymbol{A}\boldsymbol{b} & \cdots & \boldsymbol{A}^{n-1}\boldsymbol{b} \end{array} \right].$$

This is the ***controllability canonical form*** or the ***first Luenberger-Brunovsky canonical form*** for $(\boldsymbol{A}, \boldsymbol{b})$.

E2.32 Again let $(\boldsymbol{A}, \boldsymbol{b})$ be controllable. Recall from Exercise E2.31 that if

$$\boldsymbol{T}_1^{-1} = \left[\begin{array}{c|c|c|c} \boldsymbol{b} & \boldsymbol{A}\boldsymbol{b} & \cdots & \boldsymbol{A}^{n-1}\boldsymbol{b} \end{array} \right],$$

then

$$\boldsymbol{T}_1\boldsymbol{A}\boldsymbol{T}_1^{-1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & -p_0 \\ 1 & 0 & \cdots & 0 & -p_1 \\ 0 & 1 & \cdots & 0 & -p_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{T}_1\boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

Now answer the following questions.

(a) Let

$$\boldsymbol{T}_2 = \begin{bmatrix} 1 & -p_{n-1} & 0 & \cdots & 0 & 0 \\ 0 & 1 & -p_{n-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -p_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Use mathematical induction to show that

$$\boldsymbol{T}_2^{-1} = \begin{bmatrix} 1 & p_{n-1} & p_{n-1}^2 & \cdots & p_{n-1}^{n-2} & p_{n-1}^{n-1} \\ 0 & 1 & p_{n-1} & \cdots & p_{n-1}^{n-3} & p_{n-1}^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & p_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

(b) Define $\boldsymbol{T} = \boldsymbol{T}_2\boldsymbol{T}_1$ and show that

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} -p_{n-1} & -p_{n-2} & \cdots & -p_1 & -p_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

E2.33 Let $(\boldsymbol{A}, \boldsymbol{b})$ be controllable. Find an invertible matrix $\boldsymbol{T}$ so that

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} -p_{n-1} & 1 & 0 & \cdots & 0 \\ -p_{n-2} & 0 & 1 & \cdots & 0 \\ -p_{n-3} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -p_1 & 0 & 0 & \cdots & 1 \\ -p_0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

E2.34 Let $(\boldsymbol{A}, \boldsymbol{c})$ be an observable pair. Show that the representations of $\boldsymbol{A}$ and $\boldsymbol{c}$ in the basis formed from the columns of the matrix

$$\begin{bmatrix} \boldsymbol{c} & \boldsymbol{A}^t\boldsymbol{c} & \cdots & (\boldsymbol{A}^t)^{n-1}\boldsymbol{c} \end{bmatrix}^{-1}$$

are given by

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix},$$

where

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{c} & \boldsymbol{A}^t\boldsymbol{c} & \cdots & (\boldsymbol{A}^t)^{n-1}\boldsymbol{c} \end{bmatrix}.$$

This is the ***observability canonical form*** or the ***first Luenberger-Brunovsky canonical form*** for $(\boldsymbol{A}, \boldsymbol{c})$.

E2.35 Again let $(\boldsymbol{A}, \boldsymbol{c})$ be observable. Recall from Exercise E2.34 that if

$$\boldsymbol{T}_1 = \begin{bmatrix} \boldsymbol{c} & \boldsymbol{A}^t\boldsymbol{c} & \cdots & (\boldsymbol{A}^t)^{n-1}\boldsymbol{c} \end{bmatrix},$$

then

$$\boldsymbol{T}_1\boldsymbol{A}\boldsymbol{T}_1^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}_1^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}.$$

Now answer the following questions.

(a) Let

$$\boldsymbol{T}_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -p_{n-1} & 1 & 0 & \cdots & 0 & 0 \\ 0 & -p_{n-1} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -p_{n-1} & 1 \end{bmatrix}$$

and use mathematical induction to show that

$$\boldsymbol{T}_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ p_{n-1} & 1 & \cdots & 0 & 0 \\ p_{n-1}^2 & p_{n-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{n-2}^{n-2} & p_{n-1}^{n-3} & \cdots & 1 & 0 \\ p_{n-1}^{n-1} & p_{n-1}^{n-2} & \cdots & p_{n-1} & 1 \end{bmatrix}.$$

(b) Define $\boldsymbol{T} = \boldsymbol{T}_2\boldsymbol{T}_1$ and show that

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} -p_{n-1} & 1 & 0 & \cdots & 0 \\ -p_{n-2} & 0 & 1 & \cdots & 0 \\ -p_{n-3} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -p_1 & 0 & 0 & \cdots & 1 \\ -p_0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

E2.36 Let $(\boldsymbol{A}, \boldsymbol{c})$ be observable. Find an invertible matrix $\boldsymbol{T}$ so that

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} -p_{n-1} & -p_{n-2} & \cdots & -p_1 & -p_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

E2.37 Consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ of Example 2.19:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Does there exist an invertible matrix $\boldsymbol{P}$ for which

$$\boldsymbol{TAT}^{-1} = \begin{bmatrix} 0 & 1 \\ -p_0 & -p_1 \end{bmatrix}, \quad \boldsymbol{Tb} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

for some $p_0, p_1 \in \mathbb{R}$?

E2.38 Let $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ be controllable. Show that there exists a *unique* invertible $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ for which $(\boldsymbol{TAT}^{-1}, \boldsymbol{Tb})$ is in controller canonical form.

For a given matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ it may not be possible to find a vector $\boldsymbol{b} \in \mathbb{R}^n$ so that $(\boldsymbol{A}, \boldsymbol{b})$ is controllable. This is related to the Jordan canonical form for $\boldsymbol{A}$, and in the next two exercises you are asked to look into this a little.

E2.39 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system where $\boldsymbol{A}$ has repeated eigenvalues and is diagonalisable.

(a) Is $(\boldsymbol{A}, \boldsymbol{b})$ controllable?

(b) Is $(\boldsymbol{A}, \boldsymbol{c})$ observable?

(c) When is the matrix

$$\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}$$

diagonalisable?

E2.40 Define

$$\boldsymbol{A}_1 = \begin{bmatrix} \sigma & \omega & 0 & 0 \\ -\omega & \sigma & 0 & 0 \\ 0 & 0 & \sigma & \omega \\ 0 & 0 & -\omega & \sigma \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} \sigma & \omega & 1 & 0 \\ -\omega & \sigma & 0 & 1 \\ 0 & 0 & \sigma & \omega \\ 0 & 0 & -\omega & \sigma \end{bmatrix}$$

(a) Show that there is no vector $\boldsymbol{b} \in \mathbb{R}^4$ so that $(\boldsymbol{A}_1, \boldsymbol{b})$ is controllable.

(b) Let $V \subset \mathbb{R}^4$ be the subspace spanned by $\{(1,0,0,0), (0,1,0,0)\}$. Show that $(\boldsymbol{A}_2, \boldsymbol{b})$ is controllable if and only if $\boldsymbol{b} \notin V$.

Chapter 3

Transfer functions (the s-domain)

Although in the previous chapter we occasionally dealt with MIMO systems, from now on we will deal exclusively with SISO systems unless otherwise stated. Certain aspects of what we say can be generalised to the MIMO setting, but it is not our intention to do this here.

Much of what we do in this book revolves around looking at things in the "s-domain," i.e., the complex plane. This domain comes about via the use of the Laplace transform. It is assumed that the reader is familiar with the Laplace transform, but we review some pertinent aspects in Section E.3. We are a little more careful with how we use the Laplace transform than seems to be the norm. This necessitates the use of some Laplace transform terminology that may not be familiar to all students. This may make more desirable than usual a review of the material in Section E.3. In the s-domain, the things we are interested in appear as quotients of polynomials in s, and so in Appendix C we provide a review of some polynomial things you have likely seen before, but perhaps not as systematically as we shall require. The "transfer function" that we introduce in this chapter will be an essential tool in what we do subsequently.

Contents

3.1 Block diagram algebra

We have informally drawn some block diagrams, and it is pretty plain how to handle them. However, let us make sure we are clear on how to do things with block diagrams. In Section 6.1 we will be looking at a more systematic way of handling system with interconnected blocks of rational functions, but our discussion here will serve for what we need immediately, and actually serves for a great deal of what we need to do.

The blocks in a diagram will contain rational functions with indeterminate being the Laplace transform variable s. Thus when you see a block like the one in Figure 3.1 where

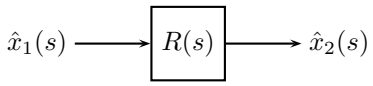


Figure 3.1 The basic element in a block diagram

$R \in \mathbb{R}(s)$ is given by

$$R(s) = \frac{p_n s^n + p_{n-1} s^{n-1} + \cdots + p_1 s + p_0}{q_k s^k + q_{k-1} s^{k-1} + \cdots + q_1 s + q_0}$$

means that $\hat{x}_2(s) = R(s)\hat{x}_1(s)$ or that x_1 and x_2 are related in the time-domain by

$$p_n x_1^{(n)} + p_{n-1} x_1^{(n-1)} + \cdots + p_1 x_1^{(1)} + p_0 x_1 = q_k x_2^{(k)} + q_{k-1} x_2^{(k-1)} + \cdots + q_1 x_2^{(1)} + q_0 x_2$$

(ignoring initial conditions). We shall shortly see just why this should form the basic element for the block diagrams we construct.

Now let us see how to assemble blocks and obtain relations. First let's look at two blocks in ***series*** as in Figure 3.2. If one wanted, one could introduce a variable $\hat{\tilde{x}}$ that represents

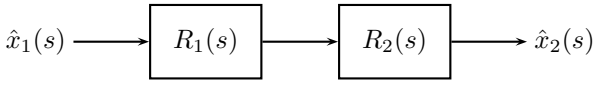


Figure 3.2 Blocks in series

the signal between the blocks and then one has

$$\begin{aligned} & \hat{\tilde{x}}(s) = R_1(s)\hat{x}_1(s), \quad \hat{x}_2(x) = R_2(s)\hat{\tilde{x}}(s) \\ \Longrightarrow \quad & \hat{x}_2(s) = R_1(s)R_2(s)\hat{x}_1(s). \end{aligned}$$

Since multiplication of rational functions is commutative, it does not matter whether we write $R_1(s)R_2(s)$ or $R_2(s)R_1(s)$.

We can also assemble blocks in ***parallel*** as in Figure 3.3. If one introduces temporary signals $\hat{\tilde{x}}_1$ and $\hat{\tilde{x}}_2$ for what comes out of the upper and lower block respectively, then we have

$$\hat{\tilde{x}}_1(s) = R_1(s)\hat{x}_1(s), \quad \hat{\tilde{x}}_2(s) = R_2(s)\hat{x}_1(s).$$

Notice that when we just split a signal like we did before piping $\hat{x}_1$ into both R_1 and R_2, the signal does not change. The temporary signals $\hat{\tilde{x}}_1$ and $\hat{\tilde{x}}_2$ go into the little circle that is a ***summer***. This does what its name implies and sums the signals. That is

$$\hat{x}_2(s) = \hat{\tilde{x}}_1(s) + \hat{\tilde{x}}_2(s).$$

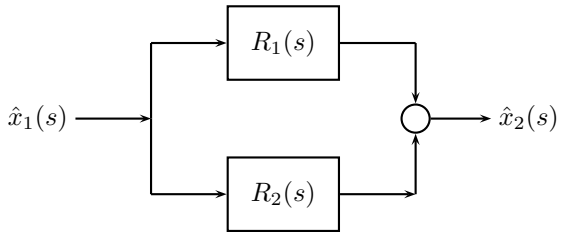


Figure 3.3 Blocks in parallel

Unless it is otherwise depicted, the summer always adds the signals. We'll see shortly what one has to do to the diagram to subtract. We can now solve for $\hat{x}_2$ in terms of $\hat{x}_1$:

$$\hat{x}_2(s) = (R_1(s) + R_2(s))\hat{x}_1(s).$$

The final configuration we examine is the ***negative feedback*** configuration depicted in Figure 3.4. Observe the minus sign attributed to the signal coming out of R_2 into the

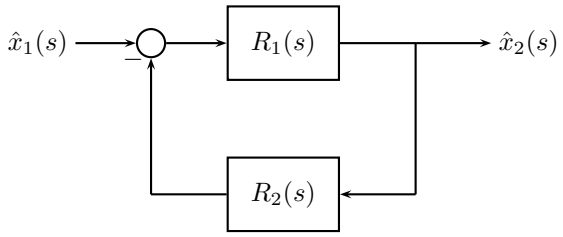


Figure 3.4 Blocks in negative feedback configuration

summer. This means that the signal going into the R_1 block is $\hat{x}_1(s) - R_2(s)\hat{x}_2(s)$. This then gives

$$\begin{aligned} & \hat{x}_2(s) = R_1(s)(\hat{x}_1(s) - R_2(s)\hat{x}_2(s)) \\ \Longrightarrow \quad & \hat{x}_2(s) = \frac{R_1(s)}{1 + R_1(s)R_2(s)}\hat{x}_1(s). \end{aligned}$$

We emphasise that when doing block diagram algebra, one need not get upset when dividing by a rational function unless the rational function is identically zero. That is, don't be thinking to yourself, "But what if this blows up when $s = 3$?" because this is just not something to be concerned about for rational function arithmetic (see Appendix C).

We shall sometimes consider the case where we have ***unity feedback*** (i.e., $R_2(s) = 1$) and to do so, we need to show that the situation in Figure 3.4 can be captured with unity feedback, perhaps with other modifications to the block diagram. Indeed, one can check that the relation between $\hat{x}_2$ and $\hat{x}_1$ is the same for the block diagram of Figure 3.5 as it is for the block diagram of Figure 3.4.

In Section 6.1 we will look at a compact way to represent block diagrams, and one that enables one to prove some general structure results on how to interconnect blocks with rational functions.

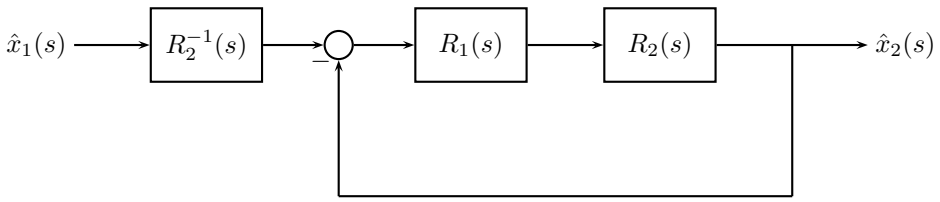


Figure 3.5 A unity feedback equivalent for Figure 3.4

3.2 The transfer function for a SISO linear system

The first thing we do is look at our linear systems formalism of Chapter 2 and see how it appears in the Laplace transform scheme.

We suppose we are given a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, and we fiddle with Laplace transforms a bit for such systems. Note that one takes the Laplace transform of a vector function of time by taking the Laplace transform of each component. Thus we can take the left causal Laplace transform of the linear system

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t) \end{aligned} \tag{3.1}$$

to get

$$\begin{aligned} s\mathscr{L}^+_{0-}(\boldsymbol{x})(s) &= \boldsymbol{A}\mathscr{L}^+_{0-}(\boldsymbol{x})(s) + \boldsymbol{b}\mathscr{L}^+_{0-}(u)(s) \\ \mathscr{L}^+_{0-}(y)(s) &= \boldsymbol{c}^t\mathscr{L}^+_{0-}(\boldsymbol{x})(s) + \boldsymbol{D}\mathscr{L}^+_{0-}(u)(s). \end{aligned}$$

It is convenient to write this in the form

$$\begin{aligned} \mathscr{L}^+_{0-}(\boldsymbol{x})(s) &= (s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}\mathscr{L}^+_{0-}(u)(s) \\ \mathscr{L}^+_{0-}(y)(s) &= \boldsymbol{c}^t\mathscr{L}^+_{0-}(\boldsymbol{x})(s) + \boldsymbol{D}\mathscr{L}^+_{0-}(u)(s). \end{aligned} \tag{3.2}$$

We should be careful how we interpret the inverse of the matrix $s\boldsymbol{I}_n - \boldsymbol{A}$. What one does is think of the entries of the matrix as being polynomials, so the matrix will be invertible provided that its determinant is not the zero polynomial. However, the determinant is simply the characteristic polynomial which is never the zero polynomial. In any case, you should not really think of the entries as being real numbers that you evaluate depending on the value of s. This is best illustrated with an example.

3.1 Example Consider the mass-spring-damper $\boldsymbol{A}$ matrix:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{d}{m} \end{bmatrix} \quad \Longrightarrow \quad s\boldsymbol{I}_2 - \boldsymbol{A} = \begin{bmatrix} s & -1 \\ \frac{k}{m} & s + \frac{d}{m} \end{bmatrix}.$$

To compute $(s\boldsymbol{I}_2 - \boldsymbol{A})^{-1}$ we use the formula (A.2):

$$(s\boldsymbol{I}_2 - \boldsymbol{A})^{-1} = \frac{1}{\det(s\boldsymbol{I}_2 - \boldsymbol{A})}\operatorname{adj}(s\boldsymbol{I}_2 - \boldsymbol{A}),$$

where adj is the adjugate defined in Section A.3.1. We compute

$$\det(s\boldsymbol{I}_2 - \boldsymbol{A}) = s^2 + \tfrac{d}{m}s + \tfrac{k}{m}.$$

This is, of course, the characteristic polynomial for $\boldsymbol{A}$ with indeterminant s! Now we use the cofactor formulas to ascertain

$$\operatorname{adj}(s\boldsymbol{I}_2 - \boldsymbol{A}) = \begin{bmatrix} s + \frac{d}{m} & 1 \\ -\frac{k}{m} & s \end{bmatrix}$$

and so

$$(s\boldsymbol{I}_2 - \boldsymbol{A})^{-1} = \frac{1}{s^2 + \frac{d}{m}s + \frac{k}{m}} \begin{bmatrix} s + \frac{d}{m} & 1 \\ -\frac{k}{m} & s \end{bmatrix}.$$

Note that we do not worry whether $s^2 + \frac{d}{m}s + \frac{k}{m}$ vanishes for certain values of s because we are only thinking of it as a polynomial, and so as long as it is not the zero polynomial, we are okay. And since the characteristic polynomial is *never* the zero polynomial, we are always in fact okay. •

Back to the generalities for the moment. We note that we may, in the Laplace transform domain, solve explicitly for the output $\mathscr{L}^+_{0-}(y)$ in terms of the input $\mathscr{L}^+_{0-}(u)$ to get

$$\mathscr{L}^+_{0-}(y)(s) = \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}\mathscr{L}^+_{0-}(u)(s) + \boldsymbol{D}\mathscr{L}^+_{0-}(u)(s).$$

Note we may write

$$T_\Sigma(s) \triangleq \frac{\mathscr{L}^+_{0-}(y)(s)}{\mathscr{L}^+_{0-}(u)(s)} = \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} + \boldsymbol{D}$$

and we call T_Σ the ***transfer function*** for the linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$. Clearly if we put everything over a common denominator, we have

$$T_\Sigma(s) = \frac{\boldsymbol{c}^t\operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} + \boldsymbol{D}P_{\boldsymbol{A}}(s)}{P_{\boldsymbol{A}}(s)}.$$

It is convenient to think of the relations (3.2) in terms of a block diagram, and we show just such a thing in Figure 3.6. One can see in the figure why the term corresponding to the

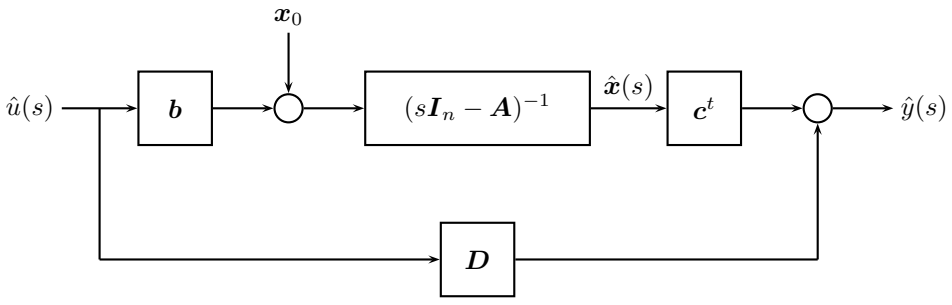


Figure 3.6 The block diagram representation of (3.2)

$\boldsymbol{D}$ matrix is called a ***feedforward*** term, as opposed to a feedback term. We have not yet included feedback, so it does not show up in our block diagram.

Let's see how this transfer function looks for some examples.

3.2 Examples We carry on with our mass-spring-damper example, but now considering the various outputs. Thus we take

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{d}{m} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}.$$

1. The first case is when we have the position of the mass as output. Thus $\boldsymbol{c} = (1, 0)$ and $\boldsymbol{D} = \boldsymbol{0}_1$, and we compute
$$T_\Sigma(s) = \frac{\frac{1}{m}}{s^2 + \frac{d}{m}s + \frac{k}{m}}.$$
2. If we take the velocity of the mass as output, then $\boldsymbol{c} = (0, 1)$ and $\boldsymbol{D} = \boldsymbol{0}_1$ and with this we compute
$$T_\Sigma(s) = \frac{\frac{s}{m}}{s^2 + \frac{d}{m}s + \frac{k}{m}}.$$
3. The final case was acceleration output, and here we had $\boldsymbol{c} = (-\frac{k}{m}, -\frac{d}{m})$ and $\boldsymbol{D} = \boldsymbol{I}_1$. We compute in this case
$$T_\Sigma(s) = \frac{\frac{s^2}{m}}{s^2 + \frac{d}{m}s + \frac{k}{m}}.$$ •

To top off this section, let's give an alternate representation for $\boldsymbol{c}^t\text{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$.

3.3 Lemma $\boldsymbol{c}^t\text{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} = \det \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix}$.

Proof By Lemma A.1 we have

$$\det \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix} = \det(s\boldsymbol{I}_n - \boldsymbol{A}) \det(\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b})$$

$$\implies \quad \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = \frac{\det \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix}}{\det(s\boldsymbol{I}_n - \boldsymbol{A})}$$

Since we also have

$$\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = \frac{\boldsymbol{c}^t\text{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}}{\det(s\boldsymbol{I}_n - \boldsymbol{A})},$$

we may conclude that

$$\boldsymbol{c}^t\text{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} = \det \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix},$$

as desired. ■

3.3 Properties of the transfer function for SISO linear systems

Now that we have constructed the transfer function as a rational function T_Σ, let us look at some properties of this transfer function. For the most part, we will relate these properties to those of linear systems as discussed in Section 2.3. It is interesting that we can infer from the transfer function some of the input/output behaviour we have discussed in the time-domain.

It is important that the transfer function be invariant under linear changes of state variable—we'd like the transfer function to be saying something about the system rather than just the set of coordinates we are using. The following result is an obvious one.

3.4 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system and let $\boldsymbol{T}$ be an invertible $n \times n$ matrix (where $\boldsymbol{A}$ is also in $\mathbb{R}^{n\times n}$). If $\Sigma' = (\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}\boldsymbol{b}, \boldsymbol{c}^t\boldsymbol{T}^{-1}, \boldsymbol{D})$ then $T_{\Sigma'} = T_\Sigma$.*

By Proposition 2.5 this means that if we make a change of coordinate $\boldsymbol{\xi} = \boldsymbol{T}^{-1}\boldsymbol{x}$ for the SISO linear system (3.1), then the transfer function remains unchanged.

3.3.1 Controllability and the transfer function

We will first concern ourselves with cases when the GCD of the numerator and denominator polynomials is not 1.

3.5 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system. If $(\boldsymbol{A}, \boldsymbol{b})$ is controllable, then the polynomials*

$$P_1(s) = \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, \quad P_2(s) = P_{\boldsymbol{A}}(s)$$

are coprime as elements of $\mathbb{R}[s]$ if and only if $(\boldsymbol{A}, \boldsymbol{c})$ is observable.

Proof Although $\boldsymbol{A}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are real, let us for the moment think of them as being complex. This means that we think of $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{C}^n$ and $\boldsymbol{A}$ as being a linear map from $\mathbb{C}^n$ to itself. We also think of $P_1, P_2 \in \mathbb{C}[s]$.

Since $(\boldsymbol{A}, \boldsymbol{b})$ is controllable, by Theorem 2.37 and Proposition 3.4 we may without loss of generality suppose that

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{3.3}$$

Let us first of all determine T_Σ with $\boldsymbol{A}$ and $\boldsymbol{b}$ of this form. Since the first $n-1$ entries of $\boldsymbol{b}$ are zero, we only need the last column of $\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})$. By definition of adj, this means we only need to compute the cofactors for the last row of $s\boldsymbol{I}_n - \boldsymbol{A}$. A tedious but straightforward calculation shows that

$$\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A}) = \begin{bmatrix} * & \cdots & * & 1 \\ * & \cdots & * & s \\ \vdots & \ddots & \vdots & \vdots \\ * & \cdots & * & s^{n-1} \end{bmatrix}.$$

Thus, if $\boldsymbol{c} = (c_0, c_1, \dots, c_{n-1})$ then it is readily seen that

$$\begin{aligned} \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} &= \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{n-1} \end{bmatrix} \\ \implies \quad P_1(s) = \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} &= c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \cdots + c_1 s + c_0. \end{aligned} \tag{3.4}$$

With these preliminaries out of the way, we are ready to proceed with the proof proper.

First suppose that $(\boldsymbol{A}, \boldsymbol{c})$ is not observable. Then there exists a nontrivial subspace $V \subset \mathbb{C}^n$ with the property that $\boldsymbol{A}(V) \subset V$ and $V \subset \ker(\boldsymbol{c}^t)$. Furthermore, we know by

Theorem 2.17 that V is contained in the kernel of $\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})$. Since V is a $\mathbb{C}$-vector space and since $\boldsymbol{A}$ restricts to a linear map on V, there is a nonzero vector $\boldsymbol{z}\in V$ with the property $A\boldsymbol{z}=\alpha\boldsymbol{z}$ for some $\alpha\in\mathbb{C}$. This is a consequence of the fact that the characteristic polynomial of $\boldsymbol{A}$ restricted to V will have a root by the fundamental theorem of algebra. Since $\boldsymbol{z}$ is an eigenvector for $\boldsymbol{A}$ with eigenvalue α, we may use (3.3) to ascertain that the components of $\boldsymbol{z}$ satisfy

$$\begin{aligned} z_2 &= \alpha z_1\\ z_3 &= \alpha z_2 = \alpha^2 z_1\\ &\vdots\\ z_{n-1} &= \alpha z_{n-2} = \alpha^{n-2} z_1\\ -p_0 z_1 - p_1 z_2 - \cdots - p_{n-1} z_n &= \alpha z_n. \end{aligned}$$

The last of these equations then reads

$$\alpha z_n + p_{n-1} z_n + \alpha^{n-2} p_{n-2} z_1 + \cdots + \alpha p_1 z_1 + p_0 z_1 = 0.$$

Using the fact that α is a root of the characteristic polynomial P_2 we arrive at

$$\alpha z_n + p_{n-1} z_n = \alpha^{n-1}(\alpha + p_{n-1}) z_1$$

from which we see that $z_n = \alpha^{n-1} z_1$ provided $\alpha \neq -p_{n-1}$. If $\alpha = -p_{n-1}$ then z_n is left free. Thus the eigenspace for the eigenvalue α is

$$\operatorname{span}((1,\alpha,\ldots,\alpha^{n-1}))$$

if $\alpha \neq -p_{n-1}$ and

$$\operatorname{span}((1,\alpha,\ldots,\alpha^{n-1}),(0,\ldots,0,1))$$

if $\alpha = -p_{n-1}$. In either case, the vector $\boldsymbol{z}_0 \triangleq (1,\alpha,\ldots,\alpha^{n-1})$ is an eigenvector for the eigenvalue α. Thus $\boldsymbol{z}_0 \in V \subset \ker(\boldsymbol{c}^t)$. Thus means that

$$\boldsymbol{c}^t \boldsymbol{z}_0 = c_0 + c_1\alpha + \cdots + c_{n-1}\alpha^{n-1} = 0,$$

and so α is a root of P_1 (by (3.4)) as well as being a root of the characteristic polynomial P_2. Thus P_1 and P_2 are not coprime.

Now suppose that P_1 and P_2 are not coprime. Since these are complex polynomials, this means that there exists $\alpha\in\mathbb{C}$ so that $P_1(s) = (s-\alpha)Q_1(s)$ and $P_2(s) = (s-\alpha)Q_2(s)$ for some $Q_1, Q_2 \in \mathbb{C}[s]$. We claim that the vector $\boldsymbol{z}_0 = (1,\alpha,\ldots,\alpha^{n-1})$ is an eigenvector for $\boldsymbol{A}$. Indeed the components of $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{z}_0$ are

$$\begin{aligned} w_1 &= \alpha\\ w_2 &= \alpha^2\\ &\vdots\\ w_{n-1} &= \alpha^{n-2}\\ w_n &= -p_0 - p_1\alpha - \cdots - p_{n-1}\alpha^{n-1}. \end{aligned}$$

However, since α is a root of P_2 the right-hand side of the last of these equations is simply α^n. This shows that $\boldsymbol{A}\boldsymbol{z}_0 = \boldsymbol{w} = \alpha\boldsymbol{z}_0$ and so $\boldsymbol{z}_0$ is an eigenvector as claimed. Now we claim that $\boldsymbol{z}_0 \in \ker(\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c}))$. Indeed, since α is a root of P_1, by (3.4) we have

$$\boldsymbol{c}^t \boldsymbol{z}_0 = c_0 + c_1 s + \cdots + c_{n-1}\alpha^{n-1} = 0.$$

Therefore, $\boldsymbol{z}_0 \in \ker(\boldsymbol{c}^t)$. Since $\boldsymbol{A}^k\boldsymbol{z}_0 = \alpha^k\boldsymbol{z}_0$ we also have $\boldsymbol{z}_0 \in \ker(\boldsymbol{c}^t\boldsymbol{A}^k)$ for any $k \geq 1$. But this ensures that $\boldsymbol{z}_0 \in \ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$ as claimed. Thus we have found a nonzero vector in $\ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$ which means that $(\boldsymbol{A}, \boldsymbol{c})$ is not observable.

To complete the proof we must now take into account the fact that, in using the Fundamental Theorem of Algebra in some of the arguments above, we have constructed a proof that only works when $\boldsymbol{A}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are thought of as complex. Suppose now that they are real, and first assume that $(\boldsymbol{A}, \boldsymbol{c})$ is not observable. The proof above shows that there is either a one-dimensional real subspace V of $\mathbb{R}^n$ with the property that $A\boldsymbol{v} = \alpha\boldsymbol{v}$ for some nonzero $\boldsymbol{v} \in V$ and some $\alpha \in \mathbb{R}$, or that there exists a two-dimensional real subspace V of $\mathbb{R}^n$ with vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in V$ with the property that

$$\boldsymbol{A}\boldsymbol{v}_1 = \sigma\boldsymbol{v}_1 - \omega\boldsymbol{v}_2, \quad \boldsymbol{A}\boldsymbol{v}_2 = \omega\boldsymbol{v}_1 + \sigma\boldsymbol{v}_2$$

for some $\sigma, \omega \in \mathbb{R}$ with $\omega \neq 0$. In the first case we follow the above proof and see that $\alpha \in \mathbb{R}$ is a root of both P_1 and P_2, and in the second case we see that $\sigma + i\omega$ is a root of both P_1 and P_2. In either case, P_1 and P_2 are not coprime.

Finally, in the real case we suppose that P_1 and P_2 are not coprime. If the root they share is $\alpha \in \mathbb{R}$ then the nonzero vector $(1, \alpha, \ldots, \alpha^{n-1})$ is shown as above to be in $\ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$. If the root they share is $\alpha = \sigma + i\omega$ then the two nonzero vectors $\mathrm{Re}(1, \alpha, \ldots, \alpha^{n-1})$ and $\mathrm{Im}(1, \alpha, \ldots, \alpha^{n-1})$ are shown to be in $\ker(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))$, and so $(\boldsymbol{A}, \boldsymbol{c})$ is not observable. ■

I hope you agree that this is a non-obvious result! That one should be able to infer observability merely by looking at the transfer function is interesting indeed. Let us see that this works in an example.

3.6 Example **(Example 2.13 cont'd)** We consider a slight modification of the example Example 2.13 that, you will recall, was not observable. We take

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & -\epsilon \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

from which we compute

$$\begin{aligned} \boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} &= 1 - s \\ \det(s\boldsymbol{I}_2 - \boldsymbol{A}) &= s^2 - \epsilon s - 1. \end{aligned}$$

Note that when $\epsilon = 0$ we have exactly the situation of Example 2.13. The controllability matrix is

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} 0 & 1 \\ 1 & \epsilon \end{bmatrix}$$

and so the system is controllable. The roots of the characteristic polynomial are

$$s = \frac{-\epsilon \pm \sqrt{4 + \epsilon^2}}{2}$$

and $\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b}$ has the single root $s = 1$. The characteristic polynomial has a root of 1 when and only when $\epsilon = 0$. Therefore, from Theorem 3.5 (which applies since $(\boldsymbol{A}, \boldsymbol{b})$ is controllable) we see that the system is observable if and only if $\epsilon \neq 0$. This can also be seen by computing the observability matrix:

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 1 & -1 \\ -1 & 1 - \epsilon \end{bmatrix}.$$

This matrix has full rank except when $\epsilon = 0$, and this is as it should be. •

Note that Theorem 3.5 holds only when $(\boldsymbol{A}, \boldsymbol{b})$ is controllable. When they are *not* controllable, the situation is somewhat disastrous, as the following result describes.

3.7 Theorem *If $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is not controllable, then for any $\boldsymbol{c} \in \mathbb{R}^n$ the polynomials*

$$P_1(s) = \boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, \quad P_2(s) = P_{\boldsymbol{A}}(s)$$

are not coprime.

Proof By Theorem 2.39 we may suppose that $\boldsymbol{A}$ and $\boldsymbol{b}$ are given by

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{n-\ell,\ell} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0}_{n-\ell} \end{bmatrix},$$

for some $\ell < n$. Therefore,

$$s\boldsymbol{I}_n - \boldsymbol{A} = \begin{bmatrix} s\boldsymbol{I}_\ell \boldsymbol{A}_{11} & -\boldsymbol{A}_{12} \\ \boldsymbol{0}_{n-\ell,\ell} & s\boldsymbol{I}_{n-\ell}\boldsymbol{A}_{22} \end{bmatrix} \implies (s\boldsymbol{I}_n - \boldsymbol{A})^{-1} = \begin{bmatrix} (s\boldsymbol{I}_\ell \boldsymbol{A}_{11})^{-1} & * \\ \boldsymbol{0}_{n-\ell,\ell} & (s\boldsymbol{I}_{n-\ell}\boldsymbol{A}_{22}^{-1} \end{bmatrix},$$

where the $*$ denotes a term that will not matter to us. Thus we have

$$(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = \begin{bmatrix} (s\boldsymbol{I}_\ell \boldsymbol{A}_{11})^{-1}\boldsymbol{b}_1 \\ \boldsymbol{0}_{n-\ell.} \end{bmatrix}$$

This means that if we write $\boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2) \in \mathbb{R}^\ell \times \mathbb{R}^{n-\ell}$ we must have

$$\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = \boldsymbol{c}_1^t(s\boldsymbol{I}_\ell \boldsymbol{A}_{11})^{-1}\boldsymbol{b}_1.$$

This shows that

$$\frac{\boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}}{\det(s\boldsymbol{I}_n - \boldsymbol{A})} = \frac{\boldsymbol{c}_1^t \operatorname{adj}(s\boldsymbol{I}_\ell - \boldsymbol{A}_{11})\boldsymbol{b}_1}{\det(s\boldsymbol{I}_\ell - \boldsymbol{A}_{11})}.$$

The denominator on the left is monic of degree n and the denominator on the right is monic and degree ℓ. This must mean that there is a monic polynomial P of degree $n - \ell$ so that

$$\frac{\boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}}{\det(s\boldsymbol{I}_n - \boldsymbol{A})} = \frac{P(s)\boldsymbol{c}_1^t \operatorname{adj}(s\boldsymbol{I}_\ell - \boldsymbol{A}_{11})\boldsymbol{b}_1}{P(s)\det(s\boldsymbol{I}_\ell - \boldsymbol{A}_{11})},$$

which means that the polynomials $\boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ and $\det(s\boldsymbol{I}_n - \boldsymbol{A})$ are not coprime. ■

This result shows that when $(\boldsymbol{A}, \boldsymbol{b})$ is not controllable, the order of the denominator in T_Σ, after performing pole/zero cancellations, will be strictly less than the state dimension. Thus the transfer function for an uncontrollable system, is *never* representing the complete state information.

Let's see how this works out in our uncontrollable example.

3.8 Example **(Example 2.19 cont'd)** We consider a slight modification of the system in Example 2.19, and consider the system

$$\boldsymbol{A} = \begin{bmatrix} 1 & \epsilon \\ 1 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The controllability matrix is given by

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} 0 & \epsilon \\ 1 & -1 \end{bmatrix},$$

which has full rank except when $\epsilon = 0$. We compute

$$\operatorname{adj}(s\boldsymbol{I}_2 - \boldsymbol{A}) = \begin{bmatrix} s+1 & \epsilon \\ 1 & s-1 \end{bmatrix} \quad \Longrightarrow \quad \operatorname{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} = \begin{bmatrix} \epsilon \\ s-1 \end{bmatrix}.$$

Therefore, for $\boldsymbol{c} = (c_1, c_2)$ we have

$$\boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} = c_2(s-1) + c_1\epsilon.$$

We also have $\det(s\boldsymbol{I}_2 - \boldsymbol{A}) = s^2 - 1 = (s+1)(s-1)$ which means that there will always be a pole/zero cancellation in T_Σ precisely when $\epsilon = 0$. This is precisely when $(\boldsymbol{A}, \boldsymbol{b})$ is not controllable, just as Theorem 3.7 predicts. •

3.3.2 Observability and the transfer function

The above relationship between observability and pole/zero cancellations in the numerator and denominator of T_Σ relies on $(\boldsymbol{A}, \boldsymbol{b})$ being controllable. There is a similar story when $(\boldsymbol{A}, \boldsymbol{c})$ is observable, and this is told by the following theorem.

3.9 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, 0_1)$ be a SISO linear system. If $(\boldsymbol{A}, \boldsymbol{c})$ is observable, then the polynomials*

$$P_1(s) = \boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, \quad P_2(s) = P_{\boldsymbol{A}}(s)$$

are coprime as elements of $\mathbb{R}[s]$ if and only if $(\boldsymbol{A}, \boldsymbol{b})$ is controllable.

Proof First we claim that

$$\begin{aligned} \boldsymbol{b}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A}^t)\boldsymbol{c} &= \boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} \\ \det(s\boldsymbol{I}_n - \boldsymbol{A}^t) &= \det(s\boldsymbol{I}_n - \boldsymbol{A}). \end{aligned} \tag{3.5}$$

Indeed, since the transpose of a 1×1 matrix, i.e., a scalar, is simply the matrix itself, and since matrix inversion and transposition commute, we have

$$\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = \boldsymbol{b}^t(s\boldsymbol{I}_n - s\boldsymbol{A}^t)^{-1}\boldsymbol{c}.$$

This implies, therefore, that

$$\frac{\boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}}{\det(s\boldsymbol{I}_n - \boldsymbol{A})} = \frac{\boldsymbol{b}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A}^t)\boldsymbol{c}}{\det(s\boldsymbol{I}_n - \boldsymbol{A}^t)}.$$

Since the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{A}^t$ agree,

$$\det(s\boldsymbol{I}_n - \boldsymbol{A}^t) = \det(s\boldsymbol{I}_n - \boldsymbol{A}),$$

and from this it follows that

$$\boldsymbol{b}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A}^t)\boldsymbol{c} = \boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}.$$

Now, since $(\boldsymbol{A}, \boldsymbol{c})$ is observable, $(\boldsymbol{A}^t, \boldsymbol{c})$ is controllable (cf. the proof of Theorem 2.38). Therefore, by Theorem 3.5, the polynomials

$$\tilde{P}_1(s) = \boldsymbol{b}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A}^t)\boldsymbol{c}, \quad \tilde{P}_2(s) = P_{\boldsymbol{A}^t}(s)$$

are coprime if and only if $(\boldsymbol{A}^t, \boldsymbol{b})$ is observable. Thus the polynomials $\tilde{P}_1$ and $\tilde{P}_2$ are coprime if and only if $(\boldsymbol{A}, \boldsymbol{b})$ is controllable. However, by (3.5) $P_1 = \tilde{P}_1$ and $P_2 = \tilde{P}_2$, and the result now follows. ■

Let us illustrate this result with an example.

3.10 Example **(Example 2.19 cont'd)** We shall revise slightly Example 2.19 by taking

$$\boldsymbol{A} = \begin{bmatrix} 1 & \epsilon \\ 1 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

We determine that

$$\begin{aligned} \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} &= s - 1 \\ \det(s\boldsymbol{I}_2 - \boldsymbol{A}) &= s^2 - \epsilon - 1. \end{aligned}$$

The observability matrix is computed as

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix},$$

so the system is observable for all ϵ. On the other hand, the controllability matrix is

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} 0 & \epsilon \\ 1 & -1 \end{bmatrix},$$

so the $(\boldsymbol{A}, \boldsymbol{b})$ is controllable if and only if $\epsilon = 0$. What's more, the roots of the characteristic polynomial are $s = \pm\sqrt{1+\epsilon}$. Therefore, the polynomials $\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b}$ and $\det(s\boldsymbol{I}_2 - \boldsymbol{A})$ are coprime if and only if $\epsilon = 0$, just as predicted by Theorem 3.9. •

We also have the following analogue with Theorem 3.7.

3.11 Theorem *If $(\boldsymbol{A}, \boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is not observable, then for any $\boldsymbol{b} \in \mathbb{R}^n$ the polynomials*

$$P_1(s) = \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, \quad P_2(s) = P_{\boldsymbol{A}}(s)$$

are not coprime.

Proof This follows immediately from Theorem 3.7, (3.5) and the fact that $(\boldsymbol{A}, \boldsymbol{c})$ is observable if and only if $(\boldsymbol{A}^t, \boldsymbol{b})$ is controllable. ■

It is, of course, possible to illustrate this in an example, so let us do so.

3.12 Example **(Example 2.13 cont'd)** Here we work with a slight modification of Example 2.13 by taking

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & -\epsilon \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

As the observability matrix is

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 1 & -1 \\ -1 & 1+\epsilon \end{bmatrix},$$

the system is observable if and only if $\epsilon = 0$. If $\boldsymbol{b} = (b_1, b_2)$ then we compute

$$\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} = (b_1 - b_2)(s-1) + \epsilon b_1.$$

We also have $\det(s\boldsymbol{I}_2 - \boldsymbol{A}) = s^2 + \epsilon s - 1$. Thus we see that indeed the polynomials $\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b}$ and $\det(s\boldsymbol{I}_2 - \boldsymbol{A})$ are not coprime for *every* $\boldsymbol{b}$ exactly when $\epsilon = 0$, i.e., exactly when the system is not observable. •

The following corollary summarises the strongest statement one may make concerning the relationship between controllability and observability and pole/zero cancellations in the transfer functions.

3.13 Corollary *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, 0_1)$ be a SISO linear system, and define the polynomials*

$$P_1(s) = \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, \quad P_2(s) = \det(s\boldsymbol{I}_n - \boldsymbol{A}).$$

The following statements are equivalent:

(i) $(\boldsymbol{A}, \boldsymbol{b})$ is controllable and $(\boldsymbol{A}, \boldsymbol{c})$ is observable;

(ii) the polynomials P_1 and P_2 are coprime.

Note that if you are only handed a numerator polynomial and a denominator polynomial that are not coprime, you can only conclude that the system is not *both* controllable and observable. From the polynomials alone, one cannot conclude that the system is, say, controllable but not observable (see Exercise E3.9).

3.3.3 Zero dynamics and the transfer function

It turns out that there is another interesting interpretation of the transfer function as it relates to the zero dynamics. The following result is of general interest, and is also an essential part of the proof of Theorem 3.15. We have already seen that $\det(s\boldsymbol{I}_n - \boldsymbol{A})$ is never the zero polynomial. This result tells us exactly when $\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ is the zero polynomial.

3.14 Lemma *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system, and let Z_Σ be the subspace constructed in Algorithm 2.28. Then $\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ is the zero polynomial if and only if $\boldsymbol{b} \in Z_\Sigma$.*

Proof Suppose that $\boldsymbol{b} \in Z_\Sigma$. By Theorem 2.31 this means that Z_Σ is $\boldsymbol{A}$-invariant, and so also is $(s\boldsymbol{I}_n - \boldsymbol{A})$-invariant. Furthermore, from the expression (A.3) we may ascertain that Z_Σ is $(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}$-invariant, or equivalently, that Z_Σ is $\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})$-invariant. Thus we must have $\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} \in Z_\Sigma$. Since $Z_\Sigma \subset \ker(\boldsymbol{c}^t)$, we must have $\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} = 0$.

Conversely, suppose that $\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} = 0$. By Exercise EE.4 this means that $\boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b} = 0$ for $t \geq 0$. If we Taylor expand $e^{\boldsymbol{A}t}$ about $t = 0$ we get

$$\boldsymbol{c} \sum_{k=0}^{\infty} \frac{t^k}{k!} \boldsymbol{A}^k \boldsymbol{b} = 0.$$

Evaluating the kth derivative of this expression with respect to t at $t = 0$ gives $\boldsymbol{c}\boldsymbol{A}^k\boldsymbol{b} = 0$, $k = 0, 1, 2, \ldots$. Given these relations, we claim that the subspaces Z_k, $k = 1, 2, \ldots$ of Algorithm 2.28 are given by $Z_k = \ker(\boldsymbol{c}^t\boldsymbol{A}^{k-1})$. Since $Z_1 = \ker(\boldsymbol{c}^t)$, the claim holds for $k = 1$. Now suppose that the claim holds for $k = m > 1$. Thus we have $Z_m = \ker(\boldsymbol{c}^t\boldsymbol{A}^{m-1})$. By Algorithm 2.28 we have

$$\begin{aligned}
Z_{m+1} &= \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \in Z_m + \mathrm{span}(\boldsymbol{b})\} \\
&= \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \in \ker(\boldsymbol{c}^t\boldsymbol{A}^{m-1}) + \mathrm{span}(\boldsymbol{b})\} \\
&= \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \in \ker(\boldsymbol{c}^t\boldsymbol{A}^{m-1})\} \\
&= \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \in \ker(\boldsymbol{c}^t\boldsymbol{A}^{m})\} \\
&= \ker(\boldsymbol{c}^t\boldsymbol{A}^{m}),
\end{aligned}$$

where, on the third line, we have used the fact that $\boldsymbol{b} \in \ker(\boldsymbol{c}^t\boldsymbol{A}^{m-1})$. Since our claim follows, and since $\boldsymbol{b} \in \ker(\boldsymbol{c}^t\boldsymbol{A}^k) = Z_{k-1}$ for $k = 0, 1, \ldots$, it follows that $\boldsymbol{b} \in Z_\Sigma$. ■

The lemma, note, gives us conditions on so-called ***invertibility*** of the transfer function. In this case we have invertibility if and only if the transfer function is non-zero.

With this, we may now prove the following.

3.15 Theorem *Consider a SISO control system of the form* $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$. *If* $\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ *is not the zero polynomial then the zeros of* $\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ *are exactly the spectrum for the zero dynamics of* Σ.

Proof Since $\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} \neq 0$, by Lemma 3.14 we have $\boldsymbol{b} \notin Z_\Sigma$. We can therefore choose a basis $\mathcal{B} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ for $\mathbb{R}^n$ with the property that $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_\ell\}$ is a basis for Z_Σ and $\boldsymbol{v}_{\ell+1} = \boldsymbol{b}$. With respect to this basis we can write $\boldsymbol{c} = (\boldsymbol{0}, \boldsymbol{c}_2) \in \mathbb{R}^\ell \times \mathbb{R}^{n-\ell}$ since $Z_\Sigma \subset \ker(\boldsymbol{c}^t)$. We can also write $\boldsymbol{b} = (\boldsymbol{0}, (1, 0, \dots, 0)) \in \mathbb{R}^\ell \times \mathbb{R}^{n-\ell}$, and we denote $\boldsymbol{b}_2 = (1, 0, \dots, 0) \in \mathbb{R}^{n-\ell}$. We write the matrix for the linear map $\boldsymbol{A}$ in this basis as

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}.$$

Since $\boldsymbol{A}(Z_\Sigma) = Z_\Sigma + \mathrm{span}(\boldsymbol{b})$, for $k = 1, \dots \ell$ we must have $\boldsymbol{A}\boldsymbol{v}_k = \boldsymbol{u}_k + \alpha_k \boldsymbol{v}_{\ell+1}$ for some $\boldsymbol{u}_k \in Z_\Sigma$ and for some $\alpha_k \in \mathbb{R}$. This means that $\boldsymbol{A}_{21}$ must have the form

$$\boldsymbol{A}_{21} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_\ell \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{3.6}$$

Therefore $\boldsymbol{f}_1 = (-\alpha_1, \dots, -\alpha_\ell)$ is the unique vector for which $\boldsymbol{b}_2\boldsymbol{f}_1^t = -\boldsymbol{A}_{21}$. We then define $\boldsymbol{f} = (\boldsymbol{f}_1, \boldsymbol{0}) \in \mathbb{R}^\ell \times \mathbb{R}^{n-\ell}$ and determine the matrix for $\boldsymbol{A} + \boldsymbol{b}\boldsymbol{f}^t$ in the basis $\mathcal{B}$ to be

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{n-\ell,\ell} & \boldsymbol{A}_{22} \end{bmatrix}.$$

Thus, by (A.3), $\boldsymbol{A} + \boldsymbol{b}\boldsymbol{f}^t$ has Z_Σ as an invariant subspace. Furthermore, by Algorithm 2.28, we know that the matrix $\boldsymbol{N}_\Sigma$ describing the zero dynamics is exactly $\boldsymbol{A}_{11}$.

We now claim that for all $s \in \mathbb{C}$ the matrix

$$\begin{bmatrix} s\boldsymbol{I}_{n-\ell} - \boldsymbol{A}_{22} & \boldsymbol{b}_2 \\ -\boldsymbol{c}_2^t & 0 \end{bmatrix} \tag{3.7}$$

is invertible. To show this, suppose that there exists a vector $(\boldsymbol{x}_2, u) \in \mathbb{R}^{n-\ell} \times \mathbb{R}$ with the property that

$$\begin{bmatrix} s\boldsymbol{I}_{n-\ell} - \boldsymbol{A}_{22} & \boldsymbol{b}_2 \\ -\boldsymbol{c}_2^t & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_2 \\ u \end{bmatrix} = \begin{bmatrix} (s\boldsymbol{I}_{n-\ell} - \boldsymbol{A}_{22})\boldsymbol{x}_2 + \boldsymbol{b}_2 u \\ -\boldsymbol{c}_2^t\boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ 0 \end{bmatrix}. \tag{3.8}$$

Define

$$Z = Z_\Sigma + \mathrm{span}((\boldsymbol{0}, \boldsymbol{x}_2)).$$

Since $Z_\Sigma \subset \ker(\boldsymbol{c}^t)$ and since $-\boldsymbol{c}_2^t\boldsymbol{x}_2 = 0$ we conclude that $Z \subset \ker(\boldsymbol{c}^t)$. Given the form of $\boldsymbol{A}_{21}$ in (3.6), we see that if $\boldsymbol{v} \in Z_\Sigma$, then $\boldsymbol{A}\boldsymbol{v} \in Z_\Sigma + \mathrm{span}(\boldsymbol{b})$. This shows that $Z \subset Z_\Sigma$, and from this we conclude that $(\boldsymbol{0}, \boldsymbol{x}_2) \in Z_\Sigma$ and so $\boldsymbol{x}_2$ must be zero. It then follows from (3.8) that $u = 0$, and this shows that the kernel of the matrix (3.7) contains only the zero vector, and so the matrix must be invertible.

Next we note that

$$\begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix} = \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ -\boldsymbol{f}^t & 1 \end{bmatrix}$$

and so

$$\det \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix} = \det \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix}. \tag{3.9}$$

We now rearrange the matrix on the left-hand side corresponding to our decomposition. The matrix for the linear map corresponding to this matrix in the basis $\mathcal{B}$ is

$$\begin{bmatrix} s\boldsymbol{I}_\ell - \boldsymbol{A}_{11} & -\boldsymbol{A}_{12} & \boldsymbol{0} \\ \boldsymbol{0}_{n-\ell,\ell} & s\boldsymbol{I}_{n-\ell} - \boldsymbol{A}_{22} & \boldsymbol{b}_2 \\ \boldsymbol{0}^t & \boldsymbol{c}_2^t & 0 \end{bmatrix}.$$

The determinant of this matrix is therefore exactly the determinant on the left-hand side of (3.9). This means that

$$\det \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \\ -\boldsymbol{c}^t & 0 \end{bmatrix} = \det \begin{bmatrix} s\boldsymbol{I}_\ell - \boldsymbol{A}_{11} & -\boldsymbol{A}_{12} & \boldsymbol{0} \\ \boldsymbol{0}_{n-\ell,\ell} & s\boldsymbol{I}_{n-\ell} - \boldsymbol{A}_{22} & \boldsymbol{b}_2 \\ \boldsymbol{0}^t & \boldsymbol{c}_2^t & 0 \end{bmatrix}.$$

By Lemma 3.3 we see that the left-hand determinant is exactly $\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$. Therefore, the values of s for which the left-hand side is zero are exactly the roots of the numerator of the transfer function. On the other hand, since the matrix (3.7) is invertible for all $s \in \mathbb{C}$, the values of s for which the right-hand side vanish must be those values of s for which $\det(s\boldsymbol{I}_\ell - \boldsymbol{A}_{11}) = 0$, i.e., the eigenvalues of $\boldsymbol{A}_{11}$. But we have already decided that $\boldsymbol{A}_{11}$ is the matrix that represents the zero dynamics, so this completes the proof. ■

This theorem is very important as it allows us to infer—at least in those cases where the transfer function is invertible—the nature of the zero dynamics from the transfer function. If there are zeros, for example, with positive real part we know our system has unstable zero dynamics, and we ought to be careful.

To further illustrate this connection between the transfer function and the zero dynamics, we give an example.

3.16 Example **(Example 2.27 cont'd)** Here we look again at Example 2.27. We have

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

We had computed $Z_\Sigma = \mathrm{span}((1,1))$, and so $\boldsymbol{b} \notin Z_\Sigma$. Thus $\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b}$ is not the zero polynomial by Lemma 3.14. Well, for pity's sake, we can just compute it:

$$\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} = 1 - s.$$

Since this is non-zero, we can apply Theorem 3.15 and conclude that the spectrum for the zero dynamics is $\{1\}$. This agrees with our computation in Example 2.29 where we computed

$$\boldsymbol{N}_\Sigma = \begin{bmatrix} 1 \end{bmatrix}.$$

Since $\mathrm{spec}(\boldsymbol{N}_\Sigma) \cap \mathbb{C}_+ \neq \emptyset$, the system is not minimum phase. •

3.17 Remark We close with an important remark. This section contains some technically demanding mathematics. If you can understand this, then that is really great, and I encourage you to try to do this. However, it is more important that you get the punchline here which is:

> *The transfer function contains a great deal of information about the behaviour of the system, and it does so in a deceptively simple manner.*

We will be seeing further implications of this as things go along. •

3.4 Transfer functions presented in input/output form

The discussion of the previous section supposes that we are given a state-space model for our system. However, this is sometimes not the case. Sometimes, all we are given is a scalar differential equation that describes how a scalar output behaves when given a scalar input. We suppose that we are handed an equation of the form

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1y^{(1)}(t) + p_0y(t) = \\ c_{n-1}u^{(n-1)}(t) + c_{n-1}u^{(n-2)}(t) + \cdots + c_1u^{(1)}(t) + c_0u(t) \quad (3.10)$$

for real constants $p_0, \dots, p_{n-1}$ and $c_0, \dots, c_{n-1}$. How might such a model be arrived at? Well, one might perform measurements on the system given certain inputs, and figure out that a differential equation of the above form is one that seems to accurately model what you are seeing. This is not a topic for this book, and is referred to as "model identification." For now, we will just suppose that we are given a system of the form (3.10). Note here that there are no states in this model! All there is is the input $u(t)$ and the output $y(t)$. Our system may possess states, but the model of the form (3.10) does not know about them. As we have already seen in the discussion following Theorem 2.37, there is a relationship between the systems we discuss in this section, and SISO linear systems. We shall further develop this relationship in this section.

For the moment, let us alleviate the nuisance of having to ever again write the expression (3.10). Given the differential equation (3.10) we define two polynomials in $\mathbb{R}[\xi]$ by

$$D(\xi) = \xi^n + p_{n-1}\xi^{n-1} + \cdots + p_1\xi + p_0 \\ N(\xi) = c_{n-1}\xi^{n-1} + c_{n-2}\xi^{n-2} + \cdots + c_1\xi + c_0.$$

Note that if we let $\xi = \frac{\mathrm{d}}{\mathrm{d}t}$ then we think of $D(\frac{\mathrm{d}}{\mathrm{d}t})$ and $N(\frac{\mathrm{d}}{\mathrm{d}t})$ as a differential operator, and we can write

$$D(\tfrac{\mathrm{d}}{\mathrm{d}t})(y) = y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1y^{(1)}(t) + p_0y(t).$$

In like manner we can think of $N(\frac{\mathrm{d}}{\mathrm{d}t})$ as a differential operator, and so we write

$$N(\tfrac{\mathrm{d}}{\mathrm{d}t})(u) = c_{n-1}u^{(n-1)}(t) + c_{n-1}u^{(n-2)}(t) + \cdots + c_1u^{(1)}(t) + c_0u(t).$$

In this notation the differential equation (3.10) reads $D(\frac{\mathrm{d}}{\mathrm{d}t})(y) = N(\frac{\mathrm{d}}{\mathrm{d}t})(u)$. With this little bit of notation in mind, we make some definitions.

3.18 Definition A ***SISO linear system in input/output form*** is a pair of polynomials (N, D) in $\mathbb{R}[s]$ with the properties

(i) D is monic and

(ii) D and N are coprime.

The ***relative degree*** of (N, D) is $\deg(D) - \deg(N)$. The system is ***proper*** (resp. ***strictly proper***) if its relative degree is nonnegative (resp. positive). If (N, D) is not proper, it is ***improper***. A SISO linear system (N, D) in input/output form is ***stable*** if D has no roots in $\overline{\mathbb{C}}_+$ and ***minimum phase*** if N has no roots in $\mathbb{C}_+$. If (N, D) is not minimum phase, it is ***nonminimum phase***. The ***transfer function*** associated with the SISO linear system (N, D) in input/output form is the rational function $T_{N,D}(s) = \frac{N(s)}{D(s)}$. •

3.19 Remarks

1. Note that in the definition we allow for the numerator to have degree greater than that of the denominator, even though this is not the case when the input/output system is derived from a differential equation (3.10). Our reason for doing this is that occasionally one does encounter transfer functions that are improper, or maybe situations where a transfer function, even though proper itself, is a product of rational functions, at least one of which is not proper. This will happen, for example, in Section 6.5 with the "derivative" part of PID control. Nevertheless, we shall for the most part be thinking of proper, or even strictly proper SISO linear systems in input/output form.
2. At this point, it is not quite clear what is the motivation behind calling a system (N, D) stable or minimum phase. However, this will certainly be clear as we discuss properties of transfer functions. This being said, a realisation of just what "stable" might mean will not be made fully until Chapter 5. •

If we take the causal left Laplace transform of the differential equation (3.10) we get simply $D(s)\mathscr{L}^+_{0-}(y)(s) = N(s)\mathscr{L}^+_{0-}(u)(s)$, provided that we suppose that both the input u and the output y are causal signals. Therefore we have

$$T_{N,D}(s) = \frac{\mathscr{L}^+_{0-}(y)(s)}{\mathscr{L}^+_{0-}(u)(s)} = \frac{N(s)}{D(s)} = \frac{c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \cdots + c_1 s + c_0}{s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0}.$$

Block diagrammatically, the situation is illustrated in Figure 3.7. We should be very clear

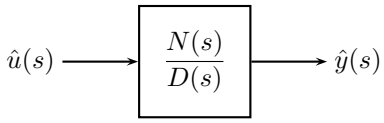


Figure 3.7 The block diagram representation of (3.10)

on why the diagrams Figure 3.6 and Figure 3.7 are different: *there are no state variables* $\boldsymbol{x}$ *in the differential equations (3.10).* All we have is an input/output relation. This raises the question of whether there is a connection between the equations in the form (3.1) and those in the form (3.10). Following the proof of Theorem 2.37 we illustrated how one can take differential equations of the form (3.1) and produce an input/output differential equation like (3.10), provided $(\boldsymbol{A}, \boldsymbol{b})$ is controllable. To go from differential equations of the form (3.10) and produce differential equations of the form (3.1) is in some sense artificial, as we would have to "invent" states that are not present in (3.10). Indeed, there are infinitely many ways to introduce states into a given input/output relation. We shall look at the one that is related to Theorem 2.37. It turns out that the best way to think of making the connection from (3.10) to (3.1) is to use transfer functions.

3.20 Theorem *Let (N, D) be a proper SISO linear system in input/output form. There exists a complete SISO linear control system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ so that $T_\Sigma = T_{N,D}$.*

Proof Let us write

$$\begin{aligned} D(s) &= s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \\ N(s) &= \tilde{c}_n s^n + \tilde{c}_{n-1}s^{n-1} + \tilde{c}_{n-2}s^{n-2} + \cdots + \tilde{c}_1 s + \tilde{c}_0. \end{aligned}$$

We may write $\frac{N(s)}{D(s)}$ as

$$\begin{aligned} \frac{N(s)}{D(s)} &= \frac{\tilde{c}_n s^n + \tilde{c}_{n-1}s^{n-1} + \tilde{c}_{n-2}s^{n-2} + \cdots + \tilde{c}_1 s + \tilde{c}_0}{s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0} \\ &= \tilde{c}_n \frac{s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0}{s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0} + \\ &\quad \frac{(\tilde{c}_{n-1} - \tilde{c}_n p_{n-1})s^{n-1} + (\tilde{c}_{n-2} - \tilde{c}_n p_{n-2})s^{n-2} + \cdots + (\tilde{c}_1 - \tilde{c}_n p_1)s + (\tilde{c}_0 - \tilde{c}_n p_0)}{s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0} \\ &= \tilde{c}_n + \frac{c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \cdots + c_1 s + c_0}{s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0}, \end{aligned}$$

where $c_i = \tilde{c}_i - \tilde{c}_n p_i$, $i = 0, \ldots, n-1$. Now define

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{n-2} \\ c_{n-1} \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} \tilde{c}_n \end{bmatrix}. \tag{3.11}$$

By Exercise E2.11 we know that $(\boldsymbol{A}, \boldsymbol{b})$ is controllable. Since D and N are by definition coprime, by Theorem 3.5 $(\boldsymbol{A}, \boldsymbol{c})$ is observable. In the proof of Theorem 3.5 we showed that

$$\frac{\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}}{\det(s\boldsymbol{I}_n - \boldsymbol{A})} = \frac{c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \cdots + c_1 s + c_0}{s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0},$$

(see equation (3.4)), and from this follows our result. ■

We shall denote the SISO linear control system Σ of the theorem, i.e., that one given by (3.11), by $\Sigma_{N,D}$ to make explicit that it comes from a SISO linear system in input/output form. We call $\Sigma_{N,D}$ the ***canonical minimal realisation*** of the transfer function $T_{N,D}$. Note that condition (ii) of Definition 3.18 and Theorem 3.5 ensure that $(\boldsymbol{A}, \boldsymbol{c})$ is observable. This establishes a way of getting a linear system from one in input/output form. However, it *not* the case that the linear system $\Sigma_{N,D}$ should be thought of as representing the physical states of the system, but rather it represents *only* the input/output relation. There are consequences of this that you need to be aware of (see, for example, Exercise E3.20).

It is possible to represent the above relation with a block diagram with each of the states $x_1, \ldots, x_n$ appearing as a signal. Indeed, you should verify that the block diagram of Figure 3.8 provides a transfer function which is exactly

$$\frac{\mathscr{L}^+_{0-}(y)s)}{\mathscr{L}^+_{0-}(u)(s)} = \frac{c_{n-1}s^{n-1} + \cdots + c_1 s + c_0}{s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0} + \boldsymbol{D}.$$

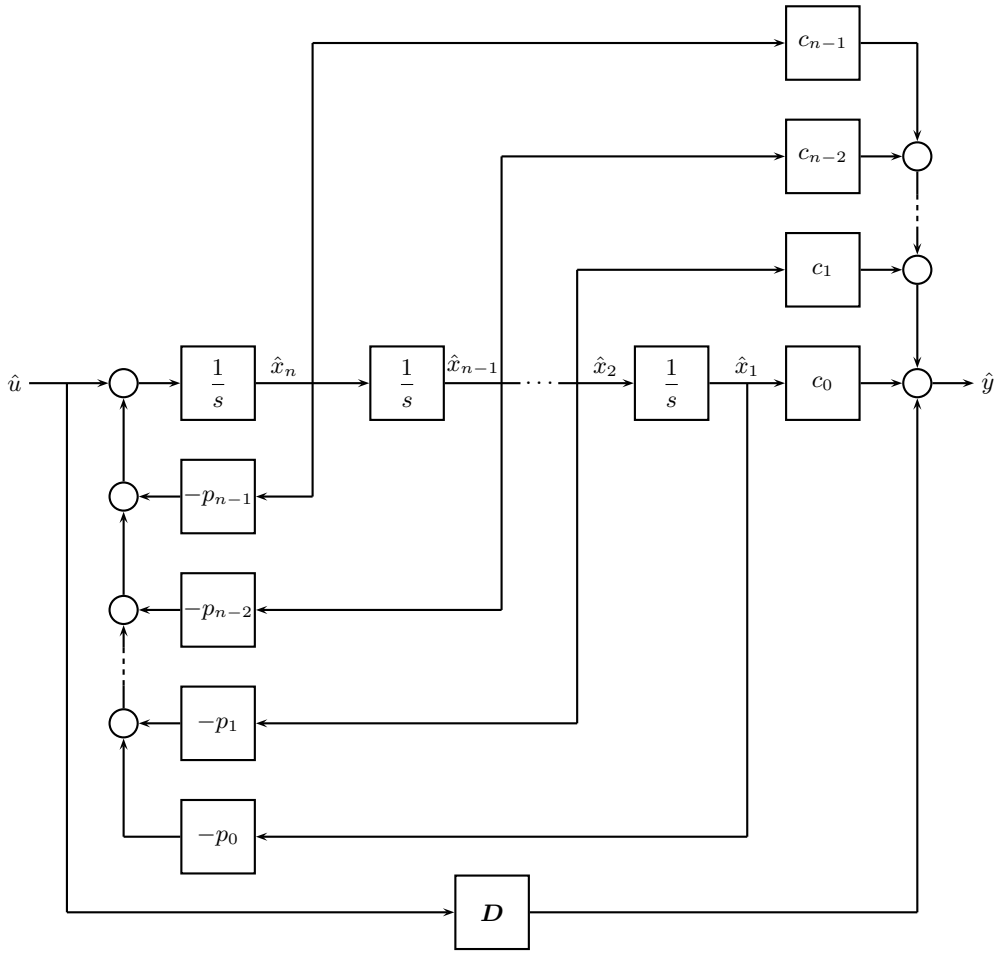


Figure 3.8 A block diagram for the SISO linear system of Theorem 3.20

Note that this also provides us with a way of constructing a block diagram corresponding to a transfer function, even though the transfer function may have been obtained from a different block diagram. The block diagram of Figure 3.8 is particularly useful if you live in mediaeval times, and have access to an ***analogue computer***...

3.21 Remark Note that we have provided a system in controller canonical form corresponding to a system in input/output form. Of course, it is also possible to give a system in observer canonical form. This is left as Exercise E3.19 for the reader. •

Theorem 3.20 allows us to borrow some concepts that we have developed for linear systems of the type (3.1), but which are not obviously applicable to systems in the form (3.10). This can be a useful thing to do. For example, motivated by Theorem 3.15, our notion that a SISO linear system (N, D) in input/output form is minimum phase if all roots of N lie in $\mathbb{C}_+$, and nonminimum phase otherwise, makes some sense.

Also, we can also use the correspondence of Theorem 3.20 to make a sensible notion of impulse response for SISO systems in input/output form. The problem with a direct definition is that if we take $u(t)$ to be a limit of inputs from $\mathscr{U}$ as described in Theorem 2.34, it is not clear what we should take for $u^{(k)}(t)$ for $k \geq 1$. However, from the transfer function point of view, this is not a problem. To wit, if (N, D) is a strictly proper SISO linear system

in input/output form, its ***impulse response*** is given by

$$h_{N,D}(t) = 1(t)\boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}$$

where $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \mathbf{0}_1) = \Sigma_{N,D}$. As with SISO linear systems, we may define the causal impulse response $h^+_{N,D}\colon [0,\infty) \to \mathbb{R}$ and the anticausal impulse response $h^-_{N,D}\colon (-\infty, 0] \to \mathbb{R}$. Also, as with SISO linear systems, it is the causal impulse response we will most often use, so we will frequently just write $h_{N,D}$, as we have already done, for $h^+_{N,D}$.

We note that it is not a simple matter to define the impulse response for a SISO linear system in input/output form that is proper but not strictly proper. The reason for this is that the impulse response is not realisable by a piecewise continuous input $u(t)$. However, if one is willing to accept the notion of a "delta-function," then one may form a suitable notion of impulse response. How this may be done *without* explicit recourse to delta-functions is outlined in Exercise E3.1.

3.5 The connection between the transfer function and the impulse response

We can make some statements about how the transfer function impinges upon the impulse response. We made some off the cuff remarks about how the impulse response contained the essential time-domain character of the system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$. We justified this in some way by stating Proposition 2.32. We will now see that the information contained in the impulse response is also directly contained in the transfer function, and further in about the simplest way possible.

3.5.1 Properties of the causal impulse response

We begin by looking at the case that is the most interesting for us; the causal impulse response. Our discussions in this section will be important for the development of certain fundamental aspects of, for example, input/output stability.

We begin by making the essential connection between the impulse and the transfer function. We state the result for both the left and right causal Laplace transform of the impulse response.

3.22 Theorem *For a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$, we have*

$$\mathscr{L}^+_{0+}(h_\Sigma)(s) = \mathscr{L}^+_{0-}(h_\Sigma)(s) = T_\Sigma(s)$$

provided that $\mathrm{Re}(s) > \sigma_{\min}(h^+_\Sigma)$.[1]

Proof Since h^+_Σ has no delta-function at $t = 0$ as we are supposing $\boldsymbol{D} = \mathbf{0}_1$, we have

$$\mathscr{L}^+_{0+}(h_\Sigma)(s) = \mathscr{L}^+_{0-}(h_\Sigma)(s) = \int_0^\infty h^+_\Sigma(t)e^{-st}\,\mathrm{d}t.$$

Thus clearly it does not matter in this case whether we use the left or right Laplace transform. We have

$$\mathscr{L}^+_{0-}(h_\Sigma)(s) = \int_0^\infty h^+_\Sigma(t)e^{-st}\,\mathrm{d}t = \int_0^\infty \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}e^{-st}\,\mathrm{d}t = \int_0^\infty \boldsymbol{c}^t e^{-st}e^{\boldsymbol{A}t}\boldsymbol{b}\,\mathrm{d}t.$$

[1] Here is an occasion where we actually think of T_Σ as a function of the complex variable s rather than as a rational function.

Recall that the matrix exponential has the property that

$$e^a \boldsymbol{x} = e^{a\boldsymbol{I}_n}\boldsymbol{x}$$

for $a \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$. Therefore we have

$$\mathscr{L}^+_{0-}(h_\Sigma)(s) = \int_0^\infty \boldsymbol{c}^t e^{-st\boldsymbol{I}_n} e^{\boldsymbol{A}t}\boldsymbol{b}\,\mathrm{d}t.$$

Now recall that if $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n\times n}$ have the property that $\boldsymbol{BC} = \boldsymbol{CB}$ then $e^{\boldsymbol{B}+\boldsymbol{C}} = e^{\boldsymbol{B}}e^{\boldsymbol{C}}$. Noting that the $n \times n$ matrices $-st\boldsymbol{I}_n$ and $\boldsymbol{A}t$ commute, we then have

$$\mathscr{L}^+_{0-}(h_\Sigma)(s) = \int_0^\infty \boldsymbol{c}^t e^{(-s\boldsymbol{I}_n+\boldsymbol{A})t}\boldsymbol{b}\,\mathrm{d}t.$$

Now we note that

$$\frac{\mathrm{d}}{\mathrm{d}t}(-s\boldsymbol{I}_n + \boldsymbol{A})^{-1}e^{(-s\boldsymbol{I}_n+\boldsymbol{A})t} = e^{(-s\boldsymbol{I}_n+\boldsymbol{A})t},$$

from which we ascertain that

$$\mathscr{L}^+_{0-}(h_\Sigma)(s) = \boldsymbol{c}^t(-s\boldsymbol{I}_n + \boldsymbol{A})^{-1}e^{(-s\boldsymbol{I}_n+\boldsymbol{A})t}\boldsymbol{b}\Big|_0^\infty.$$

Since the terms in the matrix $e^{(-s\boldsymbol{I}_n+\boldsymbol{A})t}$ satisfy an inequality like (E.2), the upper limit on the integral is zero so we have

$$\mathscr{L}^+_{0-}(h_\Sigma)(s) = -\boldsymbol{c}^t(-s\boldsymbol{I}_n + \boldsymbol{A})^{-1}\boldsymbol{b} = \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = T_\Sigma(s)$$

as claimed. ■

3.23 Remark Note that we ask that the feedforward term $\boldsymbol{D}$ be zero in the theorem. This can be relaxed provided that one is willing to think about the "delta-function." The manner in which this can be done is the point of Exercise E3.1. When one does this, the theorem no longer holds for both the left and right causal Laplace transforms, but only holds for the left causal Laplace transform. •

We should, then, be able to glean the behaviour of the impulse response by looking only at the transfer function. This is indeed the case, and this will now be elucidated. You will recall that if $f(t)$ is a positive real-valued function then

$$\limsup_{t\to\infty} f(t) = \lim_{t\to\infty}\big(\sup_{\tau>t} f(\tau)\big).$$

The idea is that $\limsup_{t\to\infty}$ will exist for bounded functions that "oscillate" whereas $\lim_{t\to\infty}$ may not exist for such functions. With this notation we have the following result.

3.24 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system with impulse response h^+_Σ and transfer function T_Σ. Write*

$$T_\Sigma(s) = \frac{N(s)}{D(s)}$$

where (N, D) is the c.f.r. of T_Σ. The following statements hold.

(i) If D has a root in $\mathbb{C}_+$ then

$$\lim_{t\to\infty}|h^+_\Sigma(t)| = \infty.$$

(ii) If all roots of D are in $\mathbb{C}_-$ then

$$\lim_{t\to\infty} |h_\Sigma^+(t)| = 0.$$

(iii) If D has no roots in $\mathbb{C}_+$, but has distinct roots on $i\mathbb{R}$, then

$$\limsup_{t\to\infty} |h_\Sigma^+(t)| = M$$

for some $M > 0$.

(iv) If D has repeated roots on $i\mathbb{R}$ then

$$\lim_{t\to\infty} |h_\Sigma^+(t)| = \infty.$$

Proof (i) Corresponding to a root with positive real part will be a term in the partial fraction expansion of T_Σ of the form

$$\frac{\beta}{(s-\sigma)^k} \quad \text{or} \quad \frac{\beta_1 s + \beta_0}{\left((s-\sigma)^2 + \omega^2\right)^k}$$

with $\sigma > 0$. By Proposition E.11, associated with such a term will be a term in the impulse response that is a linear combination of functions of the form

$$t^\ell e^{\sigma t} \quad \text{or} \quad t^\ell e^{\sigma t}\cos\omega t \quad \text{or} \quad t^\ell e^{\sigma t}\sin\omega t.$$

Such terms will clearly blow up to infinity as t increases.

(ii) If all roots lie in the negative half-plane then all terms in the partial fraction expansion for T_Σ will have the form

$$\frac{\beta}{(s+\sigma)^k} \quad \text{or} \quad \frac{\beta_1 s + \beta_0}{\left((s+\sigma)^2 + \omega^2\right)^k}$$

for $\sigma > 0$. Again by Proposition E.11, associated with these terms are terms in the impulse response that are linear combinations of functions of the form

$$t^\ell e^{-\sigma t} \quad \text{or} \quad t^\ell e^{-\sigma t}\cos\omega t \quad \text{or} \quad t^\ell e^{-\sigma t}\sin\omega t.$$

All such functions decay to zero at t increases.

(iii) The roots in the negative half-plane will give terms in the impulse response that decay to zero, as we saw in the proof for part (ii). For a distinct complex conjugate pair of roots $\pm i\omega$ on the imaginary axis, the corresponding terms in the partial fraction expansion for T_Σ will be of the form

$$\frac{\beta_1 s + \beta_0}{s^2 + \omega}$$

which, by Proposition E.11, lead to terms in the impulse response that are linear combinations of $\cos\omega t$ and $\sin\omega t$. This will give a bounded oscillatory time response as $t \to \infty$, and so the resulting lim sup will be positive.

(iv) If there are repeated imaginary roots, these will lead to terms in the partial fraction expansion for T_Σ of the form

$$\frac{\beta_1 s + \beta_0}{(s^2 + \omega^2)^k}.$$

The corresponding terms in the impulse response will be linear combinations of functions like $t^\ell \cos\omega t$ or $t^\ell \sin\omega t$. Such functions clearly are unbounded at $t \to \infty$. ■

This result is important because it tells us that we can understand a great deal about the "stability" of a system by examining the transfer function. Indeed, this is the whole idea behind classical control: one tries to make the transfer function look a certain way. In this section we have tried to ensure that we fully understand the relationship between the time-domain and the transfer function, as this relationship is a complete one most of the time (you'll notice that we have made controllability and/or observability assumptions quite often, but many systems you encounter will be both controllable and observable).

3.5.2 Things anticausal

We will have occasion to make a few computations using the anticausal impulse response, and the causal impulse response in relation to anticausal inputs. In this section we collect the relevant results.

The first result is a natural analogue of Theorem 3.22. Since the proof is exactly the same with the exception of some signs, it is omitted.

3.25 Theorem *For a SISO linear system* $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ *we have*

$$\mathscr{L}^-_{0+}(h^-_\Sigma)(s) = \mathscr{L}^-_{0-}(h^-_\Sigma)(s) = T_\Sigma(s),$$

provided that $\mathrm{Re}(s) < \sigma_{\max}(h^-_\Sigma)$.

Of course, we also have an analogue of Proposition 3.24 for the anticausal impulse response, where the statement is altered by replacing $\mathbb{C}_+$ with $\mathbb{C}_-$, and vice versa, and by replacing $\lim_{t\to\infty}$ with $\lim_{t\to-\infty}$. However, we shall not make use of this result, so do not state it. Instead, let us state a few results about causal inputs and the anticausal impulse response, and vice versa.

3.26 Proposition *Let* $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ *be a SISO linear system.* For NP suff proof

(i)

Proof ■

3.6 The matter of computing outputs

Given a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, Theorem 2.6 gives us a manner of determining the state, and therefore the output, using the matrix exponential. While this is certainly a valid way to compute the output, it is really only useful in simple cases, where the matrix exponential formalism provides a way of determining a nice symbolic expression for the output. Similarly, using the right causal Laplace transform, it is a simple matter to compute outputs in cases where the Laplace transform can be inverted. However, this is often not the case. Furthermore, the matrix exponential and Laplace transform techniques are not readily implemented numerically. Therefore, in this section we look at ways of writing scalar differential equations for determining output that *can* be readily implemented numerically.

In the course of this discussion we shall be forced to come to grips with the difference between the left and right causal Laplace transform. To be in any way accurate, *this must be done*. That this is normally glossed over is a reflection of the willingness to oversimplify the use of the Laplace transform. What we shall essentially see is a dichotomy of the following type for the use of the two forms of causal transform.

1. For solving differential equations whose right-hand side involves no delta-function, the right causal Laplace transform is the most convenient tool. Note that this precludes the convenient use of the right causal Laplace transform with the impulse response since the impulse response involves a delta-function on the right-hand side.
2. For general control theoretic discussions, the left causal Laplace transform is more convenient since it eliminates the appearance of the pesky initial condition terms present in the expressions for derivatives. Indeed, in the places in the discussion above where the Laplace transform was used, it was the left causal transform that was used.

Thus we see that there is a natural tension between whether to use the left or right causal Laplace transform. It is important to realise that the two things are different, and that their differences sometimes make one preferable over the other.

3.6.1 Computing outputs for SISO linear systems in input/output form using the right causal Laplace transform

We begin looking at systems that are given to us in input/output form. Thus we have a SISO linear system (N, D) in input/output form with $\deg(D) = n$, and we are concerned with obtaining solutions to the initial value problem

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)u(t), \quad y(0+) = y_0,\ y^{(1)}(0+) = y_1,\ \ldots,\ y^{(n-1)}(0+) = y_{n-1}, \tag{3.12}$$

for a known function $u(t)$. We shall assume that u is sufficiently differentiable that the needed derivatives exist at $t = 0+$, and that u possesses a right causal Laplace transform. Then, roughly speaking, if one wishes to use Laplace transforms one determines the right causal Laplace transform $\mathscr{L}^+_{0+}(u)$, and then inverse Laplace transforms the function

$$\mathscr{L}^+_{0+}(y)(s) = \frac{N(s)}{D(s)}\mathscr{L}^+_{0+}(u)(s)$$

to get $y(t)$. This is indeed very rough, of course, because we have lost track of the initial conditions in doing this. The following result tells us how to use the Laplace transform method to obtain the solution to (3.12), properly keeping track on initial conditions.

3.27 Proposition *Let (N, D) be a proper SISO linear system in input/output form, let $u\colon [0,\infty) \to \mathbb{R}$ possess a right causal Laplace transform, and assume that $u, u^{(1)}, \ldots, u^{(\deg(N)-1)}$ are continuous on $[0,\infty)$ and that $u^{(\deg(N))}$ is piecewise continuous on $[0,\infty)$. Then the right causal Laplace transform of the solution to the initial value problem (3.12) is*

$$\mathscr{L}^+_{0+}(y)(s) = \frac{1}{D(s)}\Big(N(s)\mathscr{L}^+_{0+}(u)(s) + \sum_{k=1}^{n}\sum_{j=0}^{k-1}\big(p_k s^j y^{(k-j-1)}(0+) - c_k s^j u^{(k-j-1)}(0+)\big)\Big),$$

where

$$\begin{aligned} N(s) &= c_n s^n + c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \cdots + c_1 s + c_0\\ D(s) &= s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0, \end{aligned}$$

and where we take $p_n = 1$.

Proof Note that by Corollary E.8 the right causal Laplace transform of $y^{(k)}(t)$ is given by

$$s^k\mathscr{L}^+_{0+}(y)(s) - \sum_{j=0}^{k-1} s^j y^{(k-j-1)}(0+), \tag{3.13}$$

with a similar formula holding, of course, for $u^{(k)}(t)$. Thus, if we take the Laplace transform of the differential equation in (3.12) and use the initial conditions given there we get the relation

$$\sum_{k=0}^{n} p_k \Big(s^k \mathscr{L}^+_{0+}(y)(s) - \sum_{j=0}^{k-1} s^j y^{(k-j-1)}(0+) \Big) = \sum_{k=0}^{n} c_k \Big(s^k \mathscr{L}^+_{0+}(u)(s) - \sum_{j=0}^{k-1} s^j u^{(k-j-1)}(0+) \Big),$$

where we take $p_n = 1$. The result now follows by a simple manipulation. ■

3.28 Remark The assumptions of differentiability on the input can be relaxed as one can see by looking at the proof that one really only needs for u to satisfy the conditions of the Laplace transform derivative theorem, Theorem E.7, for a sufficiently large number of derivatives. This can be made true for a large class of functions by using the definition of the Laplace transform on distributions. This same observation holds for results that follow and possess the same hypotheses. •

One may, if it is possible, use Proposition 3.27 to obtain the solution $y(t)$ by using the inverse Laplace transform. Let's see how this works in an example.

3.29 Example We have $(N(s), D(s)) = (1 - s, s^2 + 2s + 2)$ and we take as input $u(t) = 1(t)t$. The initial value problem we solve is

$$\ddot{y}(t) + 2\dot{y}(t) + 2y(t) = t - 1, \quad y(0+) = 1, \;\; \dot{y}(0+) = 0.$$

Taking the Laplace transform of the left-hand side of this equation gives

$$(s^2 + 2s + 2)\mathscr{L}^+_{0+}(y)(s) - (s + 2)y(0+) - \dot{y}(0+) = (s^2 + 2s + 2)\mathscr{L}^+_{0+}(y)(s) - s - 2.$$

One may verify that $(s + 2)y(0+) + \dot{y}(0+)$ is exactly the expression

$$\sum_{k=1}^{n} \sum_{j=0}^{k-1} p_k s^j y^{(k-j-1)}(0+)$$

in the statement of Proposition 3.27 with D as given. The Laplace transform of the right-hand side of the differential equation is

$$(1 - s)\mathscr{L}^+_{0+}(u)(s) + u(0+) = \frac{1 - s}{s^2},$$

using the fact that the Laplace transform of $u(t)$ is $\frac{1}{s^2}$. As with the expression involving the left-hand side of the equation, we note that $-u(0+)$ is exactly the expression

$$-\sum_{k=1}^{n} \sum_{j=0}^{k-1} c_k s^j u^{(k-j-1)}(0+)$$

in the statement of Proposition 3.27. Combining our work, the Laplace transform of the differential equation gives

$$\mathscr{L}^+_{0+}(y)(s) = \frac{1 - s}{s^2(s^2 + 2s + 2)} + \frac{s + 2}{s^2 + 2s + 2}.$$

To compute the inverse Laplace transform we perform the partial fraction expansion for the first term to get

$$\frac{1-s}{s^2(s^2+2s+2)} = \frac{1}{2s^2} - \frac{1}{s} + \frac{s+\frac{3}{2}}{s^2+2s+2}.$$

Thus we obtain

$$\mathscr{L}^+_{0+}(y)(s) = \frac{1}{2s^2} - \frac{1}{s} + \frac{2s+\frac{7}{2}}{s^2+2s+2}$$

The inverse Laplace transform of the first term is $\frac{1}{2}t$, and of the second term is -1. To determine the Laplace transform of the third term we note that the inverse Laplace of

$$\frac{1}{s^2+2s+2}$$

is $e^{-t}\sin t$, and the inverse Laplace transform of

$$\frac{s+1}{s^2+2s+2}$$

is $e^{-t}\cos t$. Putting this all together we have obtained the solution

$$y(t) = e^{-t}\big(\tfrac{3}{2}\sin t + 2\cos t\big) + \tfrac{t}{2} - 1$$

to our initial value problem. •

This example demonstrates how tedious can be the matter of obtaining "by hand" the solution to even a fairly simple initial value problem.

It is sometimes preferable to obtain the solution in the time-domain, particularly for systems of high-order since the inverse Laplace transform will typically be difficult to obtain in such cases. One way to do this is proposed in Section 3.6.3.

3.6.2 Computing outputs for SISO linear systems in input/output form using the left causal Laplace transform

Finish

As mentioned in the preamble to this section, the right causal Laplace transform is the more useful than its left brother for solving initial value problems, by virtue of its encoding the initial conditions in a convenient manner. However, it is *possible* to solve these same problems using the left causal Laplace transform. To do so, since the initial conditions at $t = 0-$ are all zero (we always work with causal inputs and outputs), it turns out that the correct way to encode the initial conditions at $t = 0+$ is to add delta-function inputs. While this is not necessarily the recommended way to solve such equations, it does make precise the connection between the left and right causal transforms in this case. Also, it allows us to better understand such things as the impulse response.

Let us begin by considering a proper SISO linear system (N, D) in input/output form. For this system, let us consider inputs of the form

$$u(t) = 1(t)u_0(t) + c\delta(t) \tag{3.14}$$

where $u_0 \colon \mathbb{R} \to \mathbb{R}$ has the property that $u_0, u_0^{(1)}, \dots, u_0^{(\deg(N)-1)}$ are continuous on $[0, \infty)$ and that $u_0^{(\deg(N))}$ is piecewise continuous on $[0, \infty)$. Thus we allow inputs which satisfy the derivative rule for the right causal Laplace transform (cf. Theorem E.7), and which additionally allows a delta-function as input. This allows us to consider the impulse response

as part of the collection of problems considered. Note that by allowing a delta-function in our input, we essentially mandate the use of the left causal Laplace transform to solve the equation. Since we are interested in causal outputs, we ask that the output have initial conditions

$$y(0-) = 0,\ y^{(1)}(0-),\ \dots,\ y^{(n-1)}(0-) = 0, \tag{3.15}$$

where, as usual, $n = \deg(D)$.

Let us first state the solution of the problem just formulated.

3.30 Proposition *Let (N, D) be a proper SISO linear system in input/output form with input u as given by (3.14). The solution y of the initial value problem*

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$$

with initial conditions (3.15) has the form $y(t) = y_0(t) + d\delta(t)$ where $d =$ and where y_0 is the solution of the initial value problem

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y_0(t) = N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)u_0(t), \quad y_0(0+) =,\ y_0^{(1)}(0+) =,\ \dots,\ y_0^{(n-1)}(0+) =,$$

where

$$N(s) = c_n s^n + c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \cdots + c_1 s + c_0$$
$$D(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0.$$

Proof Taking the left causal Laplace transform of the differential equation, using Corollary E.8, gives

$$D(s)\mathscr{L}^+_{0-}(y)(s) = N(s)\mathscr{L}^+_{0-}(u_0)(s) + \sum_{j=0}^{n} c_j c s^j.$$

Since the initial conditions at $t = 0-$ for y are all zero, we may express $y = y_0 + y_1$ where

$$D(s)\mathscr{L}^+_{0-}(y_0)(s) = N(s)\mathscr{L}^+_{0-}(u_0)(s)$$

and

$$D(s)\mathscr{L}^+_{0-}(y_1)(s) = \sum_{j=0}^{n} c_j c s^j.$$

■

3.6.3 Computing outputs for SISO linear systems in input/output form using the causal impulse response

To see how the Laplace transform method is connected with solving equations in the time-domain, we make the observation, following from Theorem 3.22, that the Laplace transform of the impulse response $h_{N,D}$ is the transfer function $T_{N,D}$ in the case where (N, D) is strictly proper. Therefore, by Exercise EE.5, the inverse Laplace transform of $T_{N,D}(s)\mathscr{L}^+_{0+}(u)(s)$ is

$$\int_0^t h_{N,D}(t-\tau)u(\tau)\,\mathrm{d}\tau. \tag{3.16}$$

Let us address the question of which initial value problem of the form (3.12) has the expression (3.16) as its solution. That is, let us determine the proper initial conditions to obtain the solution (3.16).

3.31 Proposition *Let (N, D) be a strictly proper SISO linear system in input/output form, let $u\colon [0,\infty) \to \mathbb{R}$ possess a right causal Laplace transform, and assume that $u, u^{(1)}, \dots, u^{(\deg(N)-1)}$ are continuous on $[0,\infty)$ and that $u^{(\deg(N))}$ is piecewise continuous on $[0,\infty)$. Then the integral (3.16) is the solution to the initial value problem (3.12) provided that the initial values $y(0+), y^{(1)}(0+), \dots, y^{(n-1)}(0+)$ are chosen as follows:*

(i) let $y(0+) = 0$;

(ii) recursively define $y^{(k)}(0+) = \sum_{j=1}^{k} \big(c_{n-j} u^{(k-j)}(0+) - p_{n-j} y^{(k-j)}(0+)\big)$, $k = 1, \dots, n-1$.

Proof From Proposition 3.27 it suffices to show that the initial conditions we have defined are such that

$$\sum_{k=1}^{n} \sum_{j=0}^{k-1} \big(p_k s^j y^{(k-j-1)}(0+) - c_k s^j u^{(k-j-1)}(0+)\big) = 0, \quad k = 1, \dots, n-1.$$

This will follow if the coefficient of each power of s in the preceding expression vanishes. Starting with the coefficient of s^{n-1} we see that $y(0+) = 0$. The coefficient of s^{n-2} is then determined to be

$$y^{(1)}(0+) + p_{n-1} y(0+) - c_{n-1} u(0+),$$

which gives $y^{(1)}(0+) = c_{n-1} u(0+) - p_{n-1} y(0+)$. Proceeding in this way, we develop the recursion relation as stated in the proposition. ■

Let's see how we may use this to obtain a solution to an initial value problem in the time-domain using the impulse response $h_{N,D}$.

3.32 Proposition *Let (N, D) be a strictly proper SISO linear system in input/output form, let $u\colon [0,\infty) \to \mathbb{R}$ possess a right causal Laplace transform, and assume that $u, u^{(1)}, \dots, u^{(\deg(N)-1)}$ are continuous on $[0,\infty)$ and that $u^{(\deg(N))}$ is piecewise continuous on $[0,\infty)$. Let $\tilde{y}_0, \tilde{y}_1, \dots, \tilde{y}_{n-1}$ be the initial conditions as defined in Proposition 3.31, and suppose that $y_h(t)$ solves the initial value problem*

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big) y_h(t) = 0, \quad y(0+) = y_0 - \tilde{y}_0, \ , y^{(1)}(0+) = y_1 - \tilde{y}_1, \ \dots, \ y^{(n-1)}(0+) = y_{n-1} - \tilde{y}_{n-1}.$$

Then the solution to the initial value problem (3.12) is given by

$$y(t) = y_h(t) + \int_0^t h_{N,D}(t-\tau) u(\tau)\, \mathrm{d}\tau. \tag{3.17}$$

Proof That *every* solution of (3.12) can be expressed in the form of (3.17) follows from Proposition 2.32. It thus suffices to show that the given solution satisfies the initial conditions. However, by the definition of y_h and by Proposition 3.31 we have

$$\begin{aligned}
y(0+) &= y_h(0+) + \int_0^0 h_{N,D}(-\tau) u(\tau)\, \mathrm{d}\tau = y_0 - \tilde{y}_0 = y_0 \\
y^{(1)}(0+) &= y_h^{(1)}(0+) + \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \int_0^t h_{N,D}(t-\tau) u(\tau)\, \mathrm{d}\tau = y_1 - \tilde{y}_1 + \tilde{y}_1 = y_1 \\
&\vdots \\
y^{(n-1)}(0+) &= y_h^{(1)}(0+) + \frac{\mathrm{d}^{n-1}}{\mathrm{d}t^{n-1}}\Big|_{t=0} \int_0^t h_{N,D}(t-\tau) u(\tau)\, \mathrm{d}\tau = y_{n-1} - \tilde{y}_{n-1} + \tilde{y}_{n-1} = y_{n-1},
\end{aligned}$$

which are the desired initial conditions. ■

Let's see how this works in a simple example.

3.33 Example **(Example 3.29 cont'd)** We take $(N(s), D(s)) = (1-s, s^2+2s+2)$ so the differential equation is

$$\ddot{y}(t) + 2\dot{y}(t) + 2y(t) = u(t) - \dot{u}(t).$$

As input we take $u(t) = t$ and as initial conditions we take $y(0+) = 1$ and $\dot{y}(0+) = 0$. We first obtain the homogeneous solution $y_h(t)$. We first determine the initial conditions, meaning we need to determine the initial conditions $\tilde{y}_0$ and $\tilde{y}_1$ from Proposition 3.31. These we readily determine to be $\tilde{y}_0 = 0$ and $\tilde{y}_1 = c_1u(0+) - p_1\tilde{y}_0 = 0$. Thus y_h should solve the initial value problem

$$\ddot{y}_h(t) + 2\dot{y}_h(t) + 2y_h(t) = 0, \quad y_h(0+) = 1,\ \dot{y}_h(0+) = 0.$$

Recall how to solve this equation. One first determines the roots of the characteristic polynomial that in this case is s^2+2s+2. The roots we compute to be $-1 \pm i$. This gives rise to two linearly independent solutions to the homogeneous differential equation, and these can be taken to be

$$y_1(t) = e^{-t}\cos t, \quad y_2(t) = e^{-t}\sin t.$$

Any solution to the homogeneous equation is a sum of these two solutions, so we must have $y_h(t) = C_1y_1(t) + C_2y_2(t)$ for appropriate constants C_1 and C_2. To determine these constants we use the initial conditions:

$$\begin{aligned} y_h(0+) &= C_1 = 1 \\ \dot{y}_h(0+) &= -C_1 + C_2 = 0, \end{aligned}$$

from which we ascertain that $C_1 = C_2 = 1$ so that $y_h(t) = e^{-t}(\cos t + \sin t)$.

We now need the impulse response that we can compute however we want (but see Example 3.41 for a slick way to do this) to be

$$h_{N,D}(t) = e^{-t}(2\sin t - \cos t).$$

Now we may determine

$$\int_0^t h_{N,D}(t-\tau)u(\tau)\,\mathrm{d}\tau = \int_0^t e^{-(t-\tau)}(2\sin(t-\tau) - \cos(t-\tau))\tau\,\mathrm{d}\tau = \tfrac{t}{2} - 1 + e^{-t}\big(\cos t + \tfrac{1}{2}\sin t\big)$$

Therefore, the solution to the initial value problem

$$\ddot{y}(t) + 2\dot{y}(t) + 2y(t) = t - 1, \quad y(0+) = 1,\ \dot{y}(0+) = 0$$

is

$$y(t) = e^{-t}\big(\tfrac{3}{2}\sin t + 2\cos t\big) + \tfrac{t}{2} - 1,$$

agreeing with what we determined using Laplace transforms in Example 3.29. •

In practice, one does not—at least I do not—solve simple differential equations this way. Rather, I typically use the "method of undetermined coefficients" where the idea is that after solving the homogeneous problem, the part of the solution that depends on the right-hand side is made a general function of the same "form" as the right-hand side, but with undetermined coefficients. One then resolves the coefficients by substitution into the differential equation. This method can be found in your garden variety text on ordinary differential equations, for example [Boyce and Diprima 1972]. A too quick overview is given in Section B.1.

3.6.4 Computing outputs for SISO linear systems

Next we undertake the above constructions for SISO linear systems $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{D})$. Of course, we may simply compute the transfer function T_Σ and proceed using the tools we developed above for systems in input/output form. However, we wish to see how the additional structure for state-space systems comes into play. Thus in this section we consider the initial value problem

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t), \quad \boldsymbol{x}(0+) = \boldsymbol{x}_0 \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t). \end{aligned} \tag{3.18}$$

Let us first make a somewhat trivial observation.

3.34 Lemma *Let $y(t)$ be the output for the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ subject to the input $u(t)$ and the initial condition $\boldsymbol{x}(0+) = \boldsymbol{x}_0$: thus $y(t)$ is defined by (3.18). Then there exists unique functions $y_1, y_2, y_3 \colon [0, \infty) \to \mathbb{R}$ satisfying:*

(i) $y_1(t)$ *satisfies*

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t), \quad \boldsymbol{x}(0+) = \boldsymbol{x}_0 \\ y_1(t) &= \boldsymbol{c}^t\boldsymbol{x}(t); \end{aligned}$$

(ii) $y_2(t)$ *satisfies*

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t), \quad \boldsymbol{x}(0+) = \boldsymbol{0} \\ y_1(t) &= \boldsymbol{c}^t\boldsymbol{x}(t); \end{aligned}$$

(iii) $y_3(t)$ *satisfies*

$$y_3(t) = \boldsymbol{D}u(t);$$

(iv) $y(t) = y_1(t) + y_2(t) + y_3(t)$.

Proof First note that $y_1(t)$, $y_2(t)$, and $y_3(t)$ are indeed uniquely defined by the conditions of (i), (ii), and (iii), respectively. It thus remains to show that (iv) is satisfied. However, this is a straightforward matter of checking that $y_1(t) + y_2(t) + y_3(t)$ as defined by the conditions (i), (ii), and (iii) satisfies (3.18). ■

The idea here is that to obtain the output for (3.18) we first look at the case where $\boldsymbol{D} = \boldsymbol{0}_1$ and $u(t) = 0$, obtaining the solution $y_1(t)$. To this we add the output $y_2(t)$, defined again with $\boldsymbol{D} = \boldsymbol{0}_1$, but this time with the input as the given input $u(t)$. Note that to obtain y_1 we use the given initial condition $\boldsymbol{x}_0$, but to obtain $y_2(t)$ we use the zero initial condition. Finally, to these we add $y_3(t) = \boldsymbol{D}u(t)$ to get the actual output.

Our objective in this section is to provide *scalar* initial value problems whose solutions are $y_1(t)$, $y_2(t)$, and $y_3(t)$. These are easily implemented numerically. We begin by determining the scalar initial value problem for $y_1(t)$.

3.35 Lemma *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system and let $y_1(t)$ be the output determined by condition (i) of Lemma 3.34. Then $y_1(t)$ solves the initial value problem*

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y_1(t) = 0, \quad y_1(0+) = \boldsymbol{c}^t\boldsymbol{x}_0,\ y_1^{(1)}(0+) = \boldsymbol{c}^t\boldsymbol{A}\boldsymbol{x}_0,\ \ldots,\ y_1^{(n-1)}(0+) = \boldsymbol{c}^t\boldsymbol{A}^{n-1}\boldsymbol{x}_0,$$

where (N, D) is the c.f.r. for T_Σ.

Proof Note that $y_1(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{x}_0$. Taking Laplace transforms and using Exercise EE.4 gives

$$\mathscr{L}^+_{0+}(y_1)(s) = \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{x}_0 = \frac{\boldsymbol{c}^t\operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{x}_0}{\det(s\boldsymbol{I}_n - \boldsymbol{A})}.$$

Thus we have

$$D(s)\mathscr{L}^+_{0+}(y)_1(s) = \boldsymbol{c}^t\operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{x}_0,$$

and since the right-hand side is a polynomial of degree at most $n-1$ (if $\deg(D) = n$), by (3.13) this means that there are some initial conditions for which $y_1(t)$ is a solution to $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y_1(t) = 0$. It remains to compute the initial conditions. However, this is a simple computation, giving exactly the conditions of the lemma. ■

Now we look at a scalar differential equation for $y_2(t)$.

3.36 Lemma *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system, let $u\colon [0,\infty) \to \mathbb{R}$ possess a right causal Laplace transform, and assume that $u, u^{(1)}, \dots, u^{(\deg(N)-1)}$ are continuous on $[0,\infty)$ and that $u^{(\deg(N))}$ is piecewise continuous on $[0,\infty)$. If $y_2(t)$ is the output determined by condition (ii) of Lemma 3.34, then*

$$y_2(t) = \int_0^t h_\Sigma(t-\tau)u(\tau)\,\mathrm{d}\tau,$$

and furthermore $y_2(t)$ solves the initial value problem

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y_2(t) = N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)u(t), \quad y_2(0+) = 0,\ y_2^{(1)}(0+) = \boldsymbol{c}^t\boldsymbol{b}u(0+),\ \dots,$$
$$y_2^{(k)}(0+) = \sum_{\substack{i,j\\ i+j=k-1}} \boldsymbol{c}^t\boldsymbol{A}^i\boldsymbol{b}u^{(j)}(0+),\ \dots,\ y_2^{(n-1)}(0+) = \sum_{\substack{i,j\\ i+j=n-2}} \boldsymbol{c}^t\boldsymbol{A}^i\boldsymbol{b}u^{(j)}(0+),$$

where (N, D) is the c.f.r. for T_Σ.

Proof That $y_2(t) = \int_0^t h_\Sigma(t-\tau)u(\tau)\,\mathrm{d}\tau$ is a simple consequence of the definition of $y_2(t)$ and Proposition 2.32. What's more, taking the Laplace transform of $y_2(t)$ we get

$$\mathscr{L}^+_{0+}(y_2)(s) = T_\Sigma(s)\mathscr{L}^+_{0+}(u)(s) \quad \Longrightarrow \quad D(s)\mathscr{L}^+_{0+}(y_2)(s) = N(s)\mathscr{L}^+_{0+}(u)(s),$$

using Theorem 3.22, Exercise EE.5, and the fact that (N, D) is the c.f.r. of T_Σ. This means that $y_2(t)$ is indeed a solution of the differential equation $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y_2(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$. To determine the initial conditions for $y_2(t)$, we simply differentiate the formula we have. Thus we immediately get $y_2(0+) = 0$. For the first derivative we have

$$\begin{aligned} y_2^{(1)}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_0^t h_\Sigma(t-\tau)u(\tau)\,\mathrm{d}\tau \\ &= h_\Sigma(t)u(t) + \int_0^t h_\Sigma^{(1)}(t-\tau)u(\tau)\,\mathrm{d}\tau. \end{aligned}$$

Thus $y^{(1)}(0+) = h_\Sigma(0+)u(0+)$. We may proceed, using mathematical induction if one wishes to do it properly, to derive

$$y_2^{(k)}(t) = \sum_{\substack{i,j\\ i+j=k-1}} h_\Sigma^{(i)}(0+)u^{(j)}(t) + \int_0^t h_\Sigma^{(k)}(t-\tau)u(\tau)\,\mathrm{d}\tau.$$

Now we observe that a simple computation given $h_\Sigma(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}$ demonstrates that

$$h_\Sigma^{(k)}(0+) = \boldsymbol{c}^t \boldsymbol{A}^k \boldsymbol{b},$$

and using this expression, the result follows. ■

The above two lemmas give us the "hard part" of the output for (3.18)—all that remains is to add $y_3(t) = \boldsymbol{D}u(t)$. Let us therefore summarise how to determine the output for the system (3.18) from a scalar differential equation.

3.37 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system, let $u\colon [0,\infty) \to \mathbb{R}$ possess a right causal Laplace transform, and assume that $u, u^{(1)}, \dots, u^{(\deg(N)-1)}$ are continuous on $[0,\infty)$ and that $u^{(\deg(N))}$ is piecewise continuous on $[0,\infty)$. If $y(t)$ is the output defined by (3.18) for the input $u(t)$ and initial state $\boldsymbol{x}_0$, then $y(t)$ is the solution of the differential equation*

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)\tilde{y}(t) = N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)u(t),$$

subject to the initial conditions

$$\begin{aligned}
y(0+) &= \boldsymbol{c}^t\boldsymbol{x}_0 + \boldsymbol{D}u(0+),\\
y^{(1)}(0+) &= \boldsymbol{c}^t\boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{c}^t\boldsymbol{b}u(0+) + \boldsymbol{D}u^{(1)}(0+),\\
&\vdots\\
y^{(k)}(0+) &= \boldsymbol{c}\boldsymbol{A}^k\boldsymbol{x}_0 + \sum_{\substack{i,j\\ i+j=k-1}} \boldsymbol{c}^t\boldsymbol{A}^i\boldsymbol{b}u^{(j)}(0+) + \boldsymbol{D}u^{(k)}(0+),\\
&\vdots\\
y^{(n-1)}(0+) &= \boldsymbol{c}\boldsymbol{A}^{n-1}\boldsymbol{x}_0 + \sum_{\substack{i,j\\ i+j=n-2}} \boldsymbol{c}^t\boldsymbol{A}^i\boldsymbol{b}u^{(j)}(0+) + \boldsymbol{D}u^{(n-1)}(0+).
\end{aligned}$$

Proof That y satisfies the differential equation $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$ follows since

$$\mathscr{L}^+_{0+}(y)(s) = \big(\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} + \boldsymbol{D}\big)\mathscr{L}^+_{0+}(u)(s),$$

modulo initial conditions. The initial conditions for y are derived as in Lemma 3.36, using the fact that we now have

$$y(t) = e^{\boldsymbol{A}t}\boldsymbol{x}_0 + \int_0^t h_{\tilde{\Sigma}}(t-\tau)u(\tau)\,\mathrm{d}\tau + \boldsymbol{D}u(t),$$

where $\tilde{\Sigma} = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$. ■

Now admittedly this is a lengthy result, but with any given example, it is simple enough to apply. Let us justify this by applying the result to an example.

3.38 Example **(Example 3.29 cont'd)** We take $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ with

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1\\ -2 & -2 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0\\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1\\ -1 \end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1.$$

Thus this is simply the state-space version of the system we dealt with in Examples 3.29 and 3.33. As input we take again $u(t) = t$, and as initial state we select $\boldsymbol{x}_0 = (\frac{4}{5}, \frac{1}{5})$. Following Proposition 3.37 we compute

$$\boldsymbol{c}^t\boldsymbol{x}_0 = 1, \quad \boldsymbol{c}^t\boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{c}^t\boldsymbol{b}u(0+) = 0,$$

and we compute the transfer function to be

$$T_\Sigma(s) = \frac{1-s}{s^2+2s+2}.$$

Therefore, the initial value problem given to us by Proposition 3.37 is

$$\ddot{y}(t) + 2\dot{y}(t) + 2y(t) = t - 1, \quad y(0+) = 1,\ \dot{y}(0+) = 0.$$

Well now, if this isn't the same initial value problem encountered in Examples 3.29 and 3.33! Of course, the initial state vector $\boldsymbol{x}_0$ was designed to accomplish this. In any case, we may solve this initial value problem in the manner of either of Examples 3.29 and 3.33, and you will recall that the answer is

$$y(t) = e^{-t}(\tfrac{3}{2}\sin t + 2\cos t) + \tfrac{t}{2} - 1. \qquad \bullet$$

3.6.5 Formulae for impulse, step, and ramp responses

In this section we focus on developing initial value problems for obtaining some of the basic outputs for SISO systems. We provide this for both SISO linear systems, and those in input/output form. Although the basic definitions are made in the Laplace transform domain, the essential goal of this section is to provide initial value problems in the time-domain, and in so doing provide the natural method for numerically obtaining the various responses. These responses can be obtained by various control packages available, but it is nice to know what they are doing, since they certainly are not performing inverse Laplace transforms!

3.39 Definition Let (N, D) be a SISO linear system in input/output form. The ***step response*** for (N, D) is the function $1_{N,D}(t)$ whose Laplace transform is $\frac{1}{s}T_{N,D}(s)$. The ***ramp response*** is the function $R_{N,D}(t)$ whose Laplace transform is $\frac{1}{s^2}T_{N,D}(s)$. •

Note that $\frac{1}{s}$ is the Laplace transform of the unit step $1(t)$ and that $\frac{1}{s^2}$ is the Laplace transform of the unit slope ramp input $u(t) = t1(t)$. Of course, we could define the response to an input $u(t) = t^k1(t)$ for $k \geq 2$ by noting that the Laplace transform of such an input is $\frac{k!}{s^{k+1}}$. Indeed, it is a simple matter to produce the general formulas following what we do in this section, but we shall not pursue this level of generality here.

We now wish to produce the scalar initial value problems whose solutions are the impulse, step, and ramp responses Fortunately, the hard work has been done already, and we essentially have but to state the answer.

3.40 Proposition *Let (N, D) be a proper SISO linear system in input/output form with*

$$D(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1s + p_0$$
$$N(s) = c_ns^n + c_{n-1}s^{n-1} + \cdots + c_1s + c_0,$$

and let $\Sigma_{N,D} = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be the canonical minimal realisation of (N, D). The following statements hold.

(i) *If (N, D) is strictly proper then the impulse response $h_{N,D}$ is the solution to the initial value problem*

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)h_{N,D}(t) = 0, \quad h_{N,D}(0+) = \boldsymbol{c}^t\boldsymbol{b},\ h^{(1)}_{N,D}(0+) = \boldsymbol{c}^t\boldsymbol{A}\boldsymbol{b},\ \dots,\ h^{(n-1)}_{N,D}(0+) = \boldsymbol{c}^t\boldsymbol{A}^{n-1}\boldsymbol{b}.$$

(ii) *The step response $1_{N,D}(t)$ is the solution to the initial value problem*

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)1_{N,D}(t) = c_0, \quad 1_{N,D}(0+) = \boldsymbol{D},\ 1^{(1)}_{N,D}(0+) = \boldsymbol{c}^t\boldsymbol{b},\ \dots,\ 1^{(n-1)}_{N,D}(0+) = \boldsymbol{c}^t\boldsymbol{A}^{n-2}\boldsymbol{b}.$$

(iii) *The ramp response $R_{N,D}(t)$ is the solution to the initial value problem*

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)R_{N,D}(t) = c_1 + c_0 t, \quad R_{N,D}(0+) = 0,\ R^{(1)}_{N,D}(0+) = \boldsymbol{D},\ \dots,$$
$$R^{(2)}_{N,D}(0+) = \boldsymbol{c}^t\boldsymbol{b},\ \dots,\ R^{(n-1)}_{N,D}(0+) = \boldsymbol{c}^t\boldsymbol{A}^{n-3}\boldsymbol{b}.$$

Proof Let us look first at the impulse response. Since the Laplace transform of $h_{N,D}$ is the transfer function $T_{N,D}$ we have

$$D(s)\mathscr{L}^+_{0+}(h_{N,D})(s) = N(s).$$

As we saw in the course of the proof of Proposition 3.27, the fact that the right-hand side of this equation is a polynomial in s of degree at most $n-1$ (if $\deg(D) = n$) implies that $h_{N,D}$ is a solution of the differential equation $D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)h_{N,D}(t) = 0$. The determination of the initial conditions is a simple matter of differentiating $h_{N,D}(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}$ the required number of times, and evaluating at $t = 0$.

The last two statements follow from Lemma 3.36 choosing $u(t) = 1(t)$ for the step response, and $u(t) = t1(t)$ for the ramp response. ■

The impulse response is generalised for proper systems in Exercise E3.1. One can, I expect, see how the proposition gets generalised for inputs like $u(t) = t^k 1(t)$. Let's now determine the impulse, step, and ramp response for an example so the reader can see how this is done.

3.41 Example We take $(N(s), D(s)) = (1 - s, s^2 + 2s + 2)$, the same example we have been using throughout this section. We shall merely produce the initial value problems that give the step and ramp response for this problem, and not bother with going through the details of obtaining the solutions to these initial value problems. The canonical minimal realisation is $\Sigma_{N,D} = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1.$$

We readily compute $\boldsymbol{c}^t\boldsymbol{b} = -1$ and $\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{b} = 3$.

For the impulse response, we solve the initial value problem

$$\ddot{h}_{N,D}(t) + 2\dot{h}_{N,D}(t) + 2h_{N,D}(t) = 0, \quad h_{N,D}(0+) = -1, \quad \dot{h}_{N,D}(0+) = 3.$$

We obtain the solution by looking for a solution of the form e^{st}, and so determine that s should be a root of $s^2 + 2s + 2 = 0$. Thus $s = -1 \pm i$, and so our homogeneous solution will be a linear combination of the two linearly independent solutions $y_1(t) = e^{-t}\cos t$ and $y_2(t) = e^{-t}\sin t$. That is,

$$h_{N,D}(t) = C_1 e^{-t}\cos t + C_2 e^{-t}\sin t.$$

To determine C_1 and C_2 we use the initial conditions. These give

$$h_{N,D}(0+) = C_1 = -1$$
$$\dot{h}_{N,D}(0+) = -C_1 + C_2 = 3.$$

Solving gives $C_1 = -1$ and $C_2 = 2$, and so we have

$$h_{N,D}(t) = e^{-t}(2\sin t + \cos t).$$

For the step response the initial value problem is

$$\ddot{1}_{N,D}(t) + 2\dot{1}_{N,D}(t) + 21_{N,D}(t) = 1, \quad 1_{N,D}(0+) = 0, \ \dot{1}_{N,D}(0+) = \boldsymbol{c}^t\boldsymbol{b} = -1.$$

In like manner, the initial value for the ramp response is

$$\ddot{R}_{N,D}(t) + 2\dot{R}_{N,D}(t) + 2R_{N,D}(t) = t - 1, \quad R_{N,D}(0+) = 0, \ \dot{R}_{N,D}(0+) = 0.$$

Doing the requisite calculations gives

$$1_{N,D}(t) = \tfrac{1}{2} - \tfrac{1}{2}e^{-t}(\cos t + 3\sin t), \quad R_{N,D}(t) = \tfrac{t}{2} - 1 + e^{-t}\big(\cos t + \tfrac{1}{2}\sin t\big). \qquad \bullet$$

You may wish to refer back to this section on computing step responses later in the course of reading this book, since the step response is something we shall frequently write down without saying much about how we did it.

3.42 Remark Those of you who are proponents of the Laplace transform will wonder why one does not simply obtain the impulse response or step response by obtaining the inverse Laplace transform of the transfer function or the transfer function multiplied by $\frac{1}{s}$. While this is theoretically possible, in practice it is not really a good alternative. Even numerically, determining the inverse Laplace transform is very difficult, even when it is possible.

To indicate this, let us consider a concrete example. We take $(N(s), D(s)) = (s^2 + 3s + 1, s^5 + 5s^4 + 10s^3 + 20s^2 + 10s + 5)$. The impulse response and step response were generated numerically and are shown in Figure 3.9. Both of these were generated in two ways: (1) by

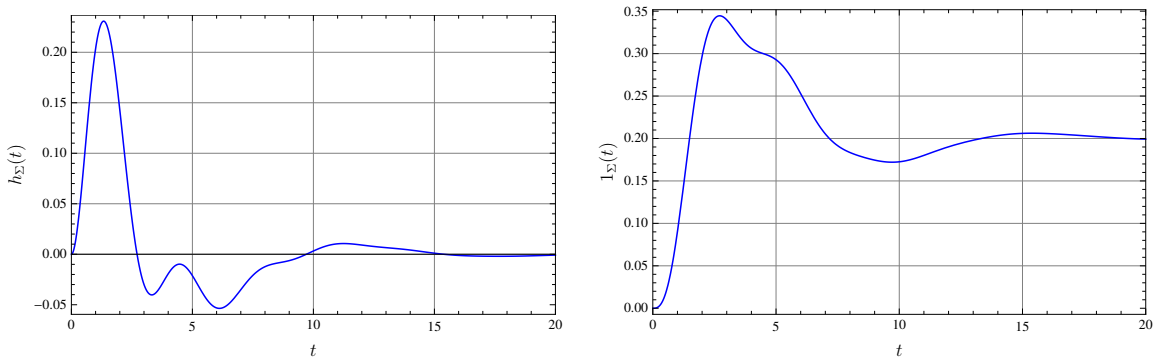


Figure 3.9 Impulse response $h_{N,D}(t)$ (left) and step response $1_{N,D}(t)$ (right) for $(N(s), D(s)) = (s^2+3s+1, s^5+5s^4+10s^3+20s^2+10s+5)$

computing numerically the inverse Laplace transform, and (2) by solving the ordinary differential equations of Proposition 3.40. In Table 3.1 can be seen a rough comparison of the time taken to do the various calculations. Obviously, the differential equation methods are far more efficient, particularly on the step response. Indeed, the inverse Laplace transform methods will sometimes not work, because they rely on the capacity to factor a polynomial. •

Table 3.1 Comparison of using inverse Laplace transform and ordinary differential equations to obtain impulse and step response

Computation	Time using inverse Laplace transform	Time taken using ode
Impulse response	14s	< 1s
Step response	5m15s	1s

3.7 Summary

This chapter is very important. A thorough understanding of what is going on here is essential and let us outline the salient facts you should assimilate before proceeding.

1. You should know basic things about polynomials, and in particular you should not hesitate at the mention of the word "coprime."
2. You need to be familiar with the concept of a rational function. In particular, the words "canonical fractional representative" (c.f.r.) will appear frequently later in this book.
3. You should be able to determine the partial fraction expansion of any rational function.
4. The definition of the Laplace transform is useful, and you ought to be able to apply it when necessary. You should be aware of the abscissa of absolute convergence since it can once in awhile come up and bite you.
5. The properties of the Laplace transform given in Proposition E.9 will see some use.
6. You should know that the inverse Laplace transform exists. You should recognise the value of Proposition E.11 since it will form the basis of parts of our stability investigation.
7. You should be able to perform block diagram algebra with the greatest of ease. We will introduce some powerful techniques in Section 6.1 that will make easier some aspects of this kind of manipulation.
8. Given a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ you should be able to write down its transfer function T_Σ.
9. You need to understand some of the features of the transfer function T_Σ; for example, you should be able to ascertain from it whether a system is observable, controllable, and whether it is minimum phase.
10. You *really really* need to know the difference between a "SISO linear system" and a "SISO linear system in input/output form." You should also know that there are relationships between these two different kinds of objects—thus you should know how to determine $\Sigma_{N,D}$ for a given strictly proper SISO linear system (N, D) in input/output form.
11. The connection between the impulse response and the transfer function is very important.
12. You ought to be able to determine the impulse response for a transfer function (N, D) in input/output form (use the partial fraction expansion).

Exercises

The impulse response $h_{N,D}$ for a strictly proper SISO linear system (N, D) in input/output form has the property that every solution of the differential equation $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$ can be written as

$$y(t) = y_h(t) + \int_0^t h_{N,D}(t-\tau)u(\tau)\,\mathrm{d}\tau$$

where $y_h(t)$ is a solution of $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y_h(t) = 0$ (see Proposition 3.32). In the next exercise, you will extend this to proper systems.

E3.1 Let (N, D) be a proper, but not necessarily strictly proper, SISO linear system in input/output form.

(a) Show that the transfer function $T_{N,D}$ for (N, D) can be written as

$$T_{N,D}(s) = T_{\tilde{N},\tilde{D}}(s) + C$$

for a uniquely defined strictly proper SISO linear system $(\tilde{N}, \tilde{D})$ in input/output form, and constant $C \in \mathbb{R}$. Explicitly determine $(\tilde{N}, \tilde{D})$ and C in terms of (N, D).

(b) Show that every solution of $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$ can be written as a linear combination of $u(t)$ and $\tilde{y}(t)$ where $\tilde{y}(t)$ is a solution of $\tilde{D}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\tilde{y}(t) = \tilde{N}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$.

(c) Conclude that the solution of $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$ can be written as

$$y(t) = y_h(t) + \int_0^t h_{\tilde{N},\tilde{D}}(t-\tau)u(\tau)\,\mathrm{d}\tau + C\int_0^t \delta(t-\tau)u(\tau)\,\mathrm{d}\tau$$

where $\delta(t)$ satisfies

$$\int_{-\infty}^{\infty} \delta(t-t_0)f(t)\,\mathrm{d}t = f(t_0)$$

for any integrable function f. Is $\delta(t)$ really a map from $\mathbb{R}$ to $\mathbb{R}$?

E3.2 Determine the transfer function from $\hat{r}$ to $\hat{y}$ for the block diagram depicted in Figure E3.1.

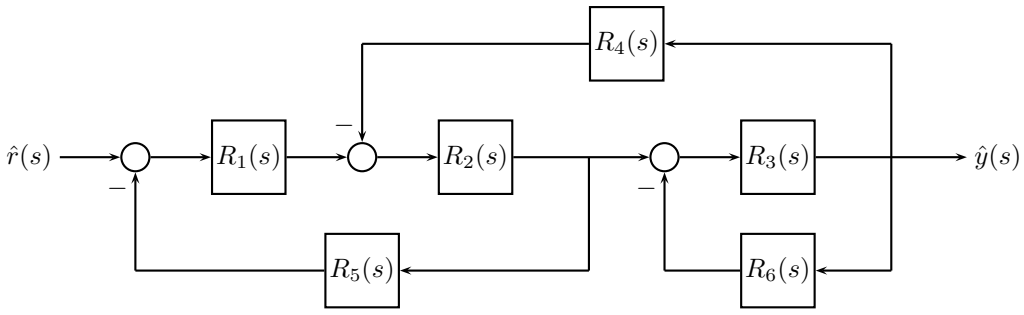


Figure E3.1 A block diagram with three loops

At various time throughout the text, we will want to have some properties of real rational functions as s becomes large. In the following simple exercise, you will show that this notion does not depend on the sense in which s is allowed to go to infinity.

E3.3 Let (N, D) be a proper SISO linear system in input/output form.

(a) Show that the limit

$$\lim_{R\to\infty} T_{N,D}(Re^{i\theta})$$

is real and independent of $\theta \in (-\pi, \pi]$. Thus it makes sense to write

$$\lim_{s\to\infty} T_{N,D}(s),$$

and this limit will be in $\mathbb{R}$.

(b) Show that the limit in part (a) is zero if (N, D) is strictly proper.

(c) If (N, D) is not strictly proper (but still proper), give an expression for the limit from part (a) in terms of the coefficients of D and N.

E3.4 For the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ with

$$\boldsymbol{A} = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

for $\sigma \in \mathbb{R}$ and $\omega > 0$, determine T_Σ.

E3.5 For the SISO linear systems connected in series in Exercise E2.1, determine the transfer function for the interconnected SISO linear system (i.e., for the $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$ determined in that exercise).

E3.6 For the SISO linear systems connected in parallel in Exercise E2.2, determine the transfer function for the interconnected SISO linear system (i.e., for the $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$ determined in that exercise).

E3.7 For the SISO linear systems connected in the negative feedback configuration of Exercise E2.3, we will determine the transfer function for the interconnected SISO linear system (i.e., for the $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$ determined in that exercise).

(a) Use Lemma A.2 to show that

$$\boldsymbol{c}^t(s\boldsymbol{I}_{n_1+n_2} - \boldsymbol{A})^{-1}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{0}^t & \boldsymbol{c}_2^t \end{bmatrix} \begin{bmatrix} s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1 & \boldsymbol{b}_1\boldsymbol{c}_2^t \\ -\boldsymbol{b}_2\boldsymbol{c}_1^t & s\boldsymbol{I}_{n_2} - \boldsymbol{A}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{c}_2^t\boldsymbol{U}\boldsymbol{b}_1,$$

where

$$\boldsymbol{U} = \big((s\boldsymbol{I}_{n_2} - \boldsymbol{A}_2) + \boldsymbol{b}_2\boldsymbol{c}_1^t(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1}\boldsymbol{b}_1\boldsymbol{c}_2^t\big)^{-1}\boldsymbol{b}_2\boldsymbol{c}_1^t(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1}.$$

(b) Use your answer from part (a), along with Lemma A.3, to show that

$$T_\Sigma(s) = (1 + T_{\Sigma_2}(s)T_{\Sigma_1}(s))^{-1}T_{\Sigma_2}(s)T_{\Sigma_1}(s).$$

(c) Thus we have

$$T_\Sigma(s) = \frac{T_{\Sigma_1}(s)T_{\Sigma_2}(s)}{1 + T_{\Sigma_1}(s)T_{\Sigma_2}(s)}.$$

Could you have deduced this otherwise?

(d) Use the computed transfer function to determine the characteristic polynomial for the system. (Thus answer Exercise E2.3(b).)

E3.8 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system with transfer function T_Σ. Show that T_Σ is proper, and is strictly proper if and only if $\boldsymbol{D} = \boldsymbol{0}_1$.

E3.9 Consider the two polynomials that are not coprime:

$$P_1(s) = s^2 + 1, \quad P_2(s) = s^3 + s^2 + s + 1.$$

Answer the following questions:

(a) Construct two SISO linear system $\Sigma_i = (\boldsymbol{A}_i, \boldsymbol{b}_i, \boldsymbol{c}_i^t, \boldsymbol{0}_1)$, $i = 1, 2$, each with state dimension 3, and with the following properties:
   1. Σ_1 is controllable but not observable;
   2. Σ_2 is observable but not controllable;
   3. $T_{\Sigma_i}(s) = \frac{P_1(s)}{P_2(s)}$, $i = 1, 2$.

   Do not take any of $\boldsymbol{A}_i, \boldsymbol{b}_i, \boldsymbol{c}_i$, $i = 1, 2$, to be zero.

(b) Is it possible to find a controllable and observable system Σ_3, with state dimension 3, for which $T_{\Sigma_3}(s) = \frac{P_1(s)}{P_2(s)}$? If so find such a system, and if not explain why not.

(c) Is it possible to find an uncontrollable and unobservable system Σ_4, with state dimension 3, for which $T_{\Sigma_3}(s) = \frac{P_1(s)}{P_2(s)}$? If so find such a system, and if not explain why not.
   Hint: *Use Theorem 2.41.*

E3.10 Consider a SISO linear system (N, D) in input/output form, and suppose that N has a real root $z > 0$ with multiplicity one. Thus, in particular, (N, D) is nonminimum phase. Let $u(t) = 1(t)\mathrm{e}^{zt}$.

(a) Show that there exists a unique set of initial conditions $y(0), y^{(1)}(0), \dots, y^{(n-1)}(0)$ for which the solution to the differential equation $D(\frac{\mathrm{d}}{\mathrm{d}t})y(t) = N(\frac{\mathrm{d}}{\mathrm{d}t})u(t)$, with these initial conditions, is identically zero.
   Hint: *Think about what Proposition 3.27 says about the method of Laplace transforms for solving ordinary differential equations.*

(b) Comment on this in light of the material in Section 2.3.3.

For the following three exercises we will consider "open-circuit" and "short-circuit" behaviour of circuits. Let us make the definitions necessary here in a general setting by talking about a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$. A pair of functions $(u(t), y(t))$ satisfying

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t) \end{aligned}$$

is ***open-circuit for Σ*** if $y(t) = 0$ for all t and is ***short-circuit for Σ*** if $u(t) = 0$ for all t.

E3.11 Consider the circuit of Exercise E1.7 with the output of Exercise E2.16.

(a) Compute the transfer function for the system.

(b) Determine all open-circuit pairs $(u(t), y(t))$.

(c) Determine all short-circuit pairs $(u(t), y(t))$.

(d) Comment on the open/short-circuit behaviour in terms of controllability, observability, and zero dynamics.

E3.12 Consider the circuit of Exercise E1.8 with the output of Exercise E2.17.

(a) Compute the transfer function for the system.

(b) Determine all open-circuit pairs $(u(t), y(t))$.

(c) Determine all short-circuit pairs $(u(t), y(t))$.

(d) Comment on the open/short-circuit behaviour in terms of controllability, observability, and zero dynamics.

E3.13 Consider the circuit of Exercise E1.9 with the output of Exercise E2.18.

(a) Compute the transfer function for the system.
(b) Determine all open-circuit pairs $(u(t), y(t))$.
(c) Determine all short-circuit pairs $(u(t), y(t))$.
(d) Comment on the open/short-circuit behaviour in terms of controllability, observability, and zero dynamics.

E3.14 For the coupled mass system of Exercises E1.4 and E2.19 (assume no damping), take as input the case of $\alpha = 0$ described in Exercise E2.19—thus the input is a force $u(t)$ applied to the leftmost mass. Determine an output that renders the system unobservable.

E3.15 Using Theorem 3.15, determine the spectrum of the zero dynamics for the pendulum/cart system of Exercises E1.5 and E2.4 for each of the following linearisations:

(a) the equilibrium point $(0, 0)$ with cart position as output;
(b) the equilibrium point $(0, 0)$ with cart velocity as output;
(c) the equilibrium point $(0, 0)$ with pendulum angle as output;
(d) the equilibrium point $(0, 0)$ with pendulum angular velocity as output;
(e) the equilibrium point $(0, \pi)$ with cart position as output;
(f) the equilibrium point $(0, \pi)$ with cart velocity as output;
(g) the equilibrium point $(0, \pi)$ with pendulum angle as output;
(h) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output.

E3.16 Consider the double pendulum of Exercises E1.6 and E2.5. In the following cases, use Theorem 3.15 to determine the spectrum of the zero dynamics:

(a) the equilibrium point $(0, 0, 0, 0)$ with the pendubot input;
(b) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;
(c) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;
(d) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;
(e) the equilibrium point $(0, 0, 0, 0)$ with the acrobot input;
(f) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;
(g) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;
(h) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input.

In each case, use the angle of the second link as output.

E3.17 Determine the spectrum of the zero dynamics for the linearised coupled tank system of Exercises E1.11 and E2.6 for the following outputs:

(a) the height in tank 1;
(b) the height in tank 2;
(c) the difference of the heights in the tanks.

E3.18 Given the SISO linear system (N, D) in input/output form with

$$D(s) = s^3 + 4s^2 + s + 1, \quad N(s) = 3s^2 + 1,$$

determine the canonical minimal realisation $\Sigma_{N,D}$. Is the triple $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})$ you found complete? Was your first impulse to answer the previous question by doing calculations? Explain why these are not necessary.

E3.19 State and prove a version of Theorem 3.20 that assigns to a SISO system (N, D) in input/output form a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ so that $\boldsymbol{A}$ and $\boldsymbol{c}$ are in observer canonical form. Also produce the block diagram analogous to Figure 3.8 in this case.

E3.20 Suppose you are handed a SISO linear system (N, D) in input/output form, and are told that it comes from a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$—that is, you are told that $T_\Sigma = T_{N,D}$. Is it possible for you to tell whether $(\boldsymbol{A}, \boldsymbol{b})$ is controllable?
Hint: *Consider Example 2.19 and Theorem 3.20.*

E3.21 For a SISO linear system (N, D) in input/output form, show that the transfer function $T_{N,D}$ has the property that $T_{N,D}(\bar{s}) = T_{N,D}(s)$ for all $s \in \mathbb{C}$. In particular, show that $s_0 \in \mathbb{C}$ is a zero or pole of $T_{N,D}$ if and only if $\bar{s}_0$ is a zero or pole, respectively.

E3.22 Verify Theorem 3.22 and Proposition 3.24 for Exercise E3.4 (recall that you had computed the impulse response for this problem in Exercise E2.26).

E3.23 For the following SISO linear systems in input/output form, use Proposition 3.32 to obtain the output corresponding to the given input and initial conditions.

(a) $(N(s), D(s)) = (1, s+3)$, $u(t) = 1(t)$ (the unit step input), and $y(0) = 1$.
(b) $(N(s), D(s)) = (1, s+3)$, $u(t) = 1(t)e^{at}$, $a \in \mathbb{R}$, and $y(0) = 0$.
(c) $(N(s), D(s)) = (s, s^3+s)$, $u(t) = 1(t)\cos t$, and $y(0) = 1$, $\dot{y}(0) = 0$, and $\ddot{y}(0) = 0$.
(d) $(N(s), D(s)) = (1, s^2)$, $u(t) = 1(t)$, and $y(0) = 0$ and $\dot{y}(0) = 1$.

E3.24 For the SISO linear systems in input/output form from Exercise E3.23 for which you obtained the solution, apply Proposition 3.27 to obtain the same solution.

E3.25 For the following SISO systems in input/output form, use Proposition 3.40 to setup the initial value problem for the step response, and use a computer package to plot the step response.

(a) $(N(s), D(s)) = (s+1, s^2+s+1)$.
(b) $(N(s), D(s)) = (s^2+2s+1, s^3+3s+1)$.
(c) $(N(s), D(s)) = (s-1, s^4+15s^3+20s^2+10s+2)$.
(d) $(N(s), D(s)) = (s^3+1, s^5+9s^4+20s^3+40s^2+50s+25)$.

E3.26 Consider the differential equation

$$\ddot{y}(t) + 4\dot{y}(t) + 8y(t) = 2\dot{u}(t) + 3u(t), \tag{E3.1}$$

where $u(t)$ is a specified function of time. If we define

$$u_\epsilon(t) = \begin{cases} \frac{1}{\epsilon}, & t \in [0, \epsilon] \\ 0, & \text{otherwise,} \end{cases}$$

and let $y_\epsilon(t)$ be the solution to the differential equation (E3.1) when $u = u_\epsilon$, determine $\lim_{\epsilon \to 0} y_\epsilon(t)$. (Note that in this problem although u_ϵ is in $\mathscr{U}$, $\dot{u}_\epsilon$ is *not* in $\mathscr{U}$. Thus to compute "directly" the solution to the differential equation (E3.1) with $u = u_\epsilon$ is not actually something you know how to do!)

E3.27 Consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ given by

$$\boldsymbol{A} = \begin{bmatrix} 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 0 \end{bmatrix},$$

and let $u(t) = 1(t)e^{t^2}$.

(a) Show that if $x(0) = 0$ then the output for the input u is $y(t) = \frac{\sqrt{\pi}}{2i}\mathrm{erf}(it)$, where erf is the ***error function*** given by

$$\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-\tau^2} \, \mathrm{d}\tau.$$

(b) Laplace transform techniques are always limited by one's ability to compute the inverse transform. Are there limitations in this example beyond the difficulty in determining the inverse transform?

Chapter 4

Frequency response (the frequency domain)

The final method we will describe for representing linear systems is the so-called "frequency domain." In this domain we measure how the system responds in the steady-state to sinusoidal inputs. This is often a good way to obtain information about how your system will handle inputs of various types.

The frequency response that we study in this section contains a wealth of information, often in somewhat subtle ways. This material, that builds on the transfer function discussed in Chapter 3, is fundamental to what we do in this course.

Contents

4.1 The frequency response of SISO linear systems

We first look at the state-space representation:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t). \end{aligned} \tag{4.1}$$

Let us first just come right out and define the frequency response, and then we can give its interpretation. For a SISO linear control system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{D})$ we let $\Omega_\Sigma \subset \mathbb{R}$ be defined by

$$\Omega_\Sigma = \{\omega \in \mathbb{R} \mid \ i\omega \text{ is a pole of } T_\Sigma\}.$$

The ***frequency response*** for Σ is the function $H_\Sigma \colon \mathbb{R}\backslash\Omega_\Sigma \to \mathbb{C}$ defined by $H_\Sigma(\omega) = T_\Sigma(i\omega)$. Note that we *do* wish to think of the frequency response as a $\mathbb{C}$-valued function, and not a rational function, because we will want to graph it. Thus when we write $T_\Sigma(i\omega)$, we intend

to evaluate the transfer function at $s = i\omega$. In order to do this, we suppose that all poles and zeroes of T_Σ have been cancelled.

The following result gives a key interpretation of the frequency response.

4.1 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$ be an complete SISO control system and let $\omega > 0$. If T_Σ has no poles on the imaginary axis that integrally divide ω,[1] then, given $u(t) = u_0 \sin \omega t$ there is a unique periodic output $y_p(t)$ with period $T = \frac{2\pi}{\omega}$ satisfying (4.1) and it is given by*

$$y(t) = u_0 \mathrm{Re}(H_\Sigma(\omega)) \sin \omega t + u_0 \mathrm{Im}(H_\Sigma(\omega)) \cos \omega t.$$

Proof We first look at the state behaviour of the system. Since $(\boldsymbol{A}, \boldsymbol{c})$ is complete, the numerator and denominator polynomials of T_Σ are coprime. Thus the poles of T_Σ are exactly the eigenvalues of $\boldsymbol{A}$. If there are no such eigenvalues that integrally divide ω, this means that there are no periodic solutions of period T for $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$. Therefore the linear equation $e^{\boldsymbol{A}T}\boldsymbol{u} = \boldsymbol{u}$ has only the trivial solution $\boldsymbol{u} = \mathbf{0}$. This means that the matrix $e^{\boldsymbol{A}T} - \boldsymbol{I}_n$ is invertible. We define

$$\boldsymbol{x}_p(t) = u_0 e^{\boldsymbol{A}t}(e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1} \int_0^T e^{-\boldsymbol{A}\tau}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau + u_0 \int_0^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau, \tag{4.2}$$

and we claim that $\boldsymbol{x}(t)$ is a solution to the first of equations (4.1) and is periodic with period $\frac{2\pi}{\omega}$. If $T = \frac{2\pi}{\omega}$ we first note that

$$\begin{aligned}
\boldsymbol{x}_p(t+T) &= u_0 e^{\boldsymbol{A}(t+T)}(e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1} \int_0^T e^{-\boldsymbol{A}\tau}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau + \\
&\quad u_0 \int_0^{t+T} e^{\boldsymbol{A}(t+T-\tau)}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau \\
&= e^{\boldsymbol{A}t}e^{\boldsymbol{A}T}(e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1} \int_0^T e^{-\boldsymbol{A}\tau}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau + \\
&\quad u_0 e^{\boldsymbol{A}t}e^{\boldsymbol{A}T} \int_0^T e^{-\boldsymbol{A}\tau}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau + u_0 \int_T^{t+T} e^{\boldsymbol{A}(t+T-\tau)}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau \\
&= e^{\boldsymbol{A}t}(e^{\boldsymbol{A}T}(e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1} + e^{\boldsymbol{A}T}) \int_0^T e^{-\boldsymbol{A}\tau}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau + \\
&\quad u_0 \int_0^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau.
\end{aligned}$$

Periodicity of $\boldsymbol{x}_p(t)$ will follow then if we can show that

$$e^{\boldsymbol{A}T}(e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1} + e^{\boldsymbol{A}T} = (e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1}.$$

But we compute

$$\begin{aligned}
e^{\boldsymbol{A}T}(e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1} + e^{\boldsymbol{A}T} &= (e^{\boldsymbol{A}T} + e^{\boldsymbol{A}T}(e^{-\boldsymbol{A}T} - \boldsymbol{I}_n))(e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1} \\
&= (e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1}.
\end{aligned}$$

Thus $\boldsymbol{x}_p(t)$ has period T. That $\boldsymbol{x}_p(t)$ is a solution to (4.1) with the $u(t) = u_0 \sin \omega t$ follows since $\boldsymbol{x}_p(t)$ is of the form

$$\boldsymbol{x}(t) = e^{\boldsymbol{A}t}\boldsymbol{x}_0 + u_0 \int_0^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b} \sin \omega\tau \, \mathrm{d}\tau$$

[1] Thus there are no poles for T_Σ of the form $i\tilde{\omega}$ where $\frac{\omega}{\tilde{\omega}} \in \mathbb{Z}$.

provided we take

$$\boldsymbol{x}_0 = (e^{-\boldsymbol{A}T} - \boldsymbol{I}_n)^{-1} \int_0^T e^{-\boldsymbol{A}\tau}\boldsymbol{b}\sin\omega\tau\,\mathrm{d}\tau.$$

For uniqueness, suppose that $\boldsymbol{x}(t)$ is a periodic solution of period T. Since it is a solution it must satisfy

$$\boldsymbol{x}(t) = e^{\boldsymbol{A}t}\boldsymbol{x}_0 + u_0 \int_0^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}\sin\omega\tau\,\mathrm{d}\tau$$

for some $\boldsymbol{x}_0 \in \mathbb{R}^n$. If $\boldsymbol{x}(t)$ has period T then we must have

$$\begin{aligned} e^{\boldsymbol{A}t}\boldsymbol{x}_0 + u_0 \int_0^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}\sin\omega\tau\,\mathrm{d}\tau &= e^{\boldsymbol{A}t}e^{\boldsymbol{A}T}\boldsymbol{x}_0 + u_0 e^{\boldsymbol{A}t}e^{\boldsymbol{A}T}\int_0^T e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}\sin\omega\tau\,\mathrm{d}\tau + \\ &\qquad u_0 \int_T^{t+T} e^{\boldsymbol{A}(t+T-\tau)}\boldsymbol{b}\sin\omega\tau\,\mathrm{d}\tau \\ &= e^{\boldsymbol{A}t}e^{\boldsymbol{A}T}\boldsymbol{x}_0 + u_0 e^{\boldsymbol{A}t}e^{\boldsymbol{A}T}\int_0^T e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}\sin\omega\tau\,\mathrm{d}\tau + \\ &\qquad u_0 \int_0^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}\sin\omega\tau\,\mathrm{d}\tau. \end{aligned}$$

But this implies that we must have

$$\boldsymbol{x}_0 = e^{\boldsymbol{A}T}\boldsymbol{x}_0 + u_0 e^{\boldsymbol{A}T}\int_0^T e^{-\boldsymbol{A}\tau}\boldsymbol{b}\sin\omega\tau\,\mathrm{d}\tau,$$

which means that $\boldsymbol{x}(t) = \boldsymbol{x}_p(t)$.

This shows that there is a unique periodic solution in state space. This clearly implies a unique output periodic output $y_p(t)$ of period T. It remains to show that $y_p(t)$ has the asserted form. We will start by giving a different representation of $\boldsymbol{x}_p(t)$ than that given in (4.2). We look for constant vectors $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ with the property that

$$\boldsymbol{x}_p(t) = \boldsymbol{x}_1 \sin\omega t + \boldsymbol{x}_2 \cos\omega t.$$

Substitution into (4.1) with $u(t) = u_0 \sin\omega t$ gives

$$\begin{aligned} & \omega\boldsymbol{x}_1\cos\omega t - \omega\boldsymbol{x}_2\sin\omega_t = \boldsymbol{A}\boldsymbol{x}_1\sin\omega t + \boldsymbol{A}\boldsymbol{x}_2\cos\omega t + u_0\boldsymbol{b}\sin\omega t \\ \Longrightarrow \quad & \omega\boldsymbol{x}_1 - \boldsymbol{A}\boldsymbol{x}_2, \quad -\boldsymbol{A}\boldsymbol{x}_1 - \omega\boldsymbol{x}_2 = u_0\boldsymbol{b} \\ \Longrightarrow \quad & i\omega(\boldsymbol{x}_1 + i\boldsymbol{x}_2) - \boldsymbol{A}(\boldsymbol{x}_1 + i\boldsymbol{x}_2) = u_0\boldsymbol{b}. \end{aligned}$$

Since $i\omega$ is not an eigenvalue for $\boldsymbol{A}$ we have

$$\begin{aligned} & \boldsymbol{x}_1 + i\boldsymbol{x}_2 = (i\omega\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} \\ \Longrightarrow \quad & \boldsymbol{c}^t\boldsymbol{x}_1 = \mathrm{Re}(H_\Sigma(\omega)), \quad \boldsymbol{c}^t\boldsymbol{x}_2 = \mathrm{Im}(H_\Sigma(\omega)). \end{aligned}$$

The result follows since $y_p(t) = \boldsymbol{c}^t\boldsymbol{x}_p(t)$. ■

4.2 Remarks

1. It turns out that *any* output from (4.1) with $u(t) = u_0 \sin \omega t$ can be written as a sum of the periodic output $y_p(t)$ with a function $y_h(t)$ where $y_h(t)$ can be obtained with zero input. This is, of course, reminiscent of the procedure in differential equations where you find a homogeneous and particular solution.
2. If the eigenvalues of $\boldsymbol{A}$ all lie in the negative half-plane, then it is easy to see that $\lim_{t\to\infty}|y_h(t)| = 0$ and so after a long enough time, we will essentially be left with the periodic solution $y_p(t)$. For this reason, one calls $y_p(t)$ the ***steady-state response*** and $y_h(t)$ the ***transient response***. Note that the steady-state response is uniquely defined (under the hypotheses of Theorem 4.1), but that there is no unique transient response—it depends upon the initial conditions for the state vector.
3. One can generalise this slightly to allow for imaginary eigenvalues $i\tilde{\omega}$ of $\boldsymbol{A}$ for which $\tilde{\omega}$ integrally divide ω, provided that $\boldsymbol{b}$ does not lie in the eigenspace of these eigenvalues. •

4.2 The frequency response for systems in input/output form

The matter of defining the frequency response for a SISO linear system in input/output form is now obvious, I hope. Indeed, if (N, D) is a SISO linear system in input/output form, then we define its ***frequency response*** by $H_{N,D}(\omega) = T_{N,D}(i\omega)$.

Let us see how one may recover the transfer function from the frequency response. Note that it is not obvious that one should be able to do this. After all, the frequency response function only gives us data on the imaginary axis. However, because the transfer function is analytic, if we know its value on the imaginary axis (as is the case when we know the frequency response), we may assert its value off the imaginary axis. To be perfectly precise on these matters requires some effort, but we can sketch how things go.

The first thing we do is indicate a direct correspondence between the frequency response and the impulse response. For this we refer to Section E.2 for a definition of the Fourier transform. With the notion of the Fourier transform in hand, we establish the correspondence between the frequency response and the impulse response as follows.

4.3 Proposition *Let (N, D) be a strictly proper SISO linear control system in input/output form, and suppose that the poles of $T_{N,D}$ are in the negative half-plane. Then $H_{N,D}(\omega) = \check{h}_{N,D}(\omega)$.*

Proof We have $H_{N,D}(\omega) = T_{N,D}(i\omega)$ and so

$$H_{N,D}(\omega) = \int_{0+}^{\infty} h_{N,D}(t)e^{-i\omega t}\,\mathrm{d}t.$$

By Exercise EE.2, $\sigma_{\min}(h_{N,D}) < 0$ since we are assuming all poles are in the negative half-plane. Therefore this integral exists. Furthermore, since $h_{N,D}(t) = 0$ for $t < 0$ we have

$$H_{N,D}(\omega) = \int_{-\infty}^{\infty} h_{N,D}(t)e^{-i\omega t}\,\mathrm{d}t = \check{h}_{N,D}(\omega).$$

This completes the proof. ■

Now we recover the transfer function $T_{N,D}$ from the frequency response $H_{N,D}$. In the following result we are thinking of the transfer function not as a rational function, but as a $\mathbb{C}$-valued function.

4.4 Proposition *Let (N, D) be a strictly proper SISO linear control system in input/output form, and suppose that the poles of $T_{N,D}$ are in the negative half-plane. Then, provided* $\mathrm{Re}(s) > \sigma_{\min}(h_{N,D})$*, we have*

$$T_{N,D}(s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{H_{N,D}(\omega)}{s - i\omega}\,\mathrm{d}\omega.$$

Proof By Proposition 4.3 we know that $h_{N,D}$ is the inverse Fourier transform of $H_{N,D}$:

$$h_{N,D}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H_{N,D}(\omega)e^{i\omega t}\,\mathrm{d}\omega.$$

On the other hand, by Theorem 3.22 the transfer function $T_{N,D}$ is the Laplace transform of $h_{N,D}$ so we have, for $\mathrm{Re}(s) > \sigma_{\min}(h_{N,D})$.

$$\begin{aligned} T_{N,D}(s) &= \int_{0+}^{\infty} h_{N,D}(t)e^{-st}\,\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{0+}^{\infty} H_{N,D}(\omega)e^{i\omega t}e^{-st}\,\mathrm{d}t\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{H_{N,D}(\omega)}{s - i\omega}\,\mathrm{d}\omega. \end{aligned}$$

This completes the proof. ■

4.5 Remarks

1. I hope you can see the importance of the results in this section. What we have done is establish the perfect correspondence between the three domains in which we work: (1) the time-domain, (2) the s-plane, and (3) the frequency domain. In each domain, one object captures the essence of the system behaviour: (1) the impulse response, (2) the transfer function, and (3) the frequency response. The relationships are summarised in Figure 4.1. Note that anything you say about one of the three objects in Figure 4.1 must

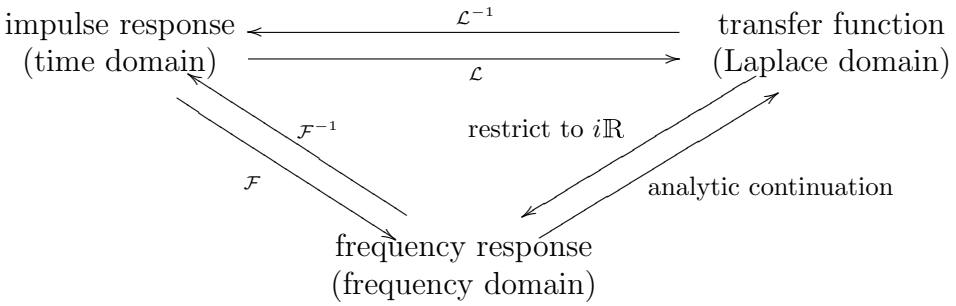


Figure 4.1 The connection between impulse response, transfer function, and frequency response

be somehow reflected in the others. We will see that this is true, and will form the centre of discussion of much of the rest of the course.

2. Of course, the results in this section may be made to apply to SISO linear systems in the form (4.1) provided that $\boldsymbol{D} = \mathbf{0}_1$ and that the polynomials $\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ and $P_{\boldsymbol{A}}(s)$ are coprime. •

4.3 Graphical representations of the frequency response

One of the reasons why frequency response is a powerful tool is that it is possible to succinctly understand its primary features by means of plotting functions or parameterised curves. In this section we see how this is done. Some of this may seem a bit pointless at present. However, as matters develop, and we get closer to design methodologies, the power of these graphical representations will become clear. The first obvious application of these ideas that we will encounter is the Nyquist criterion for stability in Chapter 7.

4.3.1 The Bode plot

What one normally does with the frequency response is plot it. But one plots it in a very particular manner. First write H_Σ in polar form:

$$H_\Sigma(\omega) = |H_\Sigma(\omega)|e^{i\angle H_\Sigma(\omega)}$$

where $|H_\Sigma(\omega)|$ is the absolute value of the complex number $H_\Sigma(\omega)$ and $\angle H_\Sigma(\omega)$ is the argument of the complex number $H_\Sigma(\omega)$. We take $-180^\circ < \angle H_\Sigma(\omega) \leq 180^\circ$. One then constructs two plots, one of $20\log|H_\Sigma(\omega)|$ as a function of $\log\omega$, and the other of $\angle H_\Sigma(\omega)$ as a function of $\log\omega$. (All logarithms we talk about here are base 10.) Together these two plots comprise the ***Bode plot*** for the frequency response H_Σ. The units of the plot of $20\log|H_\Sigma(\omega)|$ are ***decibels***.[2] One might think we are losing information here by plotting the magnitude and phase for positive values of ω (which we are restricted to doing by using $\log\omega$ as the independent variable). However, as we shall see in Proposition 4.13, we do not lose any information since the magnitude is symmetric about $\omega = 0$, and the phase is anti-symmetric about $\omega = 0$.

Let's look at the Bode plots for our mass-spring-damper system. I used Mathematica® to generate all Bode plots in this book. We will also be touching on a method for roughly determining Bode plots "by hand."

4.6 Examples In all cases we have

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{d}{m} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}.$$

We take $m = 1$ and consider the various cases of d and k as employed in Example 2.33. Here we can also consider the case when $\boldsymbol{D} \neq \boldsymbol{0}_1$.

1. We take $d = 3$ and $k = 2$.
   (a) With $\boldsymbol{c} = (1, 0)$ and $\boldsymbol{D} = \boldsymbol{0}_1$ we compute

   $$H_\Sigma(\omega) = \frac{1}{-\omega^2 + 3i\omega + 2}.$$

   The corresponding Bode plot is the first plot in Figure 4.2.
   (b) Next, with $\boldsymbol{c} = (0, 1)$ and $\boldsymbol{D} = \boldsymbol{0}_1$ we compute

   $$H_\Sigma(\omega) = \frac{i\omega}{-\omega^2 + 3i\omega + 2},$$

   and the corresponding Bode plot is the second plot in Figure 4.2.

[2] Decibels are so named after Alexander Graham Bell. The unit of "bell" was initially proposed, but when it was found too coarse a unit, the decibel was proposed.

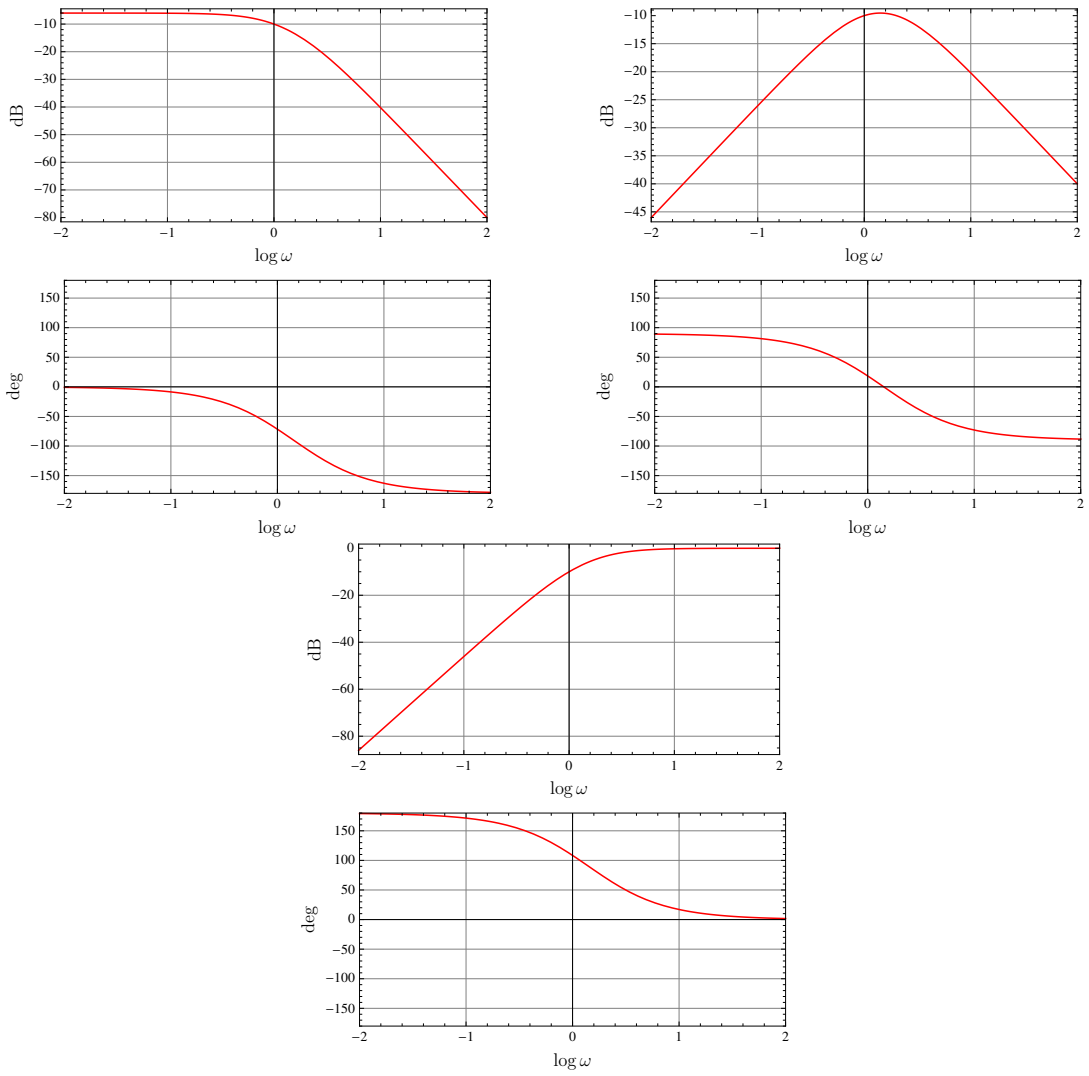


Figure 4.2 The displacement (top left), velocity (top right), and acceleration (bottom) frequency response for the mass-spring damper system when $d = 3$ and $k = 2$

(c) If we have $\boldsymbol{c} = (-\frac{k}{m}, -\frac{d}{m})$ and $\boldsymbol{D} = [1]$ we compute

$$H_\Sigma(\omega) = \frac{-\omega^2}{-\omega^2 + 3i\omega + 2},$$

The Bode plot for this frequency response function is the third plot in Figure 4.2.

2. We take $d = 2$ and $k = 1$.

(a) With $\boldsymbol{c} = (1, 0)$ and $\boldsymbol{D} = \boldsymbol{0}_1$ we compute

$$H_\Sigma(\omega) = \frac{1}{-\omega^2 + 2i\omega + 1}.$$

The corresponding Bode plot is the first plot in Figure 4.3.

(b) Next, with $\boldsymbol{c} = (0, 1)$ and $\boldsymbol{D} = \boldsymbol{0}_1$ we compute

$$H_\Sigma(\omega) = \frac{i\omega}{-\omega^2 + 2i\omega + 1},$$

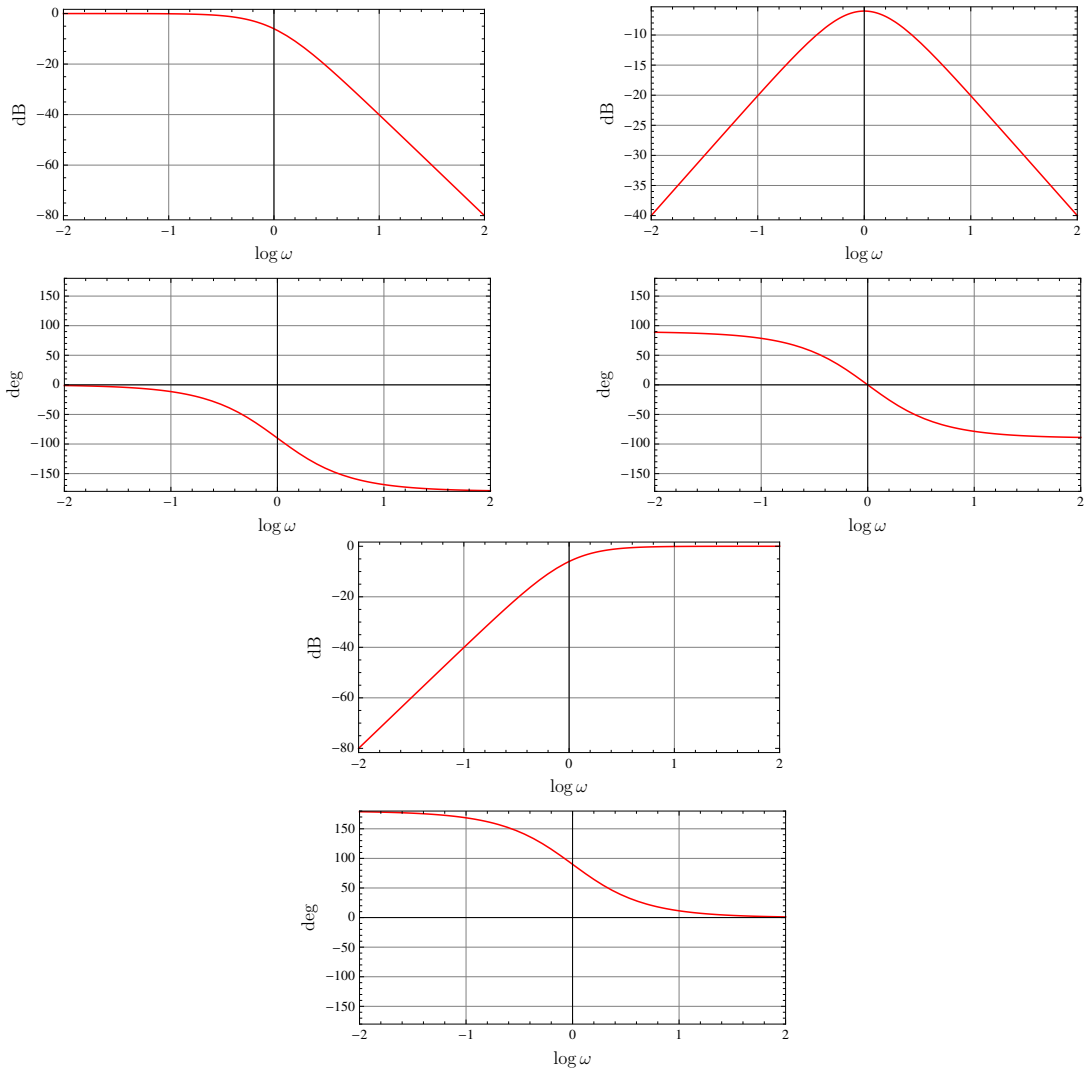


Figure 4.3 The displacement (top left), velocity (top right), and acceleration (bottom) frequency response for the mass-spring damper system when $d = 2$ and $k = 1$

and the corresponding Bode plot is the second plot in Figure 4.3.

(c) If we have $\boldsymbol{c} = (-\frac{k}{m}, -\frac{d}{m})$ and $\boldsymbol{D} = [1]$ we compute

$$H_\Sigma(\omega) = \frac{-\omega^2}{-\omega^2 + 2i\omega + 1},$$

The Bode plot for this frequency response function is the third plot in Figure 4.3.

3. We take $d = 2$ and $k = 10$.

(a) With $\boldsymbol{c} = (1, 0)$ and $\boldsymbol{D} = \mathbf{0}_1$ we compute

$$H_\Sigma(\omega) = \frac{1}{-\omega^2 + 2i\omega + 10}.$$

The corresponding Bode plot is the first plot in Figure 4.4.

(b) Next, with $\boldsymbol{c} = (0, 1)$ and $\boldsymbol{D} = \mathbf{0}_1$ we compute

$$H_\Sigma(\omega) = \frac{i\omega}{-\omega^2 + 2i\omega + 10},$$

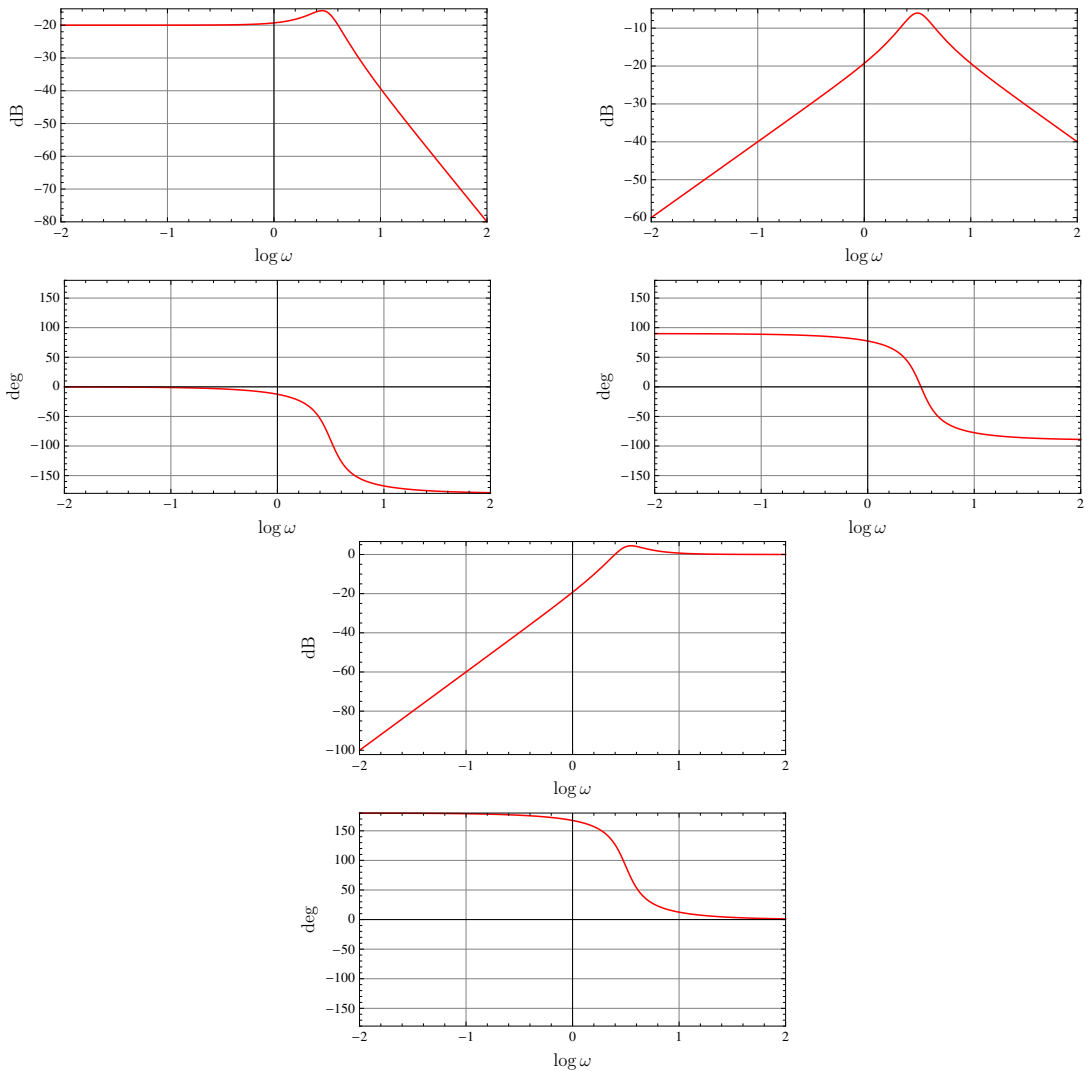


Figure 4.4 The displacement (top left), velocity (top right), and acceleration (bottom) frequency response for the mass-spring damper system when $d = 2$ and $k = 10$

and the corresponding Bode plot is the second plot in Figure 4.4.

(c) If we have $\boldsymbol{c} = (-\frac{k}{m}, -\frac{d}{m})$ and $\boldsymbol{D} = [1]$ we compute

$$H_\Sigma(\omega) = \frac{-\omega^2}{-\omega^2 + 2i\omega + 10},$$

The Bode plot for this frequency response function is the third plot in Figure 4.4.

4. We take $d = 0$ and $k = 1$.

(a) With $\boldsymbol{c} = (1, 0)$ and $\boldsymbol{D} = \mathbf{0}_1$ we compute

$$H_\Sigma(\omega) = \frac{1}{-\omega^2 + 1}.$$

The corresponding Bode plot is the first plot in Figure 4.5.

(b) Next, with $\boldsymbol{c} = (0, 1)$ and $\boldsymbol{D} = \mathbf{0}_1$ we compute

$$H_\Sigma(\omega) = \frac{i\omega}{-\omega^2 + 1},$$

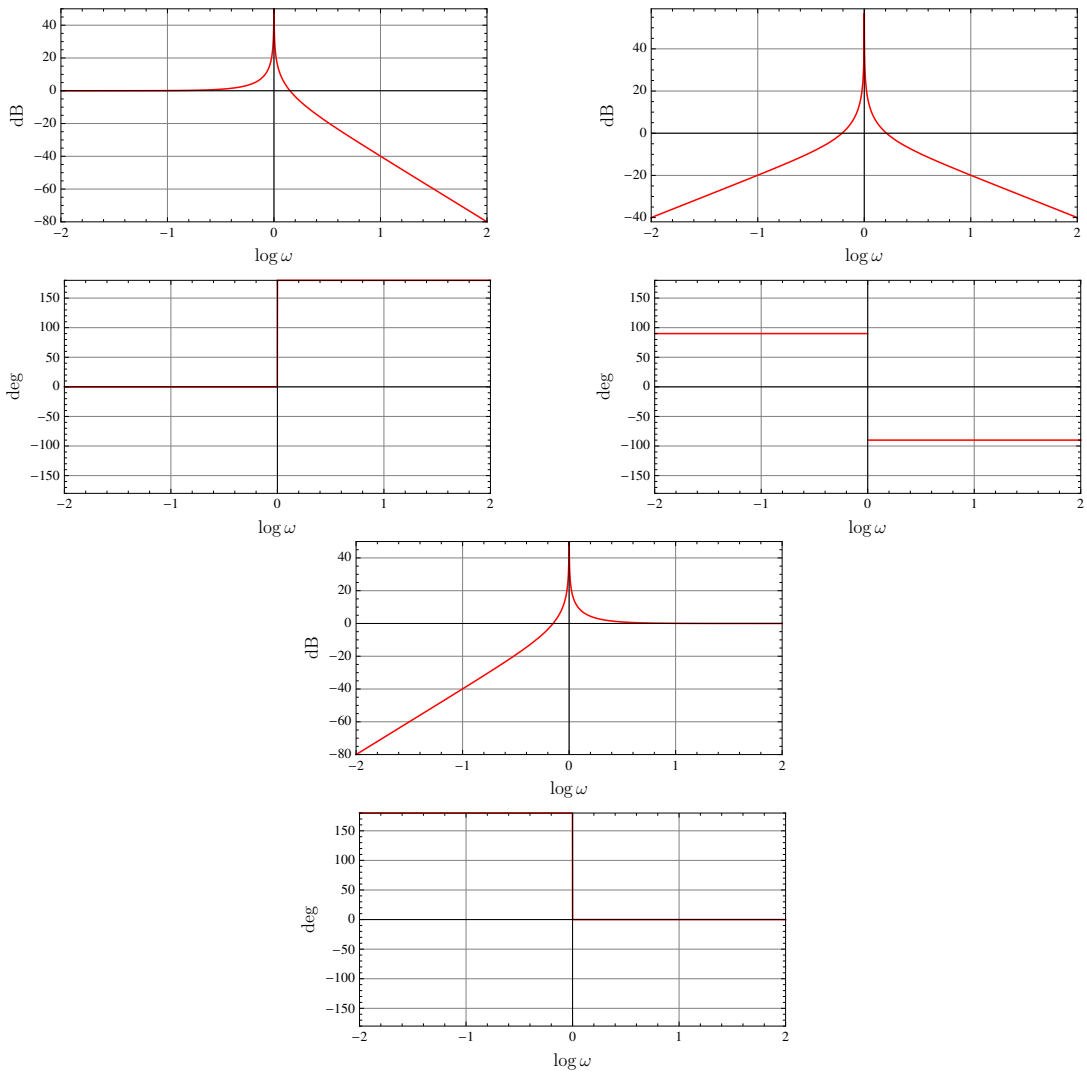


Figure 4.5 The displacement (top left), velocity (top right), and acceleration (bottom) frequency response for the mass-spring damper system when $d = 0$ and $k = 1$

and the corresponding Bode plot is the second plot in Figure 4.5.

(c) If we have $\boldsymbol{c} = (-\frac{k}{m}, -\frac{d}{m})$ and $\boldsymbol{D} = [1]$ we compute

$$H_\Sigma(\omega) = \frac{-\omega^2}{-\omega^2 + 1},$$

The Bode plot for this frequency response function is the third plot in Figure 4.5. •

4.3.2 A quick and dirty plotting method for Bode plots

It is possible, with varying levels of difficulty, to plot Bode plots by hand. The first thing we do is rearrange the frequency response in a particular way suitable to our purposes. The

form desired is

$$H(\omega) = \frac{K\displaystyle\prod_{j_1=1}^{k_1}(1+i\omega\tau_{j_1})\prod_{j_2=1}^{k_2}\bigl(1+2i\zeta_{j_2}\tfrac{\omega}{\omega_{j_2}}-(\tfrac{\omega}{\omega_{j_2}})^2\bigr)}{(i\omega)^{k_3}\displaystyle\prod_{j_4=1}^{k_4}(1+i\omega\tau_{j_4})\prod_{j_5=1}^{k_5}\bigl(1+2i\zeta_{j_5}\tfrac{\omega}{\omega_{j_5}}-(\tfrac{\omega}{\omega_{j_5}})^2\bigr)} \tag{4.3}$$

where the τ's, ω's, and ζ's are real, the ζ's are all further between -1 and 1, and the ω's are all positive. The frequency response for any stable, minimum phase system can always be put in this form. For nonminimum phase systems, or for unstable systems, the variations to what we describe here are straightforward. The form given reflects our transfer function having

1. k_1 real zeros at the points $-\frac{1}{\tau_{j_1}}$, $j_1 = 1, \dots, k_1$,
2. k_2 pairs of complex zeros at $\omega_{j_2}(-\zeta_{j_2} \pm \sqrt{1-\zeta_{j_2}^2})$, $j_2 = 1, \dots, k_2$,
3. k_3 poles at the origin,
4. k_4 real poles at the points $-\frac{1}{\tau_{j_4}}$, $j_4 = 1, \dots, k_4$, and
5. k_5 pairs of complex poles at $\omega_{j_5}(-\zeta_{j_5} \pm \sqrt{1-\zeta_{j_5}^2})$, $j_5 = 1, \dots, k_5$.

Although we exclude the possibility of having zeros or poles on the imaginary axis, one can see how to handle such functions by allowing ζ to become zero in one of the order two terms.

Let us see how to perform this in practice.

4.7 Example We consider the transfer function

$$T(s) = \frac{s+\frac{1}{10}}{s(s^2+4s+8)}.$$

To put this in the desired form we write

$$\begin{aligned} s+\tfrac{1}{10} &= \tfrac{1}{10}\bigl(1+10s\bigr)\\ s^2+4s+8 &= 8\bigl(1+\tfrac{1}{2}s+\tfrac{1}{8}s^2\bigr)\\ &= 8\bigl(1+2\tfrac{1}{4}\sqrt{8}\tfrac{s}{\sqrt{8}}+(\tfrac{s}{\sqrt{8}})^2\bigr)\\ &= 8\bigl(1+2\tfrac{1}{\sqrt{2}}\tfrac{s}{\sqrt{8}}+(\tfrac{s}{\sqrt{8}})^2\bigr). \end{aligned}$$

Thus we have a real zero at $-\frac{1}{10}$, a pole at 0, and a pair of complex poles at $-2 \pm 2i$. Thus we write

$$T(s) = \frac{1}{80}\frac{1+10s}{s(1+2\frac{1}{\sqrt{2}}\frac{s}{\sqrt{8}}+(\frac{s}{\sqrt{8}})^2)}$$

and so

$$H(\omega) = \frac{1}{80}\frac{1+i10\omega}{i\omega(1+2i\frac{1}{\sqrt{2}}\frac{\omega}{\sqrt{8}}-(\frac{\omega}{\sqrt{8}})^2)}.$$

I find it easier to work with transfer functions first to avoid imaginary numbers as long as possible. You may do as you please, of course. •

One can easily imagine that one of the big weaknesses of our computer-absentee plots is that we have to find roots by hand...

Let us see what a Bode plot looks like for each of the basic elements. The idea is to see what the magnitude and phase looks like for small and large ω, and to "fill in the gaps" in between these ***asymptotes***.

1. $H(\omega) = K$: The Bode plot here is simple. It takes the magnitude $20 \log K$ for all values of $\log \omega$. The phase is 0° for all $\log \omega$ if K is positive, and 180° otherwise.
2. $H(\omega) = 1 + i\omega\tau$: For ω near zero the frequency response looks like 1 and so has the value of 0dB. For large ω the frequency response looks like $i\omega\tau$, and so the log of the magnitude will look like $\log \omega\tau$. Thus the magnitude plot for large frequencies we have $|H(\omega)| \approx 20 \log \omega\tau$dB. These asymptotes meet when $20 \log \omega\tau = 0$ or when $\omega = \frac{1}{\tau}$. This point is called the ***break frequency***. Note that the slope of the frequency response for large ω is independent of τ (since $\log \omega\tau = \log \omega + \log \tau$), but its break frequency does depend on τ. The phase plot starts at 0 for ω small and for large ω, since the frequency response is predominantly imaginary, becomes 90°. This Bode plot is shown in Figure 4.6 for $\tau = 1$.

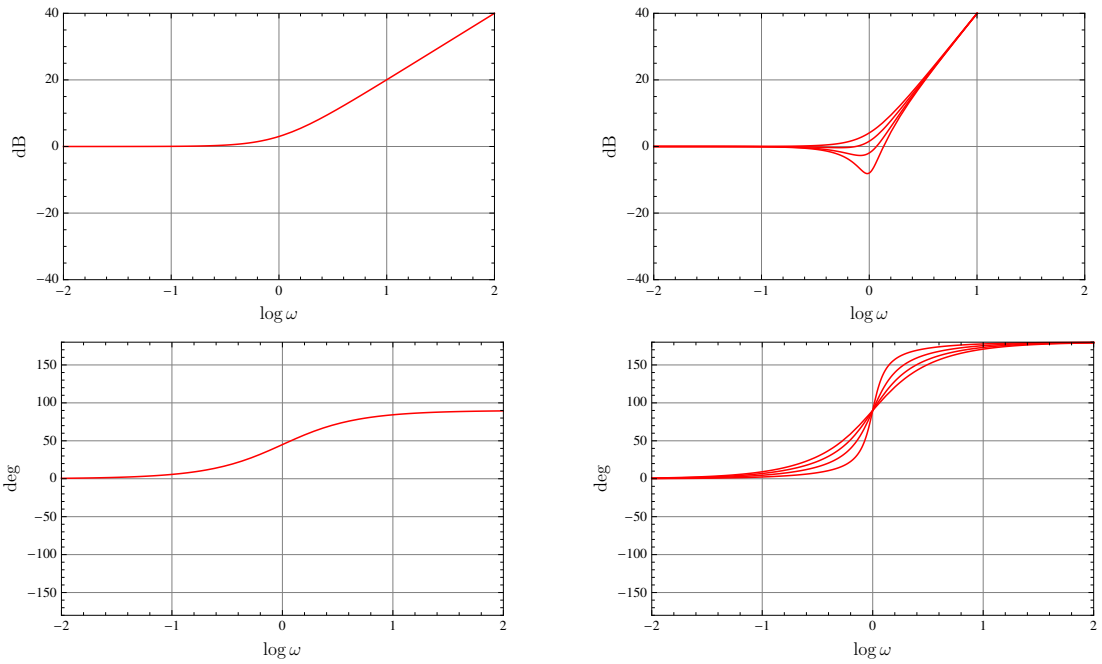


Figure 4.6 Bode plot for $H(\omega) = 1 + i\omega$ (left) and for $H(\omega) = 1 + 2i\zeta\omega - \omega^2$ for $\zeta = 0.2, 0.4, 0.6, 0.8$ (right)

3. $H(\omega) = 1 + 2i\zeta\frac{\omega}{\omega_0} - (\frac{\omega}{\omega_0})^2$: For small ω the magnitude is 1 or 0dB. For large ω the frequency response looks like $-(\frac{\omega}{\omega_0})^2$ and so the magnitude looks like $40 \log \frac{\omega}{\omega_0}$. The two asymptotes meet when $40 \log \frac{\omega}{\omega_0} = 0$ or when $\omega = \omega_0$. One has to be a bit more careful with what is happening around the frequency ω_0. The behaviour here depends on the value of ζ, and various plots are shown in Figure 4.6 for $\omega_0 = 1$. As ζ decreases, the undershoot increases. The phase starts out at 0° and goes to 180° as ω increases.
4. $H(\omega) = (i\omega)^{-1}$: The magnitude is $\frac{1}{\omega}$ over the entire frequency range which gives $|H(\omega)| = -20 \log \omega$dB. The phase is -90° over the entire frequency range, and the simple Bode plot is shown in Figure 4.7.

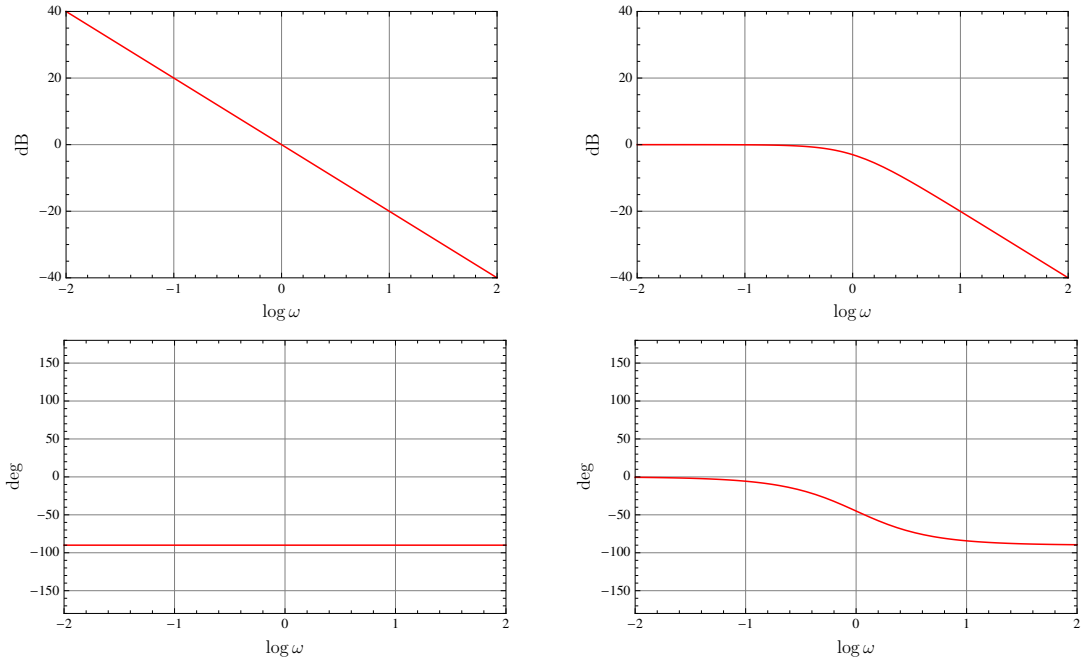


Figure 4.7 Bode plot for $H(\omega) = (i\omega)^{-1}$ (left) and for $H(\omega) = (1 + i\omega)^{-1}$ (right)

5. $H(\omega) = (1 + i\omega\tau)^{-1}$: The analysis here is just like that for a real zero except for signs. The Bode plot is shown in for $\tau = 1$.
6. $H(\omega) = (1 + 2i\zeta\frac{\omega}{\omega_0} - (\frac{\omega}{\omega_0})^2)^{-1}$: The situation here is much like that for a complex zero with sign reversal. The Bode plots are shown in Figure 4.8 for $\omega_0 = 1$.

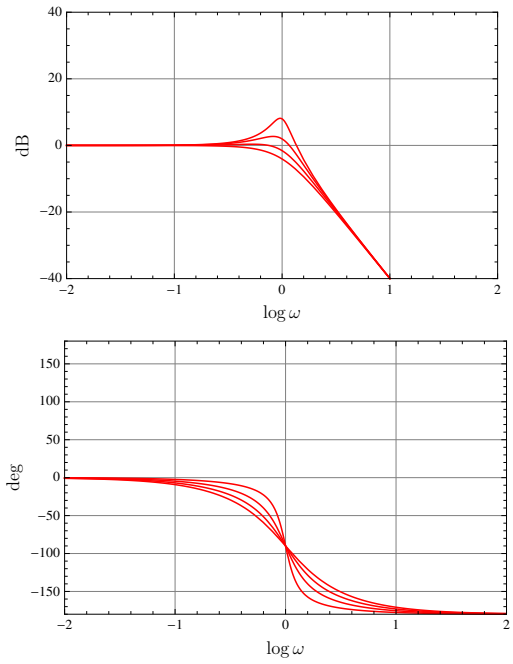


Figure 4.8 Bode plot for $H(\omega) = (1 + 2i\zeta\omega - \omega^2)^{-1}$ for $\zeta = 0.2, 0.4, 0.6, 0.8$

4.8 Remark We note that one often sees the language "20dB/decade." With what we have done above for the typical elements in a frequency response function. A first-order element in the numerator increases like $20\log\tau\omega$ for large frequencies. Thus as ω increases by a factor of 10, the magnitude will increase by 20dB. This is where "20dB/decade" comes from. If the first-order element is in the denominator, then the magnitude decreases at 20dB/decade. Second-order elements in the numerator increase at 40dB/decade, and second-order elements in the denominator decrease at 40dB/decade. In this way, one can ascertain the relative degree of the numerator and denominator of a frequency response function by looking at its slope for large frequencies. Indeed, we have the following rule:

> *The slope of the magnitude Bode plot for $H_{N,D}$ at large frequencies is* $20(\deg(N)-\deg(D))$dB/decade. •

To get a rough idea of how to sketch a Bode plot, the above arguments illustrate that the asymptotes are the most essential feature. Thus we illustrate these asymptotes in Figure 4.19 (see the end of the chapter) for the essential Bode plots in the above list. From these one can determine the character of most any Bode plot. The reason for this is that in (4.3) we have ensured that any frequency response is a product of the factors we have individually examined. Thus when we take logarithms as we do when generating a Bode plot, the graphs simply add! And the same goes for phase plots. So by plotting each term individually by the above rules, we end up with a pretty good rough approximation by adding the Bode plots.

Let us illustrate how this is done in an example.

4.9 Example **(Example 4.7 cont'd)** We take the frequency response

$$H(\omega)=\frac{1}{80}\frac{1+i10\omega}{i\omega\big(1+2i\frac{1}{\sqrt{2}}\frac{\omega}{\sqrt{8}}-(\frac{\omega}{\sqrt{8}})^2\big)}.$$

Four essential elements will comprise the frequency response:

1. $H_1(\omega)=\frac{1}{80}$;
2. $H_2(\omega)=1+i10\omega$;
3. $H_3(\omega)=(i\omega)^{-1}$;
4. $H_4(\omega)=\big(1+2i\frac{1}{\sqrt{2}}\frac{\omega}{\sqrt{8}}-(\frac{\omega}{\sqrt{8}})^2\big)^{-1}$.

Let's look first at the magnitudes.

1. H_1 will contribute $20\log\frac{1}{80}\approx -38.1$dB. The asymptotes for H_1 are shown in Figure 4.9.
2. H_2 has a break frequency of $\omega=\frac{1}{10}$ or $\log\omega=-1$. The asymptotes for H_2 are shown in Figure 4.9.
3. H_3 gives $-20\log\omega$ across the board. The asymptotes for H_3 are shown in Figure 4.10.
4. H_4 has a break frequency of $\omega=\sqrt{8}$ or $\log\omega\approx 0.45$. The asymptotes for H_4 are shown in Figure 4.10. Note that here we have $\zeta=\frac{1}{\sqrt{2}}$, which is a largish value. Thus we do not need to adjust the magnitude peak too much around the break frequency when we use the asymptotes to approximate the actual Bode plot.

Now the phase angles.

1. H_1 has phase exactly 0 for all frequencies.

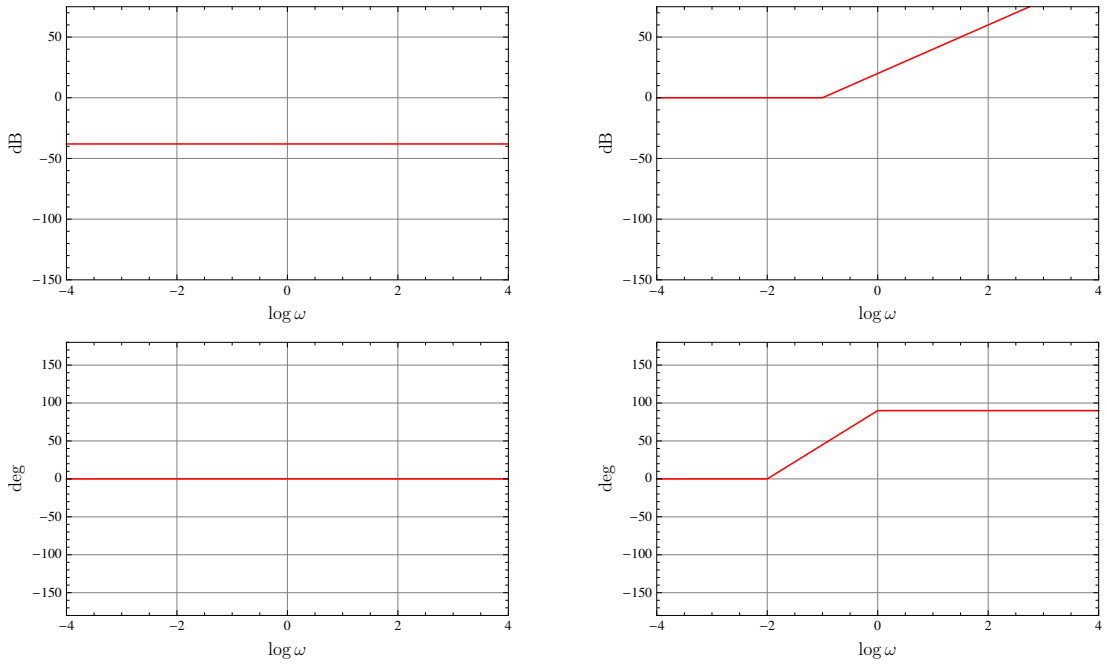


Figure 4.9 Asymptotes for Example 4.7: H_1 (left) and H_2 (right)

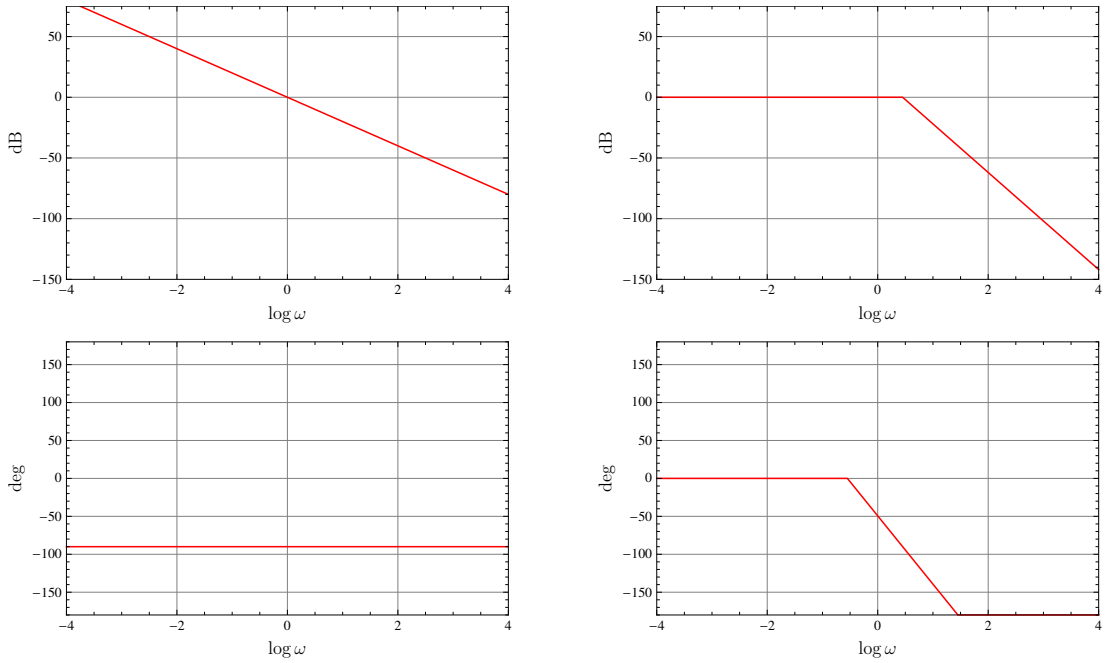


Figure 4.10 Asymptotes for Example 4.7: H_3 (left) and H_4 (right)

2. For H_2, the phase is approximately $0°$ for $\log 10\omega < -1$ or $\log \omega < -2$. For $\log 10\omega > 1$ (or $\log \omega > 0$) the phase is approximately $90°$. Between the frequencies $\log \omega = -2$ and $\log \omega = 0$ we interpolate linearly between the two asymptotic phase angles.
3. The phase for H_3 is $-90°$ for all frequencies.
4. For H_4, the phase is $0°$ for $\log \frac{\omega}{\sqrt{8}} < -1$ or $\log \omega < \log \sqrt{8} - 1 \approx -0.55$. For $\log \frac{\omega}{\sqrt{8}} > 1$,

$(\log\omega > \log\sqrt{8}+1 \approx 1.45)$, the phase is approximately $-180°$.

The sum of the asymptotes are plotted in Figure 4.11 along with the actual Bode plot so you

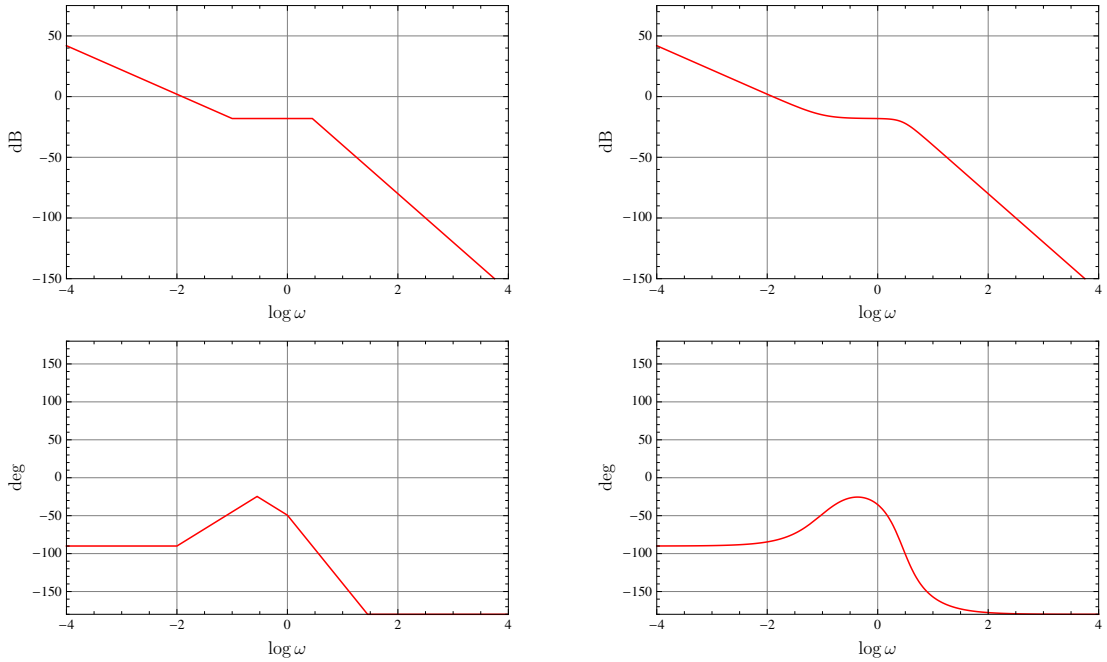


Figure 4.11 The sum of the asymptotes (left) and the actual Bode plot (right) for Example 4.7

can see how the Bode plot is essentially the sum of the individual Bode plots. You should take care that you always account for ζ, however. In our example, the value of ζ in H_4 is quite large, so not much of an adjustment had to be made. If ζ were small, we have to add a little bit of a peak in the magnitude around the break frequency $\log\omega = \log\sqrt{8}$, and also make the change in the phase a bit steeper. •

4.3.3 The polar frequency response plot

We will encounter in Chapter 12 another representation of the frequency response H. The idea here is that rather than plotting magnitude and phase as one does in a Bode plot, one plots the real and imaginary part of the frequency response as a curve in the complex plane parameterised by $\omega \in (0,\infty)$. Doing this yields the ***polar plot*** for the frequency response. One could do this, for example, by taking the Bode plot, and for each point ω on the independent variable axis, put a point at a distance $|H(\omega)|$ from the origin in the direction $\angle H(\omega)$. Indeed, given the Bode plot, one can typically make a pretty good approximation of the polar plot by noting (1) the maxima and minima of the magnitude response, and the phase at these maxima and minima, and (2) the magnitude when the phase is 0, $\pm 90°$, or $\pm 180°$.

In Figure 4.12 are shown the polar plots for the basic frequency response functions. Recall that in (4.3) we indicated that any frequency response will be a product of these basic elements, and so one can determine the polar plot for a frequency response formed by the product of such elements by performing complex multiplication that, you will recall, is done in polar coordinates merely by multiplying radii, and adding angles.

For a lark, let's look at the minimum/nonminimum phase example in polar form.

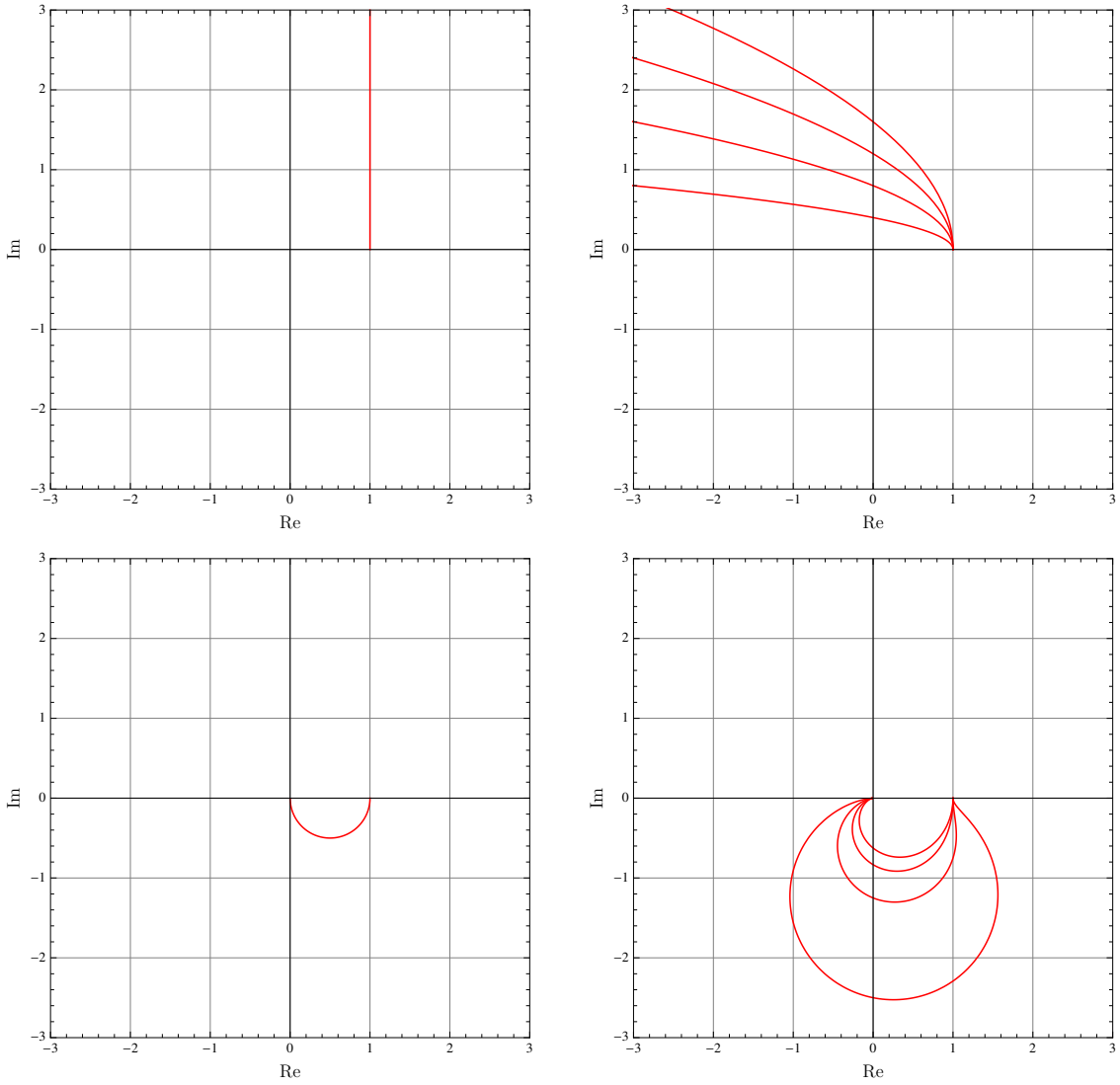


Figure 4.12 Polar plots for $H(\omega) = 1 + i\omega$ (top left), $H(\omega) = 1 + 2i\zeta\omega - \omega^2$, $\zeta = 0.2, 0.4, 0.6, 0.8$ (top right), $H(\omega) = (1 + i\omega)^{-1}$ (bottom left), and $H(\omega) = (1 + 2i\zeta\omega - \omega^2)^{-1}$, $\zeta = 0.2, 0.4, 0.6, 0.8$

4.10 Example **(Example 4.12)** Recall that we contrasted the two transfer functions

$$H_{\Sigma_1}(\omega) = \frac{1 + i\omega}{-\omega^2 + i\omega + 1}, \quad H_{\Sigma_2}(\omega) = \frac{1 - i\omega}{-\omega^2 + i\omega + 1}.$$

We contrast the polar plots for these frequency responses in Figure 4.13. Note that, as expected, the minimum phase system undergoes a smaller phase change if we follow it along its parameterised polar curve. We shall see the potential dangers of this in Chapter 12. •

Note that when making a polar plot, the thing one looses is frequency information. That is, one can no longer read from the plot the frequency at which, say, the magnitude of the frequency response is maximum. For this reason, it is not uncommon to place at intervals along a polar plot the frequencies corresponding to various points.

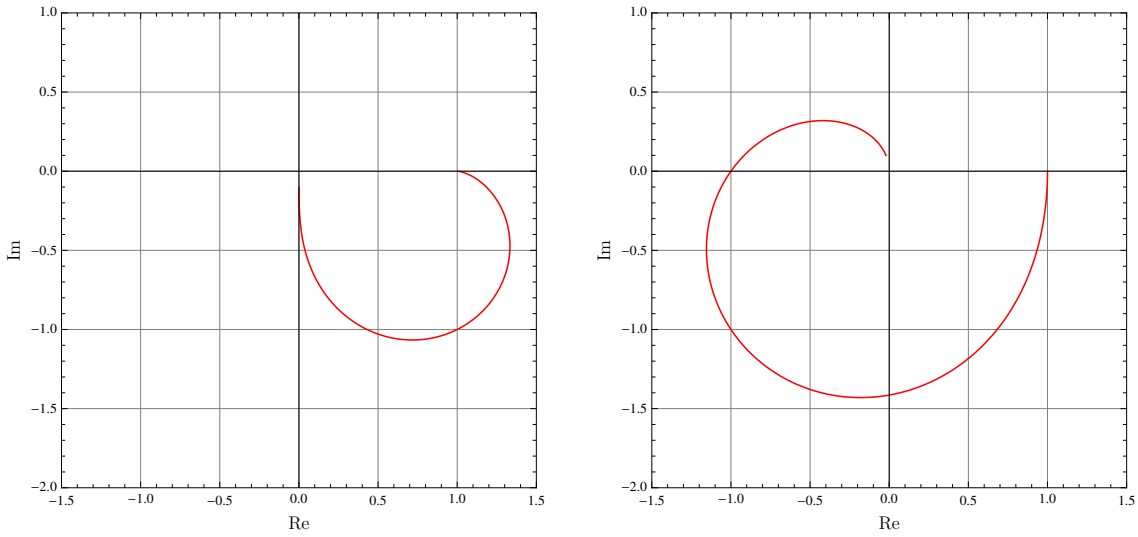


Figure 4.13 Polar plots for minimum phase (left) and nonminimum phase (right) systems

4.4 Properties of the frequency response

It turns out that in the frequency response can be seen some of the behaviour we have encountered in the time-domain and in the transfer function. We set out in this section to scratch the surface behind interpreting the frequency response. However, this is almost an art as much as a science, so plain experience counts for a lot here.

4.4.1 Time-domain behaviour reflected in the frequency response

We have seen in Section 3.2 we saw that some of the time-domain properties discussed in Section 2.3 were reflected in the transfer function. We anticipate being able to see these same features reflected in the frequency response, and ergo in the Bode plot. In this section we explore these expected relationships. We do this by looking at some examples.

4.11 Example The first example we look at is one where we have a pole/zero cancellation. As per Theorem 3.5 this indicates a lack of observability in the system. It is most beneficial to look at what happens when the pole and zero do not actually cancel, and compare it to what happens when the pole and zero really do cancel. We take

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & -\epsilon \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{4.4}$$

and we compute

$$H_\Sigma(\omega) = \frac{1 + i\omega}{-\omega^2 + i\epsilon\omega - 1}.$$

The Bode plots for three values of ϵ are shown in Figure 4.14. What are the essential features here? Well, by choosing the values of $\epsilon \neq 0$ to deviate significantly from zero, we can see accentuated two essential points. Firstly, when $\epsilon \neq 0$ the magnitude plot has two regions where the magnitude drops off at different slopes. This is a consequence of there being two different exponents in the characteristic polynomial. When $\epsilon = 0$ the plot tails off at one slope, indicating that the system is first-order and has only one characteristic exponent. This

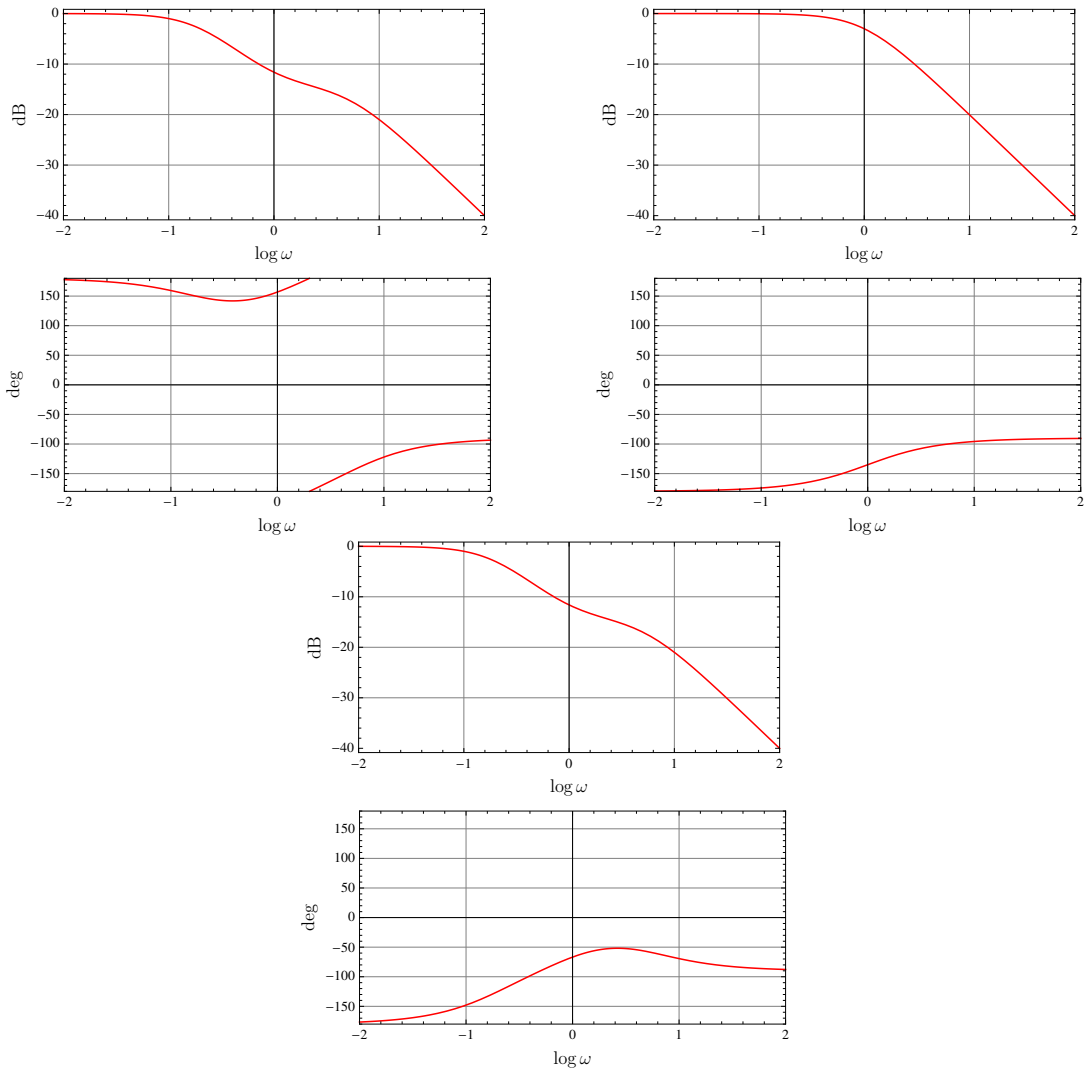


Figure 4.14 Bode plots for (4.4) for $\epsilon = -5$, $\epsilon = 0$, and $\epsilon = 5$

is a consequence of the pole/zero cancellation that occurs when $\epsilon = 0$. Note, however, that we cannot look at one Bode plot and ascertain whether or not the system is observable. •

There is also an effect that can be observed in the Bode plot for a system that is not minimum phase. An example illustrates this well.

4.12 Example We consider two SISO linear systems, both with

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{4.5}$$

The two output vectors we look at are

$$\boldsymbol{c}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{c}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{4.6}$$

Let us then denote $\Sigma_1 = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}_1, \boldsymbol{0}_1)$ and $\Sigma_2 = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}_2, \boldsymbol{0}_1)$. The two frequency response functions are

$$H_{\Sigma_1}(\omega) = \frac{1 + i\omega}{-\omega^2 + i\omega + 1}, \quad H_{\Sigma_2}(\omega) = \frac{1 - i\omega}{-\omega^2 + i\omega + 1}.$$

The Bode plots are shown in Figure 4.15. What should one observe here? Note that the

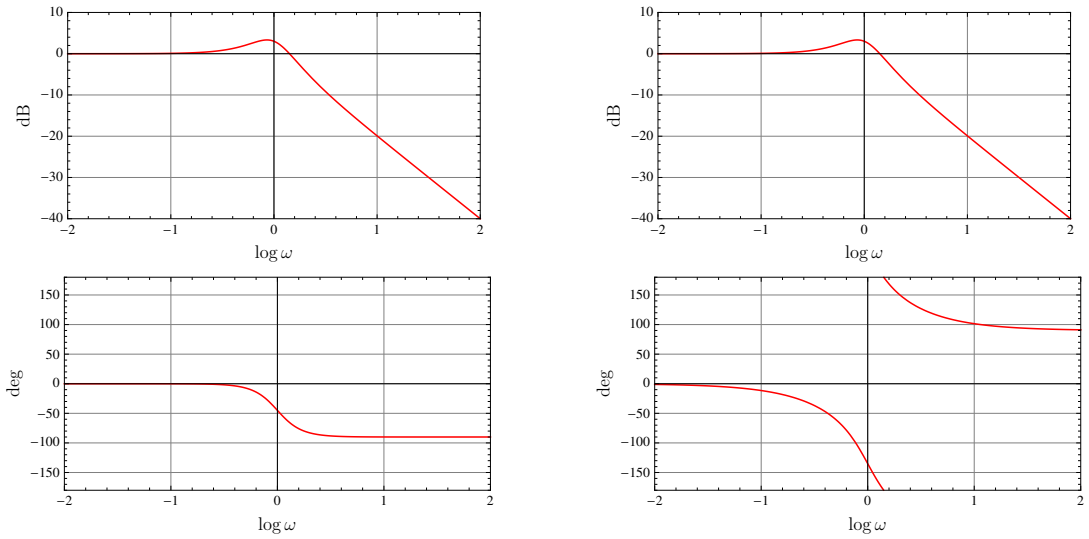


Figure 4.15 Bode plots for the two systems of (4.5) and (4.6)—H_{Σ_1} is on the left and H_{Σ_2} is on the right

magnitude plots are the same, and this can be verified by looking at the expressions for H_{Σ_1} and H_{Σ_2}. The differences occur in the phase plots. Note that the phase angle varies only slightly for Σ_1 across the frequency range, but it varies more radically for Σ_2. It is from this behaviour that the term "minimum phase" is derived. •

4.4.2 Bode's Gain/Phase Theorem

In Bode's book (1945) one can find a few chapters on some properties of the frequency response. In this section we begin this development, and from it derive Bode's famous "Gain/Phase Theorem." The material in this section relies on some ideas from complex function theory

that we review in Appendix D. We start by examining some basic properties of frequency response functions. Here we begin to see that the real and imaginary parts of a frequency response function are not arbitrary functions.

4.13 Proposition *Let (N, D) be a SISO linear system in input/output form with $H_{N,D}$ the frequency response. The following statements hold:*

(i) $\mathrm{Re}(H_{N,D}(-\omega)) = \mathrm{Re}(H_{N,D}(\omega))$;

(ii) $\mathrm{Im}(H_{N,D}(-\omega)) = -\mathrm{Im}(H_{N,D}(\omega))$;

(iii) $|H_{N,D}(-\omega)| = |H_{N,D}(\omega)|$;

(iv) $\measuredangle H_{N,D}(-\omega) = -\measuredangle H_{N,D}(\omega)$ *provided* $\measuredangle H_{N,D}(\omega) \in (-\pi, \pi)$.

Proof We prove (i) and (ii) together. It is certainly true that the real parts of $D(i\omega)$ and $N(i\omega)$ will involve terms that have even powers of ω and that the imaginary parts of $D(i\omega)$ and $N(i\omega)$ will involve terms that have odd powers of ω. Therefore, if we denote by $D_1(\omega)$ and $D_2(\omega)$ the real and imaginary parts of $D(i\omega)$ and $N_1(\omega)$ and $N_2(\omega)$ the real and imaginary parts of $N(i\omega)$, we have

$$N_1(-\omega) = N_1(\omega), \quad D_1(-\omega) = D_1(\omega), \quad N_2(-\omega) = -N_2(\omega), \quad D_2(-\omega) = -D_2(\omega). \quad (4.7)$$

We also have

$$\begin{aligned} H_{N,D}(\omega) &= \frac{N_1(\omega) + iN_2(\omega)}{D_1(\omega) + iD_2(\omega)} \\ &= \frac{N_1(\omega)D_1(\omega) + N_2(\omega)D_2(\omega)}{D_1^2(\omega) + D_2^2(\omega)} + i\frac{N_2(\omega)D_1(\omega) - N_1(\omega)D_2(\omega)}{D_1^2(\omega) + D_2^2(\omega)}, \end{aligned}$$

so that

$$\begin{aligned} \operatorname{Re}(H_{N,D}(\omega)) &= \frac{N_1(\omega)D_1(\omega) + N_2(\omega)D_2(\omega)}{D_1^2(\omega) + D_2^2(\omega)}, \\ \operatorname{Im}(H_{N,D}(\omega)) &= \frac{N_2(\omega)D_1(\omega) - N_1(\omega)D_2(\omega)}{D_1^2(\omega) + D_2^2(\omega)}. \end{aligned}$$

Using the relations (4.7) we see that

$$\begin{aligned} N_1(-\omega)D_1(-\omega) + N_2(-\omega)D_2(-\omega) &= N_1(\omega)D_1(\omega) + N_2(\omega)D_2(\omega), \\ N_2(-\omega)D_1(-\omega) - N_1(-\omega)D_2(-\omega) &= N_2(\omega)D_1(\omega) - N_1(\omega)D_2(\omega), \end{aligned}$$

and from this the assertions (i) and (ii) obviously follow.

(iii) This is a consequence of (i) and (ii) and the definition of $|\cdot|$.

(iv) This follows from (i) and (ii) and the properties of arctan. ■

Now we turn our attention to the crux of the material in this section—a look at how the magnitude and phase of the frequency response are related. To relate these quantities, it is necessary to represent the frequency response in the proper manner. To this end, for a SISO linear system (N, D) in input/output form, we define $\operatorname{ZP}_{N,D} \subset \mathbb{C}$ to be the set of zeros and poles of $T_{N,D}$. We may then define $S_{N,D} \colon \mathbb{C} \setminus \operatorname{ZP}_{N,D} \to \mathbb{C}$ by

$$S_{N,D}(s) = \ln(T_{N,D}(s)),$$

noting that $S_{N,D}$ is analytic on $\mathbb{C} \setminus \operatorname{ZP}_{N,D}$. We recall that from the properties of the complex logarithm we have

$$\operatorname{Re}(S_{N,D}(s)) = \ln|T_{N,D}(s)|, \quad \operatorname{Im}(S_{N,D}(s)) = \measuredangle T_{N,D}(s)$$

for $s \in \mathbb{C} \setminus \operatorname{ZP}_{N,D}$. Along similar lines we define $\overline{\operatorname{ZP}}_{N,D} \subset \mathbb{R}$ by

$$\overline{\operatorname{ZP}}_{N,D} = \{\omega \in \mathbb{R} \mid \ i\omega \in \operatorname{ZP}_{N,D}\}.$$

Then we may define $G_{N,D} \colon \mathbb{R} \setminus \overline{\operatorname{ZP}}_{N,D} \to \mathbb{C}$ by

$$G_{N,D}(\omega) = S_{N,D}(i\omega) = \ln(H_{N,D}(\omega)). \tag{4.8}$$

We can employ our previous use of the properties of the complex logarithm to assert that

$$\operatorname{Re}(G_{N,D}(\omega)) = \ln|H_{N,D}(\omega)|, \quad \operatorname{Im}(G_{N,D}(\omega)) = \measuredangle H_{N,D}(\omega).$$

Note that these are almost the quantities one plots in the Bode plot. The phase is precisely what is plotted in the Bode plot (against a logarithmic frequency scale), and $\operatorname{Re}(G_{N,D}(\omega))$ is related to the magnitude plotted in the Bode plot by

$$\operatorname{Re}(G_{N,D}(\omega)) = \ln 10 \log|H_{N,D}(\omega)| = \frac{\ln 10}{20}|H_{N,D}(\omega)| \text{ dB}.$$

Note that the relationship is a simple one as it only involves scaling by a constant factor. The quantity $\operatorname{Re}(G_{N,D}(\omega))$ is measured in the charming units of ***neppers***.

Recall that (N, D) is stable if all roots of D lie in $\mathbb{C}_-$ and is minimum phase if all roots of N lie in $\overline{\mathbb{C}}_-$. Here we will require the additional assumption that N have no roots on the imaginary axis. Let us say that a SISO linear system (N, D) for which all roots of N lie in $\mathbb{C}_-$ is ***strictly minimum phase***.

4.14 Proposition *Let (N, D) be a proper SISO linear system in input/output form that is stable and strictly minimum phase, and let $\omega_0 > 0$. We then have*

$$\operatorname{Im}(G_{N,D}(\omega_0)) = \frac{2\omega_0}{\pi} \int_0^\infty \frac{\operatorname{Re}(G_{N,D}(\omega)) - \operatorname{Re}(G_{N,D}(\omega_0))}{\omega^2 - \omega_0^2} \, \mathrm{d}\omega.$$

Proof Throughout the proof we denote by $G_1(\omega)$ the real part of $G_{N,D}(\omega)$ and by $G_2(\omega)$ the imaginary part of $G_{N,D}(\omega)$.

Let $U_{\omega_0} \subset \mathbb{C}$ be the open subset defined by

$$U_{\omega_0} = \{s \in \mathbb{C} \mid \operatorname{Re}(s) > 0, \ \operatorname{Im}(s) \neq \pm\omega_0\}.$$

We now define a closed contour whose interior contains points in U_{ω_0}. We do this in parts. First, for $R > \omega_0$ define a contour Γ_R in U_{ω_0} by

$$\Gamma_R = \{Re^{i\theta} \mid \ -\tfrac{\pi}{2} \leq \theta \leq \tfrac{\pi}{2}\}.$$

Now for $r > 0$ define two contours

$$\Gamma_{r,1} = \{i\omega_0 + re^{i\theta} \mid \ -\tfrac{\pi}{2} \leq \theta \leq \tfrac{\pi}{2}\}, \quad \Gamma_{r,2} = \{-i\omega_0 + re^{i\theta} \mid \ -\tfrac{\pi}{2} \leq \theta \leq \tfrac{\pi}{2}\}.$$

Finally define a contours $\Gamma_{r,j}$, $j = 3, 4, 5$, by

$$\begin{aligned} \Gamma_{\omega_0,3} &= \{i\omega \mid \ -\infty < \omega_0 \leq -\omega_0 - r\} \\ \Gamma_{\omega_0,4} &= \{i\omega \mid \ -\omega_0 + r \leq \omega_0 \leq \omega_0 - r\} \\ \Gamma_{\omega_0,5} &= \{i\omega \mid \ \omega + r \leq \omega_0 < \infty\}. \end{aligned}$$

The closed contour we take is then

$$\Gamma_{R,r} = \Gamma_R \bigcup_{j=1}^{5} \Gamma_{r,j}.$$

We show this contour in Figure 4.16. Now define a function $F_{\omega_0} \colon U_{\omega_0} \to \mathbb{C}$ by

$$F_{\omega_0}(s) = \frac{2i\omega_0(S_{N,D}(s) - G_1(\omega_0))}{s^2 + \omega_0^2}.$$

Since (N, D) is stable and strictly minimum phase, $S_{N,D}$ is analytic on $\mathbb{C}_+$, and so F_{ω_0} is analytic on U_{ω_0}, and so we may apply Cauchy's Integral Theorem to the integral of F_{ω_0} around the closed contour $\Gamma_{R,r}$.

Let us evaluate that part of this contour integral corresponding to the contour Γ_R as R get increasingly large. We claim that

$$\lim_{R\to\infty} \int_{\Gamma_R} F_{\omega_0}(s) \, \mathrm{d}s = 0.$$

Indeed we have

$$F_{\omega_0}(Re^{i\theta}) = \frac{2i\omega_0 S_{N,D}(Re^{i\theta})}{-R^2e^{2i\theta} + \omega_0^2} - \frac{2i\omega_0 G_1(\omega_0)}{-R^2e^{2i\theta} + \omega_0^2}.$$

Since $\deg(N) < \deg(D)$ the first term on the right will behave like R^{-2} as $R \to \infty$ and the second term will also behave like R^{-2} as $R \to \infty$. Since $\mathrm{d}s = iRe^{i\theta}$ on Γ_R, the integrand in $\int_{\Gamma_R} F_{\omega_0}(s) \, \mathrm{d}s$ will behave like R^{-1} as $R \to \infty$, and so our claim follows.

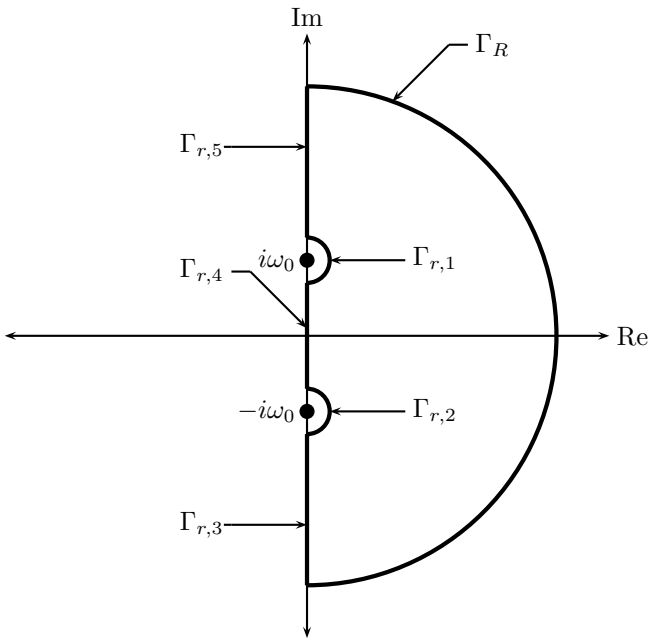


Figure 4.16 The contour $\Gamma_{R,r}$ used in the proof of Proposition 4.14

Now let us examine what happens as we let $r \to 0$. To evaluate the contributions of $\Gamma_{r,1}$ and $\Gamma_{r,2}$ we write

$$F_{\omega_0}(s) = \frac{S_{N,D}(s) - G_1(\omega_0)}{s - i\omega_0} - \frac{S_{N,D}(s) - G_1(\omega_0)}{s + i\omega_0}.$$

On $\Gamma_{r,1}$ we have $s = i\omega_0 + re^{i\theta}$ so that on $\Gamma_{r,1}$ we have

$$F_{\omega_0}(i\omega_0 + re^{i\theta}) = \frac{S_{N,D}(i\omega_0 + re^{i\theta}) - G_1(\omega_0)}{re^{i\theta}} - \frac{S_{N,D}(i\omega_0 + re^{i\theta}) - G_1(\omega_0)}{2i\omega_0 + re^{i\theta}}.$$

We parameterise $\Gamma_{r,1}$ with the curve $c\colon [-\frac{\pi}{2}, \frac{\pi}{2}] \to \mathbb{C}$ defined by $c(t) = i\omega_0 + re^{it}$ so that $c'(t) = ire^{it}$. Thus, as $r \to 0$, $F_{\omega_0}(s)\,\mathrm{d}s$ behaves like

$$F_{\omega_0}(s)\,\mathrm{d}s \approx \frac{(S_{N,D}(i\omega_0) - G_1(\omega_0))ire^{it}}{re^{it}} = i(G_{N,D}(\omega_0) - G_1(\omega_0)),$$

using the parameterisation specified by c. Integrating gives

$$\int_{\Gamma_{r,1}} F_{\omega_0}(s)\,\mathrm{d}s = i\pi(G_{N,D}(\omega_0) - G_1(\omega_0)).$$

In similar fashion one obtains

$$\int_{\Gamma_{r,2}} F_{\omega_0}(s)\,\mathrm{d}s = -i\pi(G_{N,D}(-\omega_0) - G_1(\omega_0)).$$

Now we use Proposition 4.13 to assert that

$$G_{N,D}(\omega_0) - G_{N,D}(-\omega_0) = 2iG_2(\omega_0).$$

Therefore

$$\begin{aligned}\int_{\Gamma_{r,1}} F_{\omega_0}(s)\,\mathrm{d}s + \int_{\Gamma_{r,2}} F_{\omega_0}(s)\,\mathrm{d}s &= i\pi(G_{N,D}(\omega_0) - G_1(\omega_0) - (G_{N,D}(-\omega_0) - G_1(\omega_0)))\\ &= -2\pi G_2(\omega_0).\end{aligned}$$

Finally, we look at the integrals along the contours $\Gamma_{r,3}$, $\Gamma_{r,4}$, and $\Gamma_{r,5}$ as $r \to 0$ and for fixed $R > \omega_0$. These contour integrals in the limit will yield a single integral along the a portion of the imaginary axis:

$$\int_{i[-R,R]} F_{\omega_0}(s)\,\mathrm{d}s.$$

We can parameterise the contour in this integral by the curve $c\colon [-R, R] \to \mathbb{C}$ defined by $c(t) = it$. Thus $c'(t) = i$, giving

$$\begin{aligned}\int_{i[-R,R]} F_{\omega_0}(s)\,\mathrm{d}s &= \int_{-R}^{R} \frac{2i\omega_0(S_{N,D}(it) - G_1(\omega_0))}{\omega_0^2 - t^2} i\,\mathrm{d}t\\ &= \int_{-R}^{R} \frac{2\omega_0(S_{N,D}(it) - G_1(\omega_0))}{t^2 - \omega_0^2}\,\mathrm{d}t.\end{aligned}$$

Using Proposition 4.13 we can write this as

$$\int_{i[-R,R]} F_{\omega_0}(s)\,\mathrm{d}s = \int_0^R \frac{4\omega_0(G_1(t) - G_1(\omega_0))}{t^2 - \omega_0^2}\,\mathrm{d}t.$$

Collecting this with our expression for the integrals along $\Gamma_{r,1}$ and $\Gamma_{r,2}$, as well as noting our claim that the integral along Γ_R vanishes as $R \to \infty$, we have shown that

$$\lim_{r\to 0}\lim_{R\to\infty} \int_{\Gamma_{R,r}} F_{\omega_0}(s)\,\mathrm{d}s = \int_0^\infty \frac{4\omega_0(G_1(\omega) - G_1(\omega_0))}{\omega^2 - \omega_0^2}\,\mathrm{d}\omega - 2\pi G_2(\omega_0).$$

By Cauchy's Integral Theorem, this integral should be zero, and our result now follows from straightforward manipulation. ■

The following result gives an important property of stable strictly minimum phase systems.

4.15 Theorem **(Bode's Gain/Phase Theorem)** *Let (N, D) be a proper, stable, strictly minimum phase SISO linear system in input/output form, and let $G_{N,D}(\omega)$ be as defined in (4.8). For $\omega_0 > 0$ define $M_{N,D}^{\omega_0}\colon \mathbb{R} \to \mathbb{R}$ by $M_{N,D}^{\omega_0}(u) = \mathrm{Re}(G_{N,D}(\omega_0 e^U))$. Then we have*

$$\measuredangle H_{N,D}(\omega_0) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\mathrm{d}M_{N,D}^{\omega_0}}{\mathrm{d}u} \ln\coth\left|\tfrac{u}{2}\right| \mathrm{d}u.$$

Proof The theorem follows fairly easily from Proposition 4.14. By that result we have

$$\measuredangle H_{N,D}(\omega_0) = \frac{2\omega_0}{\pi}\int_0^\infty \frac{\mathrm{Re}(G_{N,D}(\omega)) - \mathrm{Re}(G_{N,D}(\omega_0))}{\omega^2 - \omega_0^2}\,\mathrm{d}\omega. \tag{4.9}$$

We make the change of variable $\omega = \omega_0 e^u$, upon which the integral in (4.9) becomes

$$\frac{2}{\pi}\int_{-\infty}^{\infty} \frac{M_{N,D}^{\omega_0}(u) - M_{N,D}^{\omega_0}(0)}{e^u - e^{-u}}\,\mathrm{d}u = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{M_{N,D}^{\omega_0}(u) - M_{N,D}^{\omega_0}(0)}{\sinh u}\,\mathrm{d}u.$$

We note that

$$\int \frac{\mathrm{d}u}{\sinh u} = \ln \coth \tfrac{u}{2},$$

and we may use this formula, combined with integration by parts, to determine that

$$\frac{1}{\pi}\int_0^\infty \frac{M_{N,D}^{\omega_0}(u) - M_{N,D}^{\omega_0}(0)}{\sinh u}\,\mathrm{d}u = \frac{1}{\pi}\int_0^\infty \frac{\mathrm{d}M_{N,D}^{\omega_0}}{\mathrm{d}u}\ln\coth\tfrac{u}{2} - \frac{1}{\pi}(M_{N,D}^{\omega_0}(u) - M_{N,D}^{\omega_0}(0))\ln\coth\tfrac{u}{2}\Big|_0^\infty, \quad (4.10)$$

and

$$\frac{1}{\pi}\int_{-\infty}^0 \frac{M_{N,D}^{\omega_0}(u) - M_{N,D}^{\omega_0}(0)}{\sinh u}\,\mathrm{d}u = \frac{1}{\pi}\int_\infty^0 \frac{\mathrm{d}M_{N,D}^{\omega_0}}{\mathrm{d}u}\ln\coth\tfrac{-u}{2} + \frac{1}{\pi}(M_{N,D}^{\omega_0}(u) - M_{N,D}^{\omega_0}(0))\ln\coth\tfrac{-u}{2}\Big|_0^\infty. \quad (4.11)$$

Let us look at the first term in each of these integrals. At the limit $u = 0$ we may compute

$$\coth\tfrac{u}{2} \approx \frac{2}{u} + \frac{u}{6} - \frac{u^3}{360} + \cdots$$

so that near $u = 0$, $\ln\coth\frac{u}{2} \approx -\ln\frac{u}{2}$. Also, since $M_{N,D}^{\omega_0}(u)$ is analytic at $u = 0$, $M_{N,D}^{\omega_0}(u) - M_{N,D}^{\omega_0}(0)$ will behave linearly in u for u near zero. Therefore,

$$\lim_{u\to 0} M_{N,D}^{\omega_0}(u) - M_{N,D}^{\omega_0}(0))\ln\coth\tfrac{u}{2} \approx -u\ln\frac{u}{2}.$$

Recalling that $\lim_{u\to 0} u\ln u = 0$, we see that the lower limit in the above integrated expressions is zero. At the other limits as $u \to \pm\infty$, $\ln\coth\frac{u}{2}$ behaves like e^{-u} as $u \to +\infty$ and like e^u as $u \to -\infty$. This, combined with the fact that $M_{N,D}^{\omega_0}(u)$ behaves like $\ln u$ as $u \to \infty$, implies the vanishing of the upper limits in the integrated expressions in both (4.10) and (4.11). Thus we have shown that these integrated terms both vanish. From this the result follows easily. ■

It is not perhaps perfectly clear what is the import of this theorem, so let us examine it for a moment. In this discussion, let us fix $\omega_0 > 0$. Bode's Gain/Phase Theorem is telling us that the phase angle for the frequency response can be determined from the slope of the Bode plot with decibels plotted against the logarithm of frequency. The contribution to the phase of the slope at some $u = \omega_0 \ln\omega$ is determined by the weighting factor $\ln\coth\left|\frac{u}{2}\right|$ which we plot in Figure 4.17. From the figure we see that the slopes near $u = 0$, or $\omega = \omega_0$, contribute most to the phase angle. But keep in mind that this only works for stable strictly minimum phase systems. What it tells us is that for such systems, if one wishes to specify a certain magnitude characteristic for the frequency response, the phase characteristic is *completely* determined.

Let's see how this works in an example.

4.16 Example Suppose we have a Bode plot with $y = |H_{N,D}(\omega)|\text{dB}$ versus $x = \log\omega$, and that $y = 20kx + b$ for $k \in \mathbb{Z}$ and $b \in \mathbb{R}$. Thus the magnitude portion of the Bode plot is linear.

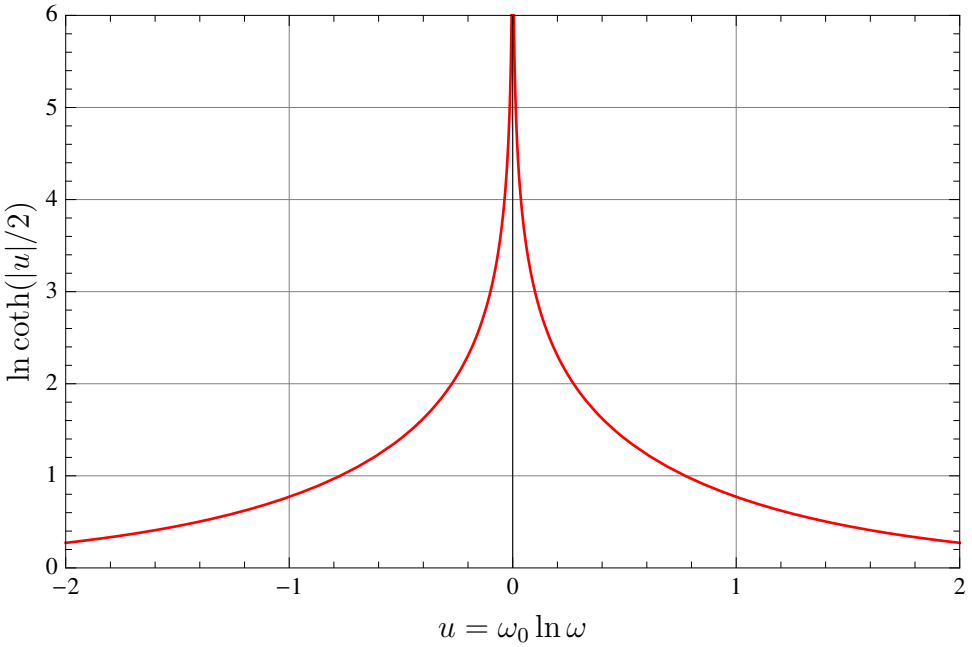


Figure 4.17 The weighting factor in Bode's Gain/Phase Theorem

To employ the Gain/Phase Theorem we should convert this to a relation in $\tilde{y} = \ln|H_{N,D}(\omega)|$ versus $\tilde{x} = \ln \frac{\omega}{\omega_0}$. The coordinates are then readily seen to be related by

$$x = \frac{\tilde{x}}{\ln 10} + \log \omega_0, \quad y = \frac{20}{\ln 10}\tilde{y}.$$

Therefore the relation $y = 20kx + b$ becomes

$$\tilde{y} = k\tilde{x} + k \ln \omega_0 + \frac{b \ln 10}{20}.$$

In the terminology of Theorem 4.15 we thus have

$$M_{N,D}^{\omega_0}(u) = ku + k \ln \omega_0 + \frac{b \ln 10}{20}$$

so that $\frac{\mathrm{d}M_{N,D}^{\omega_0}}{\mathrm{d}u} = k$. The Gain/Phase Theorem tells us that we may obtain the phase at any frequency ω_0 as

$$\measuredangle H_{N,D}(\omega_0) = \frac{k}{\pi} \int_{-\infty}^{\infty} \ln \coth \left|\tfrac{u}{2}\right| \mathrm{d}u.$$

The integral is one that can be looked up (Mathematica® evaluated it for me) to be $\frac{\pi^2}{2}$, so that we have $\measuredangle H_{N,D}(\omega_0) = \frac{k\pi}{2}$.

Let's see if this agrees with what we have seen before. We know, after all, a transfer function whose magnitude Bode plot is $20k \log \omega$. Indeed, one can check that choosing the transfer function $T_{N,D}(s) = s^k$ gives $20 \log|H_{N,D}(\omega)| = 20k \log \omega$, and so this transfer function is of the type for which we are considering in the case when $b = 0$. For systems of this type we can readily determine the phase to be $\frac{k\pi}{2}$, which agrees with Bode's Gain Phase Theorem. •

Despite the fact that it might be possible to explicitly apply Bode's Gain/Phase Theorem in some examples to directly compute the phase characteristic for a certain magnitude characteristic, the value of the theorem lies not in the ability to do such computations, but rather in its capacity to give us approximate information for stable and strictly minimum phase systems. Indeed, it is sometimes useful to make the approximation

$$\measuredangle H_{N,D}(\omega_0) = \frac{\mathrm{d}M^{\omega_0}_{N,D}}{\mathrm{d}u}\Big|_{u=0} \frac{\pi}{2}, \tag{4.12}$$

and this approximation becomes better when one is in a region where the slope of the magnitude characteristic on the Bode plot is large at ω_0 compared to the slope at other frequencies.

Uses of Gain/Phase theorem

4.5 Uncertainly in system models

As hinted at in Section 1.2 in the context of the simple DC servo motor example, robustness of a design to uncertainties in the model is a desirable feature. In recent years, say the last twenty years, rigorous mathematical techniques have been developed to handle model uncertainty, these going under the name of "robust control." These matters will be touched upon in this book, and in this section, we look at the first aspect of this: representing uncertainty in system models. The reader will observe that this uncertainty representation is done in the frequency domain. The reason for this is merely that the tools for controller design that have been developed up to this time rely on such a description. In the context of this book, this culminates in Chapter 15 with a systematic design methodology keeping robustness concerns foremost.

In this section it is helpful to introduce the $\mathbf{H}_\infty$***-norm*** for a rational function. Given $R \in \mathbb{R}(s)$ we denote

$$\|R\|_\infty = \sup_{\omega \in \mathbb{R}} \{|R(i\omega)|\}.$$

This will be investigated rather more systematically in Section 5.3.2, but for now the meaning is rather pedestrian: it is the maximum value of the magnitude Bode plot.

4.5.1 Structured and unstructured uncertainty

The reader may wish to recall our general control theoretic terminology from Section 1.1. In particular, recall that a "plant" is that part of a control system that is given to the control designer. What is given to the control designer is a model, hopefully in something vaguely resembling one of the three forms we have thus far discussed: a state-space model, a transfer function, of a frequency response function. Of course, this model cannot be expected to be perfect. If one is uncertain about a plant model, one should make an attempt to come up with a mathematical description of what this means. There are many possible candidates, and they can essentially be dichotomised as ***structured uncertainty*** and ***unstructured uncertainty***. The idea with structured uncertainty is that one has a specific type of plant model in mind, and parameters in that plant model are regarded as uncertain. In this approach, one wishes to design a controller that has desired properties for all possible values of the uncertain parameters.

4.17 Example Suppose a mass is moving under a the influence of a control force u. The precise value of the mass could be unknown, say $m \in [m_1, m_2]$. In this case, a control design should be thought of as being successful if it accomplishes stated objectives (the reader does not

know what these might be at this point in the book!) for *all* possible values of the mass parameter m.

Typically, structured uncertainty can be expected to be handled on a case by case basis, depending on the nature of the uncertainty. An approach to structured uncertainty is put forward by Doyle [1982]. For unstructured uncertainty, the situation is typically different as one considers a set of plant transfer functions $\mathscr{P}$ that are close to a nominal plant $\bar{R}_P$ in some way. Again the objective is to design a controller that works for every plant in the set of allowed plants.

4.18 Example **(Example 4.17 cont'd)** Let us look at the mass problem above in a different manner. Let us suppose that we choose a nominal plant $\bar{R}_P(s) = \frac{1/m}{s^2}$ and define a set of candidate plants

$$\mathscr{P} = \{R_P \in \mathbb{R}(s) \mid \|R_P - \bar{R}_P\|_\infty \leq \epsilon\} \tag{4.13}$$

for some $\epsilon > 0$. In this case, we have clearly allowed a much larger class of uncertainty that was allowed in Example 4.17. Indeed, not only is the mass no longer uncertain, even the form of the transfer function is uncertain. •

Thus we see that unstructured uncertainty generally forces us to consider a larger class of plants, and so is a more stringent and, therefore, conservative manner for modelling uncertainty. That is to say, by designing a controller that will work for *all* plants in $\mathscr{P}$, we are designing a controller for plants that are almost certainly *not* valid models for the plant under consideration. Nevertheless, it turns out that this conservatism of design is made up for by the admission of a consistent design methodology that goes along with unstructured uncertainty. We shall now turn our attention to describing how unstructured uncertainty may arise in examples.

4.5.2 Unstructured uncertainty models

We shall consider four unstructured uncertainty models, although others are certainly possible. Of the four types of uncertainty we present, only two will be treated in detail. A general account of uncertainty models is the subject of Chapter 8 in [Dullerud and Paganini 1999]. Consistent with our keep our treatment of robust control to a tolerable level of simplicity, we shall only look at rather straightforward types of uncertainty.

The first type we consider is called ***multiplicative uncertainty***. We start with a nominal plant $\bar{R}_P \in \mathbb{R}(s)$ and let $W_u \in \mathbb{R}(s)$ be a proper rational function with no poles in $\overline{\mathbb{C}}_+$. Denote by $\mathscr{P}_\times(\bar{R}_P, W_u)$ the set of rational functions $R_P \in \mathbb{R}(s)$ with the properties

1. R_P and $\bar{R}_P$ have the same number of poles in $\mathbb{C}_+$,
2. R_P and $\bar{R}_P$ have the same imaginary axis poles, and
3. $\left|\frac{R_P(i\omega)}{\bar{R}_P(i\omega)} - 1\right| \leq |W_u(i\omega)|$ for all $\omega \in \mathbb{R}$.

Another way to write this set of plants is to note that $R_P \in \mathscr{P}_\times(\bar{R}_P, W_u)$ if and only if

$$R_P = (1 + \Delta W_u)\bar{R}_P$$

where $\|\Delta\|_\infty \leq 1$. Of course, Δ is not arbitrary, even given $\|\Delta\|_\infty \leq 1$. Indeed, this condition only ensures condition 3 above. If Δ further satisfies the first two conditions, it is said to be ***allowable***. In this representation, it is perhaps more clear where the term multiplicative uncertainty comes up.

The following example is often used as one where multiplicative uncertainty is appropriate.

4.19 Example Recall from Exercise EE.6 that the transfer function for the time delay of a function g by T is $e^{-Ts}\hat{g}(s)$. Let us suppose that we have a plant transfer function

$$R_P(s) = e^{-Ts}R(s)$$

for some $R(s) \in \mathbb{R}(s)$. We wish to ensure that this plant is modelled by multiplicative uncertainty. To do so, we note that the first two terms in the Taylor series for e^{-Ts} are $1 - Ts$, and thus we suppose that when T is small, a nominal plant of the form

$$\bar{R}_P(s) = \frac{R(s)}{1 - Ts}$$

will do the job. Thus we are charged with finding a rational function W_u so that

$$\left|\frac{R_P(i\omega)}{\bar{R}(i\omega)} - 1\right| \leq |W_u(i\omega)|, \quad \omega \in \mathbb{R}.$$

Let us suppose that $T = \frac{1}{10}$. In this case, the condition on W_u becomes

$$|W_u(i\omega)| \geq \left|\frac{e^{-0.1i\omega}}{1 - \frac{s}{10}} - 1\right|, \quad \omega \in \mathbb{R}.$$

From Figure 4.18 (the solid curve) one can see that the magnitude of

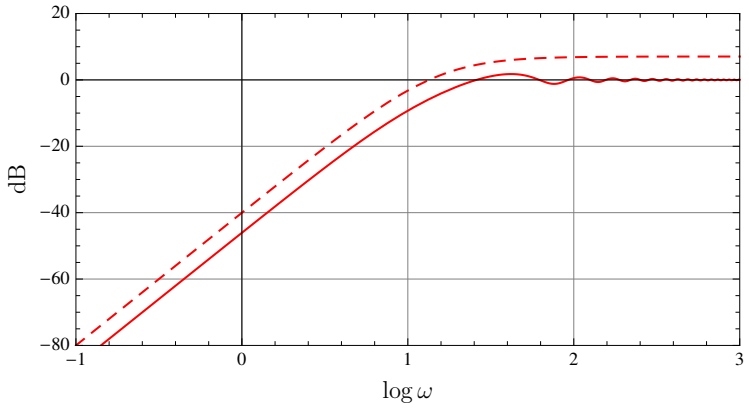


Figure 4.18 $\left|\frac{e^{-\frac{i\omega}{10}}}{1-\frac{s}{10}} - 1\right|$ (solid) and $W_u = \frac{\frac{1}{100}}{(1-\frac{\omega^2}{15})}$ (dashed)

$$\left|\frac{e^{-\frac{i\omega}{10}}}{1 - \frac{s}{10}} - 1\right|$$

has the rough behaviour of tailing off at 40dB/decade at low frequency and having constant magnitude at high frequency. Thus a model of the form

$$W_u(s) = \frac{Ks^2}{\tau s + 1}$$

is a likely candidate, although other possibilities will work as well, as long as they capture the essential features. Some fiddling with the Bode plot yields $\tau = \frac{1}{15}$ and $K = \frac{1}{100}$ as acceptable choices. Thus this choice of W_u will include the time delay plant in its set of plants. •

The next type of uncertainty we consider is called ***additive uncertainty***. Again we start with a nominal plant $\bar{R}_P$ and a proper $W_u \in \mathbb{R}(s)$ having no poles in $\overline{\mathbb{C}}_+$. Denote by $\mathscr{P}_\times(\bar{R}_P, W_u)$ the set of rational functions $R_P \in \mathbb{R}(s)$ with the properties

1. R_P and $\bar{R}_P$ have the same number of poles in $\mathbb{C}_+$,
2. R_P and $\bar{R}_P$ have the same imaginary axis poles, and
3. $|R_P(i\omega) - \bar{R}_P(i\omega)| \leq |W_u(i\omega)|$ for all $\omega \in \mathbb{R}$.

A plant in $\mathscr{P}_+(\bar{R}_P, W_u)$ will have the form

$$R_P = \bar{R}_P + \Delta W_u$$

where $\|\Delta\|_\infty \leq 1$. As with multiplicative uncertainty, Δ will be ***allowable*** if properties 1 and 2 above are met.

4.20 Example This example has a generic flavour. Suppose that we make measurements of our plant to get magnitude information about its frequency response at a finite set of frequencies $\{\omega_1, \ldots, \omega_k\}$. If the test is repeated at each frequency a number of times, we might try to find a nominal plant transfer function $\bar{R}_P$ with the property that at the measured frequencies its magnitude is roughly the average of the measured magnitudes. One then can determine W_u so that it covers the spread in the data at each of the measured frequencies. Such a W_u should have the property, by definition, that

$$|R_P(i\omega_j) - \bar{R}_P(i\omega_j)| \leq |W_u(i\omega_j)|, \quad j = 1, \ldots, k.$$

One could then hope that at the frequencies where data was not taken, the actual plant data is in the data of the set $\mathscr{P}_+(\bar{R}_P, W_u)$. •

The above two classes of uncertainty models, $\mathscr{P}_\times(\bar{R}_P, W_u)$ and $\mathscr{P}_+(\bar{R}_P, W_u)$ are the two for which analysis will be carried out in this book. The main reason for this choice is convenience of analysis. Fortunately, many interesting cases of plant uncertainty can be modelled as multiplicative or additive uncertainty. However, for completeness, we shall give two other types of uncertainty representations.

The first situation we consider where the set of plants are related to the nominal plant as

$$R_P = \frac{\bar{R}_P}{1 + \Delta W_u \bar{R}_P}, \quad \|\Delta\|_\infty \leq 1. \tag{4.14}$$

This uncertainty representation can arise in practice.

4.21 Example Suppose that we have a plant of the form

$$R_P(s) = \frac{1}{s^2 + as + 1}$$

where all we know is that $a \in [a_{\min}, a_{\max}]$. By taking

$$\bar{R}_P(s) = \frac{1}{s^2 + a_{\text{avg}} s + 1}, \quad W_u(s) = \tfrac{1}{2}\delta a s,$$

where $a_{\text{avg}} = \frac{1}{2}(a_{\max} + a_{\min})$ and $\delta a = a_{\max} - a_{\min}$, then the set of plant is exactly as in (4.14), if Δ is a number between -1 and 1. If Δ is allowed to be a rational function satisfying $\|\Delta\|_\infty \leq 1$ then we have embedded our actual set of plants inside the larger uncertainty set described by (4.14). One can view this example of one where the structured uncertainty is included as part of an unstructured uncertainty model. •

The final type of uncertainty model we present allows plants that are related to the nominal plant as

$$R_P = \frac{\bar{R}_P}{1 + \Delta W_u}, \quad \|\Delta\|_\infty \leq 1.$$

Let us see how this type of uncertainty can come up in an example.

4.22 Example • finish

4.23 Remarks

1. In our definitions of $\mathscr{P}_\times(\bar{R}_P, W_u)$ and $\mathscr{P}_+(\bar{R}_P, W_u)$ we made some assumptions about the poles of the nominal plant and the poles of the plants in the uncertainty set. The reason for these assumptions is not evident at this time, and indeed they can be relaxed with the admission of additional complexity of the results of Section 7.3, and those results that depend on these results. In practice, however, these assumptions are not inconvenient to account for.
2. All of our choices for uncertainty modelling share a common defect. They allow plants that will almost definitely *not* be possible models for our actual plant. That is to say, our sets of plants are very large. This has something of a drawback in our employment of these uncertainty models for controller design—they will lead to a too conservative design. However, this is mitigated by the existence of effective analysis tools to do robust controller design for such uncertainty models.
3. When deciding on a rational function W_u with which to model plant uncertainty with one of the above schemes, one typically will not want W_u to tend to zero as $s \to \infty$. The reason for this is that at higher frequencies is typically where model uncertainty is the greatest. This becomes a factor when choosing starting point for W_u. •

4.6 Summary

In this section we have introduced a nice piece of equipment—the frequency response function—and a pair of slick representations of it—the Bode plot and the polar plot. Here are the pertinent things you should know from this chapter.

1. For a given SISO linear system Σ, you should be able to compute Ω_Σ and H_Σ.
2. The interpretation of the frequency response given in Theorem 4.1 is fundamental in understanding just what it is that the frequency response means.
3. The complete equivalence of the impulse response, the transfer function, and the frequency response is important to realise. One should understand that something that one observes in one of these must also be reflected somehow in the other two.
4. The Bode plot as a representation of the frequency response is extremely important. Being able to look at it and understand things about the system in question is part of the "art" of control engineering, although there is certainly a lot of science in this too.
5. You should be able to draw by hand the Bode plot for any system whose numerator and denominator polynomials you can factor. Really the only subtle thing here is the dependence on ζ for the second-order components of the frequency response. If you think of ζ as damping, the interpretation here becomes straightforward since one expects larger magnitudes for lower damping when the second-order term is in the denominator.

6. The polar plot as a representation of the frequency response will be useful to us in Chapter 12. You should at least be able to sketch a polar plot given the corresponding Bode plot.
7. You should be aware of why minimum phase system have the name they do, and be able to identity these in "obvious" cases.
8. The developments of Section 4.4.2 are somehow essential, and at the same time somewhat hard. If one is to engage in controller design using frequency response, clearly the fact that there are essential restrictions on how the frequency response may behave is important.
9. It might be helpful on occasion to apply the approximation (4.12).

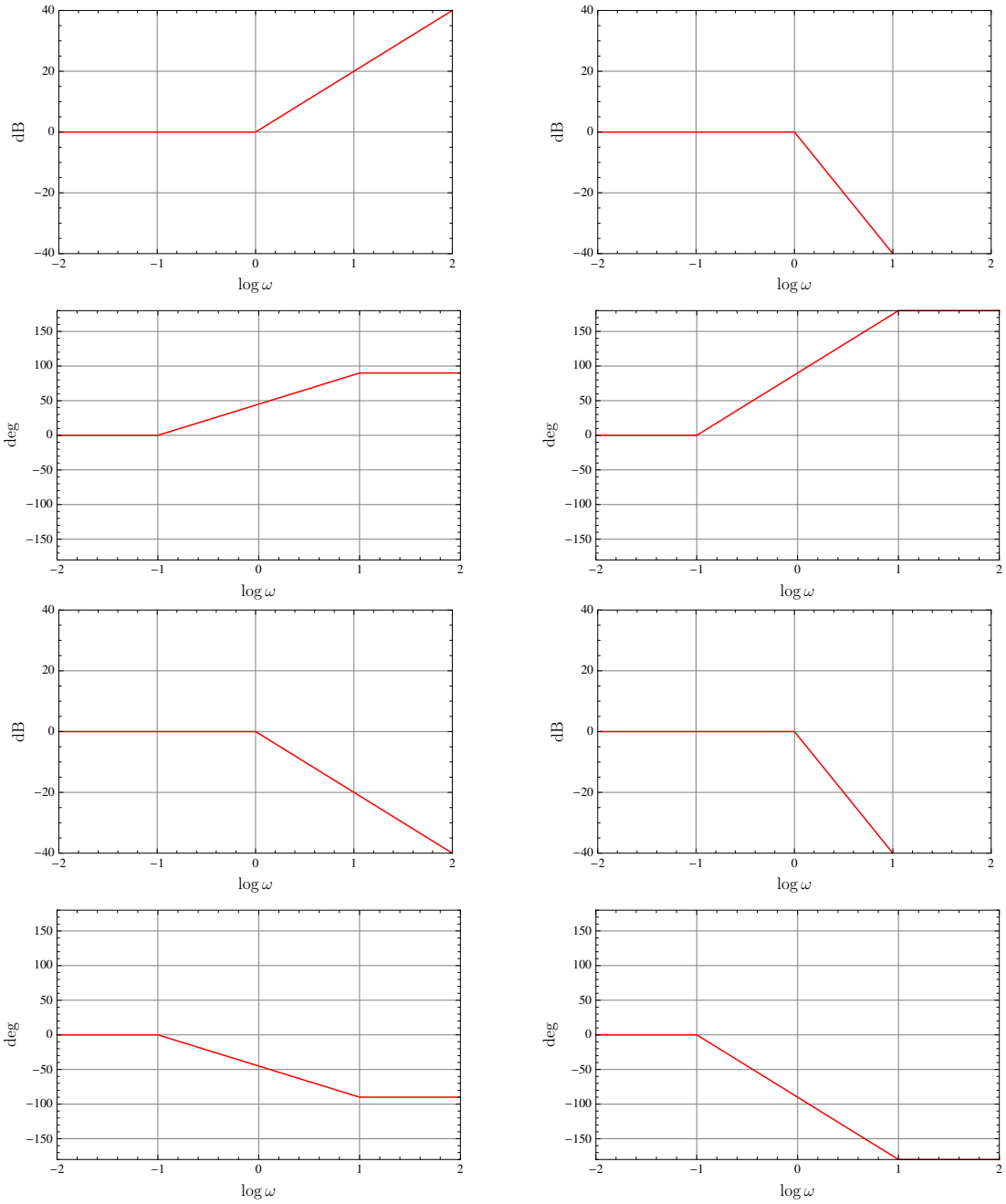


Figure 4.19 Asymptotes for magnitude and phase plots for $H(\omega) = 1 + i\omega$ (top left), $H(\omega) = 1 + 2i\zeta\omega - \omega^2$ (top right), $H(\omega) = (1+i\omega)^{-1}$ (bottom left), and $H(\omega) = (1+2i\zeta\omega-\omega^2)^{-1}$ (bottom right)

Exercises

E4.1 Consider the vector initial value problem

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + u_0 \sin \omega t \boldsymbol{b}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

where there are no eigenvalues for $\boldsymbol{A}$ of the form $i\tilde{\omega}$ where $\tilde{\omega}$ integrally divides ω. Let $\boldsymbol{x}_p(t)$ be the unique periodic solution constructed in the proof of Theorem 4.1, and

write the solution to the initial value problem as $\boldsymbol{x}(t) = \boldsymbol{x}_p(t) + \boldsymbol{x}_h(t)$. Determine an expression for $\boldsymbol{x}_h(t)$. Check that $\boldsymbol{x}_h(t)$ is a solution to the homogeneous equation.

E4.2 Consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ with

$$\boldsymbol{A} = \begin{bmatrix} \sigma_0 & \omega_0 \\ -\omega_0 & \sigma_0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

for $\sigma_0 \in \mathbb{R}$ and $\omega_0 > 0$.

(a) Determine Ω_Σ and compute H_Σ.

(b) Take $u(t) = u_0 \sin \omega t$, and determine when the system satisfies the hypotheses of Theorem 4.1, and determine the unique periodic output $y_p(t)$ guaranteed by that theorem.

(c) Do periodic solutions exist when the hypotheses of Theorem 4.1 are *not* satisfied?

(d) Plot the Bode plot for H_Σ for various values of $\sigma_0 \leq 0$ and $\omega_0 > 0$. Make sure you get all cases where the Bode plot assumes a different "character."

We have two essentially differing characterisations of the frequency response, one as the way in which sinusoidal outputs appear under sinusoidal inputs (cf. Theorem 4.1) and one involving Laplace and Fourier transforms (cf. Proposition 4.3). In the next exercise you will explore the differing domains of validity for the two interpretations.

E4.3 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be complete. Show that the Fourier transform of h_Σ^+ is defined if and only if all eigenvalues of $\boldsymbol{A}$ lie in $\mathbb{C}_-$. Does the characterisation of the frequency response provided in Theorem 4.1 share this restriction?

E4.4 For the SISO linear system $(N(s), D(s)) = (1, s + 1)$ in input/output form, verify explicitly the following statements, stating hypotheses on arguments of the functions where necessary.

(a) $T_{N,D}$ is the Laplace transform of $h_{N,D}$.

(b) $h_{N,D}$ is the inverse Laplace transform of $T_{N,D}$.

(c) $H_{N,D}$ is the Fourier transform of $h_{N,D}$.

(d) $h_{N,D}$ is the inverse Fourier transform of $H_{N,D}$.

(e) $T_{N,D}(s) = \dfrac{1}{2\pi} \displaystyle\int_{-\infty}^{\infty} \frac{H_{N,D}(\omega)}{s - i\omega} \, \mathrm{d}\omega$.

(f) $H_{N,D}(\omega) = T_{N,D}(i\omega)$.

When discussing the impulse response, it was declared that to obtain the output for an arbitrary input, one can use a convolution with the impulse response (plus a bit that depends upon initial conditions). In the next exercise, you will come to an understanding of this in terms of Fourier transforms.

E4.5 Suppose that $\check{f}$ and $\check{g}$ are functions of ω, and that they are the Fourier transforms of functions $f, g \colon (-\infty, \infty) \to \mathbb{R}$. You know from your course on transforms that the inverse Fourier transform of the product $\check{f}\check{g}$ is the convolution

$$(f * g)(t) = \int_{-\infty}^{\infty} f(t - \tau) g(\tau) \, \mathrm{d}\tau.$$

Now consider a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ with causal impulse response h_Σ^+. Recall that the output for zero state initial condition corresponding to the input $u \colon [0, \infty) \to \mathbb{R}$ is

$$y(t) = \int_0^t h_\Sigma^+(t - \tau) u(\tau) \, \mathrm{d}\tau.$$

Make sense of the statement, "The output y is equal to the convolution $h^+_\Sigma * u$." Note that you are essentially being asked to resolve the conflict in the limits of integration between convolution in Fourier and Laplace transforms.

E4.6 We will consider in detail here the mass-spring-damper system whose Bode plots are produced in Example 4.6. We begin by scaling the input u by the spring constant k so that, upon division by m, the governing equations are

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2 x = \omega_0^2 u,$$

where $\omega_0 = \sqrt{\frac{k}{m}}$ and $\zeta = \frac{d}{2\sqrt{km}}$. As output we shall take $y = x$. We allow d to be zero, but neither m nor k can be zero.

(a) Write this system as a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ (that is, determine $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$), and determine the transfer function T_Σ.

(b) Determine Ω_Σ for the various values of the parameters ω_0 and ζ, and then determine the frequency response H_Σ.

(c) Show that

$$|H_\Sigma(\omega)| = \frac{\omega_0^2}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4\zeta^2\omega_0^2\omega^2}}.$$

(d) How does $|H_\Sigma(\omega)|$ behave for $\omega \ll \omega_0$? for $\omega \gg \omega_0$?

(e) Show that $\frac{\mathrm{d}}{\mathrm{d}\omega}|H_\Sigma(\omega)| = 0$ if and only if $\frac{\mathrm{d}}{\mathrm{d}\omega}|H_\Sigma(\omega)|^2 = 0$.

(f) Using the previous simplification, show that for $\zeta < \frac{1}{\sqrt{2}}$ there is a maximum for $H_\Sigma(\omega)$. The maximum you determine should occur at the frequency

$$\omega_{\max} = \omega_0\sqrt{1 - 2\zeta^2},$$

and should take the value

$$|H_\Sigma(\omega_{\max})| = \frac{1}{2\zeta\sqrt{1 - \zeta^2}}.$$

(g) Show that

$$\measuredangle H_\Sigma(\omega) = \operatorname{atan2}(\omega_0^2 - \omega^2, -2\zeta\omega_0\omega),$$

where atan2 is the smart inverse tangent function that knows in which quadrant you are.

(h) How does $\measuredangle H_\Sigma(\omega)$ behave for $\omega \ll \omega_0$? for $\omega \gg \omega_0$?

(i) Determine an expression for $\measuredangle H_\Sigma(\omega_{\max})$.

(j) Use your work above to give an accurate sketch of the Bode plot for Σ in cases when $\zeta \leq \frac{1}{\sqrt{2}}$, making sure to mark on your plot all the features you determined in the previous parts of the question. What happens to the Bode plot as ζ is decreased? What would the Bode plot look like when $\zeta = 0$?

E4.7 Construct Bode plots by hand for the following first-order SISO linear systems in input/output form:

(a) $(N(s), D(s)) = (1, s + 1)$;

(b) $(N(s), D(s)) = (s, s + 1)$;

(c) $(N(s), D(s)) = (s - 1, s + 1)$;

(d) $(N(s), D(s)) = (2, s + 1)$.

E4.8 Construct Bode plots by hand for the following second-order SISO linear systems in input/output form:

(a) $(N(s), D(s)) = (1, s^2 + 2s + 2)$;
(b) $(N(s), D(s)) = (s, s^2 + 2s + 2)$;
(c) $(N(s), D(s)) = (s - 1, s^2 + 2s + 2)$;
(d) $(N(s), D(s)) = (s + 2, s^2 + 2s + 2)$.

E4.9 Construct Bode plots by hand for the following third-order SISO linear systems in input/output form:

(a) $(N(s), D(s)) = (1, s^3 + 3s^2 + 4s + 2)$;
(b) $(N(s), D(s)) = (s, s^3 + 3s^2 + 4s + 2)$;
(c) $(N(s), D(s)) = (s^2 - 4, s^3 + 3s^2 + 4s + 2)$;
(d) $(N(s), D(s)) = (s^2 + 1, s^3 + 3s^2 + 4s + 2)$.

E4.10 For each of the SISO linear systems given below do the following:

1. calculate the eigenvalues of $\boldsymbol{A}$;
2. calculate the transfer function;
3. sketch the Bode plots of the magnitude and phase of the frequency response;
4. by trial and error playing with the parameters, on your sketch, indicate the rôle played by the parameters of the system (e.g., ω_0 in parts (a) and (b));
5. try to justify the name of the system by looking at the shapes of the Bode plots.

Here are the systems (assume all parameters are positive):

(a) ***low-pass filter***:

$$\boldsymbol{A} = \begin{bmatrix} -\omega_0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} \omega_0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 0 \end{bmatrix};$$

(b) ***high-pass filter***:

$$\boldsymbol{A} = \begin{bmatrix} -\omega_0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} -\omega_0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 1 \end{bmatrix};$$

(c) ***notch filter***:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -\delta\omega_0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ -\delta\omega_0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 1 \end{bmatrix};$$

(d) ***bandpass filter***:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -\delta\omega_0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ \delta\omega_0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 0 \end{bmatrix}.$$

E4.11 Consider the coupled mass system of Exercises E1.4, E2.19, and E3.14. Assume no damping and take the input from Exercise E2.19 in the case when $\alpha = 0$. In Exercise E3.14, you constructed an output vector $\boldsymbol{c}_0$ for which the pair $(\boldsymbol{A}, \boldsymbol{c}_0)$ was unobservable.

(a) Construct a family of output vectors $\boldsymbol{c}_\epsilon$ by defining $\boldsymbol{c}_\epsilon = \boldsymbol{c}_0 + \epsilon \boldsymbol{c}_1$ for $\epsilon \in \mathbb{R}$ and $\boldsymbol{c}_1 \in \mathbb{R}^4$. Make sure you choose $\boldsymbol{c}_1$ so that $\boldsymbol{c}_\epsilon$ is observable for $\epsilon \neq 0$.
(b) Determine Ω_Σ, and the frequency response H_Σ using the output vector $\boldsymbol{c}_\epsilon$ (allow ϵ to be an arbitrary real number).

(c) Choose the parameters $m = 1$ and $k = 1$ and determine the Bode plot for values of ϵ around zero. Do you notice anything different in the character of the Bode plot when $\epsilon = 0$?

E4.12 Consider the pendulum/cart system of Exercises E1.5, E2.4, and E3.15. For each of the following linearisations:

(a) the equilibrium point $(0, 0)$ with cart position as output;
(b) the equilibrium point $(0, 0)$ with cart velocity as output;
(c) the equilibrium point $(0, 0)$ with pendulum angle as output;
(d) the equilibrium point $(0, 0)$ with pendulum angular velocity as output;
(e) the equilibrium point $(0, \pi)$ with cart position as output;
(f) the equilibrium point $(0, \pi)$ with cart velocity as output;
(g) the equilibrium point $(0, \pi)$ with pendulum angle as output;
(h) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output,

do the following:

1. determine Ω_Σ, and the frequency response H_Σ for the system;
2. for parameters $M = 1\frac{1}{2}$, $m = 1$, $g = 9.81$, and $\ell = \frac{1}{2}$, produce the Bode plot for the pendulum/cart system;
3. can you see reflected in your Bode plot the character of the spectrum of the zero dynamics as you determined in Exercise E3.15?

E4.13 Consider the double pendulum system of Exercises E1.6, E2.5, E1.6 and E3.16. For each of the following linearisations:

(a) the equilibrium point $(0, 0, 0, 0)$ with the pendubot input;
(b) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;
(c) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;
(d) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;
(e) the equilibrium point $(0, 0, 0, 0)$ with the acrobot input;
(f) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;
(g) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;
(h) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input,

do the following:

1. determine Ω_Σ, and the frequency response H_Σ for the system;
2. for parameters $m_1 = 1$, $m_2 = 2$, $\ell_1 = \frac{1}{2}$, and $\ell_2 = \frac{1}{3}$, produce the Bode plot for the double pendulum;
3. can you see reflected in your Bode plot the character of the spectrum of the zero dynamics as you determined in Exercise E3.16?

In each case, use the angle of the second link as output.

E4.14 Consider the coupled tank system of Exercises E1.11, E2.6, and E3.17. For the linearisations in the following cases:

(a) the output is the level in tank 1;
(b) the output is the level in tank 2;
(c) the output is the difference in the levels,

do the following:

1. determine Ω_Σ, and the frequency response H_Σ for the system;

2. for parameters $\alpha = \frac{1}{3}$, $\delta_1 = 1$, $A_1 = 1$, $A_2 = \frac{1}{2}$, $a_1 = \frac{1}{10}$, $a_2 = \frac{1}{20}$, and $g = 9.81$, produce the Bode plot for the tank system;
3. can you see reflected in your Bode plot the character of the spectrum of the zero dynamics as you determined in Exercise E3.17?

E4.15 Suppose you are shown a Bode plot for a stable SISO linear system Σ. Can you tell from the character of the plot whether Σ is controllable? observable? minimum phase?

E4.16 Construct two transfer functions, one minimum phase and the other not, whose frequency response magnitudes are the same (you cannot use the one in the book). Make Bode plots for each system, and make the relevant observations.

E4.17 Let (N, D) be a proper, stable SISO linear system of relative degree m. Characterise the total phase shift,

$$\angle H_{N,D}(\infty) - \angle H_{N,D}(0),$$

in terms of the roots of N, assuming that N has no roots on $\mathrm{i}\mathbb{R}$.

In the next exercise, you will provide a rigorous justification for the term "minimum phase."

E4.18 Let (N, D) be a SISO linear system in input/output form, and suppose that it is minimum phase (thus N has no roots in $\mathbb{C}_+$). Let $M(N, D)$ denote the collection of SISO linear systems $(\tilde{N}, \tilde{D})$ in input/output form for which

$$|H_{N,D}(\omega)| = |H_{\tilde{N},\tilde{D}}(\omega)|, \quad \omega \in \mathbb{R}.$$

Thus the magnitude Bode plots for all systems in $M(N, D)$ are exactly the magnitude Bode plot for (N, D).

(a) How are systems in $M(N, D)$ related to (N, D)? This boils down, of course, to identifying SISO linear systems that have a plot magnitude of 1 at all frequencies.

For $(\tilde{N}, \tilde{D}) \in M(N, D)$ denote $\phi_0(\tilde{N}, \tilde{D}) = \lim_{\omega \to 0} \angle H_{\tilde{N},\tilde{D}}(\omega)$ and $\phi_\infty(\tilde{N}, \tilde{D}) = \lim_{\omega \to \infty} \angle H_{\tilde{N},\tilde{D}}(\omega)$. Now let

$$\Delta\phi(\tilde{N}, \tilde{D}) = \phi_\infty(\tilde{N}, \tilde{D}) - \phi_0(\tilde{N}, \tilde{D}).$$

With this notation prove the following statement.

(b) $\Delta\phi(N, D) \leq \Delta\phi(\tilde{N}, \tilde{D})$ for any $(\tilde{N}, \tilde{D}) \in M(N, D)$.

E4.19 Construct polar plots corresponding to the Bode plots you made in Exercise E4.7.

E4.20 Construct polar plots corresponding to the Bode plots you made in Exercise E4.8.

E4.21 Construct polar plots corresponding to the Bode plots you made in Exercise E4.9.

E4.22 For the SISO linear system Σ of Exercise E4.2, plot the polar plots for various $\sigma_0 \leq 0$ and $\omega_0 > 0$.

It is not uncommon to encounter a scheme for control design that relies on a plant being stable, i.e., having all poles in $\mathbb{C}_-$. This is in conflict with many plants—for example, the simple mass— that have a pole at $s = 0$. In the next exercise, you will investigate a commonly employed hack to get around plants having poles at $s = 0$.

E4.23 The transfer function

$$T(s) = \frac{K}{\tau s + 1}$$

is put forward as providing a "stable approximation to $\frac{1}{s}$."

(a) Comment on how fiddling K and τ, particularly τ, render the approximation better or worse.

(b) Explain in what sense this approximation is valid, and also how it is invalid.

E4.24 For the following SISO linear systems in input/output form, determine whether they satisfy the hypotheses of Bode's Gain/Phase Theorem. If the system does satisfy the hypotheses, verify explicitly that the theorem holds, and determine how good is the approximation (4.12) at various frequencies. If the system does not satisfy the hypotheses, determine explicitly whether the theorem does in fact hold.

(a) $(N(s), D(s)) = (1, s)$.

(b) $(N(s), D(s)) = (s, 1)$.

(c) $(N(s), D(s)) = (1, s + 1)$.

(d) $(N(s), D(s)) = (1, s - 1)$.

(e) $(N(s), D(s)) = (s + 1, 1)$.

(f) $(N(s), D(s)) = (s - 1, 1)$.

E4.25 Exercise on the approximate version of Bode's Gain/Phase Theorem. complete

Part II

System analysis

Chapter 5

Stability of control systems

We will be trying to stabilise unstable systems, or to make an already stable system even more stable. Although the essential goals are the same for each class of system we have encountered thus far (i.e., SISO linear systems and SISO linear systems in input/output form), each has its separate issues. We first deal with these. In each case, we will see that stability boils down to examining the roots of a polynomial. In Section 5.5 we give algebraic criteria for determining when the roots of a polynomial all lie in the negative complex plane.

Contents

5.1 Internal stability

Internal stability is a concept special to SISO linear systems i.e., those like

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t). \end{aligned} \tag{5.1}$$

Internal stability refers to the stability of the system without our doing anything with the controls. We begin with the definitions.

5.1 Definition A SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is

(i) ***internally stable*** if

$$\limsup_{t\to\infty} \|\boldsymbol{x}(t)\| < \infty$$

for every solution $\boldsymbol{x}(t)$ of $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$;

(ii) ***internally asymptotically stable*** if

$$\lim_{t\to\infty} \|\boldsymbol{x}(t)\| = 0$$

for every solution $\boldsymbol{x}(t)$ of $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$;

(iii) ***internally unstable*** if it is not internally stable. •

Of course, internal stability has nothing to do with any part of Σ other than the matrix $\boldsymbol{A}$. If one has a system that is subject to the problems we discussed in Section 2.3, then one may want to hope the system at hand is one that is internally stable. Indeed, all the bad behaviour we encountered there was a direct result of my intentionally choosing systems that were *not* internally stable—it served to better illustrate the problems that can arise.

Internal stability can almost be determined from the spectrum of $\boldsymbol{A}$. The proof of the following result, although simple, relies on the structure of the matrix exponential as we discussed in Section B.2. We also employ the notation

$$\begin{aligned} \mathbb{C}_- &= \{z \in \mathbb{C} \mid \operatorname{Re}(z) < 0\}, \quad \mathbb{C}_+ = \{z \in \mathbb{C} \mid \operatorname{Re}(z) > 0\}, \\ \overline{\mathbb{C}}_- &= \{z \in \mathbb{C} \mid \operatorname{Re}(z) \leq 0\}, \quad \overline{\mathbb{C}}_+ = \{z \in \mathbb{C} \mid \operatorname{Re}(z) \geq 0\}, \\ i\mathbb{R} &= \{z \in \mathbb{C} \mid \operatorname{Re}(z) = 0\}. \end{aligned}$$

With this we have the following result, recalling notation concerning eigenvalues and eigenvectors from Section A.5.

5.2 Theorem *Consider a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$. The following statements hold.*

(i) *Σ is internally unstable if $\operatorname{spec}(\boldsymbol{A}) \cap \mathbb{C}_+ \neq \emptyset$.*

(ii) *Σ is internally asymptotically stable if $\operatorname{spec}(\boldsymbol{A}) \subset \mathbb{C}_-$.*

(iii) *Σ is internally stable if $\operatorname{spec}(\boldsymbol{A}) \cap \mathbb{C}_+ = \emptyset$ and if $m_g(\lambda) = m_a(\lambda)$ for $\lambda \in \operatorname{spec}(\boldsymbol{A}) \cap (i\mathbb{R})$.*

(iv) *Σ is internally unstable if $m_g(\lambda) < m_a(\lambda)$ for $\lambda \in \operatorname{spec}(\boldsymbol{A}) \cap (i\mathbb{R})$.*

Proof (i) In this case there is an eigenvalue $\alpha + i\omega \in \mathbb{C}_+$ and a corresponding eigenvector $\boldsymbol{u} + i\boldsymbol{v}$ which gives rise to real solutions

$$\boldsymbol{x}_1(t) = e^{\alpha t}(\cos \omega t \boldsymbol{u} - \sin \omega t \boldsymbol{v}), \quad \boldsymbol{x}_2(t) = e^{\alpha t}(\sin \omega t \boldsymbol{u} + \cos \omega t \boldsymbol{v}).$$

Clearly these solutions are unbounded as $t \to \infty$ since $\alpha > 0$.

(ii) If all eigenvalues lie in $\mathbb{C}_-$, then any solution of $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$ will be a linear combination of n linearly independent vector functions of the form

$$t^k e^{-\alpha t}\boldsymbol{u} \quad \text{or} \quad t^k e^{-\alpha t}(\cos \omega t \boldsymbol{u} - \sin \omega t \boldsymbol{v}) \quad \text{or} \quad t^k e^{-\alpha t}(\sin \omega t \boldsymbol{u} + \cos \omega t \boldsymbol{v}) \tag{5.2}$$

for $\alpha > 0$. Note that all such functions tend in length to zero as $t \to \infty$. Suppose that we have a collection $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n(t)$ of such vector functions. Then, for any solution $\boldsymbol{x}(t)$ we have,

for some constants $c_1, \dots, c_n$,

$$\begin{aligned}\lim_{t\to\infty}\|\boldsymbol{x}(t)\| &= \lim_{t\to\infty}\|c_1\boldsymbol{x}_1(t) + \cdots + c_n\boldsymbol{x}_n(t)\| \\ &\le |c_1|\lim_{t\to\infty}\|\boldsymbol{x}_1(t)\| + \cdots + |c_n|\lim_{t\to\infty}\|\boldsymbol{x}_n(t)\| \\ &= 0,\end{aligned}$$

where we have used the triangle inequality, and the fact that the solutions $\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_n(t)$ all tend to zero as $t \to \infty$.

(iii) If $\operatorname{spec}(\boldsymbol{A}) \cap \mathbb{C}_+ = \emptyset$ and if further $\operatorname{spec}(\boldsymbol{A}) \subset \mathbb{C}_-$, then we are in case (ii), so Σ is internally asymptotically stable, and so internally stable. Thus we need only concern ourselves with the case when we have eigenvalues on the imaginary axis. In this case, provided all such eigenvalues have equal geometric and algebraic multiplicities, all solutions will be linear combinations of functions like those in (5.2) or functions like

$$\sin\omega t\boldsymbol{u} \quad \text{or} \quad \cos\omega t\boldsymbol{u}. \tag{5.3}$$

Let $\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_\ell(t)$ be ℓ linearly independent functions of the form (5.2), and let $\boldsymbol{x}_{\ell+1}(t), \dots, \boldsymbol{x}_n(t)$ be linearly independent functions of the form (5.3) so that $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ forms a set of linearly independent solutions for $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$. Thus we will have, for some constants $c_1, \dots, c_n$,

$$\begin{aligned}\limsup_{t\to\infty}\|\boldsymbol{x}(t)\| &= \limsup_{t\to\infty}\|c_1\boldsymbol{x}_1(t) + \cdots + c_n\boldsymbol{x}_n(t)\| \\ &\le |c_1|\limsup_{t\to\infty}\|\boldsymbol{x}_1(t)\| + \cdots + |c_\ell|\limsup_{t\to\infty}\|\boldsymbol{x}_\ell(t)\| + \\ &\quad |c_{\ell+1}|\limsup_{t\to\infty}\|\boldsymbol{x}_{\ell+1}(t)\| + \cdots + |c_n|\limsup_{t\to\infty}\|\boldsymbol{x}_n(t)\| \\ &= |c_{\ell+1}|\limsup_{t\to\infty}\|\boldsymbol{x}_{\ell+1}(t)\| + \cdots + |c_n|\limsup_{t\to\infty}\|\boldsymbol{x}_n(t)\|.\end{aligned}$$

Since each of the terms $\|\boldsymbol{x}_{\ell+1}(t)\|, \dots, \|\boldsymbol{x}_n(t)\|$ are bounded, their lim sup's will exist, which is what we wish to show.

(iv) If $\boldsymbol{A}$ has an eigenvalue $\lambda = i\omega$ on the imaginary axis for which $m_g(\lambda) < m_a(\lambda)$ then there will be solutions of $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$ that are linear combinations of vector functions of the form $t^k \sin\omega t\boldsymbol{u}$ or $t^k \cos\omega t\boldsymbol{v}$. Such functions are unbounded as $t \to \infty$, and so Σ is internally unstable. ■

5.3 Remarks

1. A matrix $\boldsymbol{A}$ is ***Hurwitz*** if $\operatorname{spec}(\boldsymbol{A}) \subset \mathbb{C}_-$. Thus $\boldsymbol{A}$ is Hurwitz if and only if $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is internally asymptotically stable.
2. We see that internal stability is almost completely determined by the eigenvalues of $\boldsymbol{A}$. Indeed, one says that Σ is ***spectrally stable*** if $\boldsymbol{A}$ 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 internal stability.
3. One does not generally want to restrict oneself to systems that are internally stable. Indeed, one often wants to stabilise an unstable system with feedback. In Theorem 6.49 we shall see, in fact, that for controllable systems it is *always* possible to choose a "feedback vector" that makes the "closed-loop" system internally stable. •

The notion of internal stability is in principle an easy one to check, as we see from an example.

5.4 Example We look at a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix}.$$

The form of $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$ does not concern us when talking about internal stability. The eigenvalues of $\boldsymbol{A}$ are the roots of the characteristic polynomial $s^2 + as + b$, 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 internal 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 internally stable, but not internally asymptotically stable.
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 stable, but not asymptotically stable.
5. $a > 0$ and $b > 0$: One may readily ascertain (in Section 5.5 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 internally asymptotically stable.
6. $a > 0$ and $b < 0$: In this case both eigenvalues are real, one being positive and the other negative. Thus the system is internally 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 internal 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 internally unstable. •

Note that one cannot really talk about internal stability for a SISO linear system (N, D) in input/output form. After all, systems in input/output form do not have built into them a notion of state, and internal stability has to do with states. In principle, one could define the internal stability for a proper system as internal stability for $\Sigma_{N,D}$, but this is best handled by talking directly about input/output stability which we now do.

5.2 Input/output stability

We shall primarily be interested in this course in input/output stability. That is, we want nice inputs to produce nice outputs. In this section we demonstrate that this property is

intimately related with the properties of the impulse response, and therefore the properties of the transfer function.

5.2.1 BIBO stability of SISO linear systems

We begin by talking about input/output stability in the context of SISO linear systems. When we have understood this, it is a simple matter to talk about SISO linear systems in input/output form.

5.5 Definition A SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is ***bounded input, bounded output stable*** (***BIBO stable***) if there exists a constant $K > 0$ so that the conditions (1) $\boldsymbol{x}(0) = \boldsymbol{0}$ and (2) $|u(t)| \leq 1$, $t \geq 0$ imply that $y(t) \leq K$ where $u(t)$, $\boldsymbol{x}(t)$, and $y(t)$ satisfy (5.1). •

Thus BIBO stability is our way of saying that a bounded input will produce a bounded output. You can show that the formal definition means exactly this in Exercise E5.8.

The following result gives a concise condition for BIBO stability in terms of the impulse response.

5.6 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system and define $\tilde{\Sigma} = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$. Then Σ is BIBO stable if and only if $\lim_{t\to\infty}|h_{\tilde{\Sigma}}(t)| = 0$.*

Proof Suppose that $\lim_{t\to\infty}|h_{\tilde{\Sigma}}(t)| \neq 0$. Then, by Proposition 3.24, it must be the case that either (1) $h_{\tilde{\Sigma}}(t)$ blows up exponentially as $t \to \infty$ or that (2) $h_{\tilde{\Sigma}}(t)$ is a sum of terms, one of which is of the form $\sin \omega t$ or $\cos \omega t$. For the first case we can take the bounded input $u(t) = 1(t)$. Using Proposition 2.32 and Proposition 3.24 we can then see that

$$y(t) = \int_0^\infty h_{\tilde{\Sigma}}(t - \tau)\, \mathrm{d}\tau + \boldsymbol{D}u(t).$$

Since $h_{\tilde{\Sigma}}(t)$ blows up exponentially, so too will $y(t)$ if it is so defined. Thus the bounded input $u(t) = 1(t)$ produces an unbounded output. For case (2) we choose $u(t) = \sin \omega t$ and compute

$$\int_0^t \sin \omega(t - \tau) \sin \omega\tau \, \mathrm{d}\tau = \tfrac{1}{2}\big(\tfrac{1}{\omega} \sin \omega t - t \cos \omega t\big), \quad \int_0^t \cos \omega(t - \tau) \sin \omega\tau \, \mathrm{d}\tau = \tfrac{1}{2} t \sin \omega t.$$

Therefore, $y(t)$ will be unbounded for the bounded input $u(t) = \sin \omega t$. We may then conclude that Σ is not BIBO stable.

Now suppose that $\lim_{t\to\infty}|h_{\tilde{\Sigma}}(t)| = 0$. By Proposition 3.24 this means that $h_{\tilde{\Sigma}}(t)$ dies off exponentially fast as $t \to \infty$, and therefore we have a bound like

$$\int_0^\infty |h_{\tilde{\Sigma}}(t - \tau)|\, \mathrm{d}\tau \leq M$$

for some $M > 0$. Therefore, whenever $u(t) \leq 1$ for $t \geq 0$, we have

$$\begin{aligned}
|y(t)| &= \Big| \int_0^t h_{\tilde{\Sigma}}(t - \tau)u(\tau)\, \mathrm{d}\tau + \boldsymbol{D}u(t) \Big| \\
&\leq \int_0^t |h_{\tilde{\Sigma}}(t - \tau)u(\tau)|\, \mathrm{d}\tau + |\boldsymbol{D}| \\
&\leq \int_0^t |h_{\tilde{\Sigma}}(t - \tau)||u(\tau)|\, \mathrm{d}\tau + |\boldsymbol{D}| \\
&\leq \int_0^t |h_{\tilde{\Sigma}}(t - \tau)|\, \mathrm{d}\tau + |\boldsymbol{D}| \\
&\leq M + |\boldsymbol{D}|.
\end{aligned}$$

This means that Σ is BIBO stable. ■

This result gives rise to two easy corollaries, the first following from Proposition 3.24, and the second following from the fact that if the real part of all eigenvalues of $\boldsymbol{A}$ are negative then $\lim_{t\to\infty}|h_\Sigma(t)| = 0$.

5.7 Corollary *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system and write*

$$T_\Sigma(s) = \frac{N(s)}{D(s)}$$

where (N, D) is the c.f.r. Then Σ is BIBO stable if and only if D has roots only in the negative half-plane.

5.8 Corollary $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ *is BIBO stable if* $\operatorname{spec}(\boldsymbol{A}) \subset \mathbb{C}_-$.

The matter of testing for BIBO stability of a SISO linear system is straightforward, so let's do it for a simple example.

5.9 Example **(Example 5.4 cont'd)** We continue with the case where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix},$$

and we now add the information

$$\boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1.$$

We compute

$$T_\Sigma(s) = \frac{1}{s^2 + as + b}.$$

From Example 5.4 we know that we have BIBO stability if and only if $a > 0$ and $b > 0$.

Let's probe the issue a bit further by investigating what actually happens when we do not have $a, b > 0$. The cases when Σ is internally unstable are not altogether interesting since the system is "obviously" not BIBO stable in these cases. So let us examine the cases when we have no eigenvalues in $\mathbb{C}_+$, but at least one eigenvalue on the imaginary axis.

1. $a = 0$ and $b > 0$: Here the eigenvalues are $\pm i\sqrt{b}$, and we compute

$$h_\Sigma(t) = \frac{\sin\sqrt{b}t}{\sqrt{b}}.$$

Thus the impulse response is bounded, but does not tend to zero as $t \to \infty$. Theorem 5.6 predicts that there will be a bounded input signal that produces an unbounded output signal. In fact, if we choose $u(t) = \sin\sqrt{b}t$ and zero initial condition, then one verifies that the output is

$$y(t) = \int_0^t \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{b}\sin(\sqrt{b}\tau)\,\mathrm{d}\tau = \frac{\sin(\sqrt{b}t)}{2b} - \frac{t\cos(\sqrt{b}t)}{2\sqrt{b}}.$$

Thus a bounded input gives an unbounded output.

2. $a > 0$ and $b = 0$: The eigenvalues here are $\{0, -a\}$. One may determine the impulse response to be
$$h_\Sigma(t) = \frac{1 - e^{-at}}{a}.$$
This impulse response is bounded, but again does not go to zero as $t \to \infty$. Thus there ought to be a bounded input that gives an unbounded output. We have a zero eigenvalue, so this means we should choose a constant input. We take $u(t) = 1$ and zero initial condition and determine the output as
$$y(t) = \int_0^t \boldsymbol{c}^t e^{\boldsymbol{A}(t-\tau)} \boldsymbol{b} \, \mathrm{d}\tau = \frac{t}{a} - \frac{1 - e^{-at}}{a^2}.$$
Again, a bounded input provides an unbounded output. •

As usual, when dealing with input/output issues for systems having states, one needs to exercise caution for the very reasons explored in Section 2.3. This can be demonstrated with an example.

5.10 Example Let us choose here a specific example (i.e., one without parameters) that will illustrate problems that can arise with fixating on BIBO stability while ignoring other considerations. We consider the system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ with
$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$
We determine that $h_\Sigma(t) = -e^{-2t}$. From Theorem 5.6 we determine that Σ is BIBO stable.

But is everything *really* okay? Well, no, because this system is actually not observable. We compute
$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix},$$
and since this matrix has rank 1 the system is not observable. How is this manifested in the system behaviour? In exactly the way one would predict. Thus let us look at the state behaviour for the system with a bounded input. We take $u(t) = 1(t)$ as the unit step input, and take the zero initial condition. The resulting state behaviour is defined by
$$\boldsymbol{x}(t) = \int_0^t e^{\boldsymbol{A}(t-\tau)} \boldsymbol{b} \, \mathrm{d}\tau = \begin{bmatrix} \frac{1}{3}e^t + \frac{1}{6}e^{-t} - \frac{1}{2} \\ \frac{1}{3}e^t - \frac{1}{3}e^{-2t} \end{bmatrix}.$$
We see that the state is behaving poorly, even though the output may be determined as
$$y(t) = \tfrac{1}{2}(e^{-2t} - 1),$$
which is perfectly well-behaved. But we have seen this sort of thing before. •

Let us state a result that provides a situation where one can make a precise relationship between internal and BIBO stability.

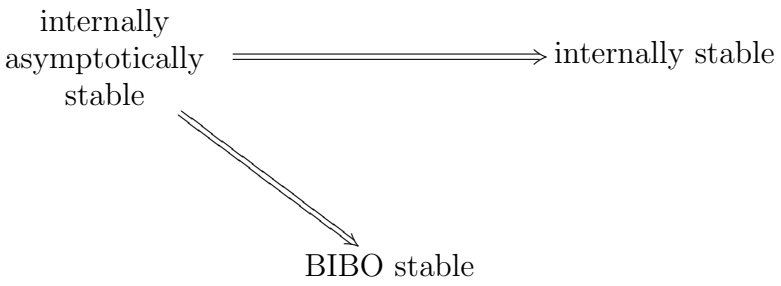


Figure 5.1 Summary of various stability types for SISO linear systems

5.11 Proposition *If a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}', \boldsymbol{D})$ is controllable and observable, then the following two statements are equivalent:*

(i) *Σ is internally asymptotically stable;*

(ii) *Σ is BIBO stable.*

Proof When Σ is controllable and observable, the poles of T_Σ are *exactly* the eigenvalues of $\boldsymbol{A}$. ■

When Σ is not both controllable and observable, the situation is more delicate. The diagram in Figure 5.1 provides a summary of the various types of stability, and which types imply others. Note that there are not many arrows in this picture. Indeed, the only type of stability which implies all others is internal asymptotic stability. This does *not* mean that if a system is only internally stable or BIBO stable that it is *not* internally asymptotically stable. It only means that one cannot generally infer internal asymptotic stability from internal stability or BIBO stability. What's more, when a system is internally stable but not internally asymptotically stable, then one can make some negative implications, as shown in Figure 5.2. Again, one should be careful when interpreting the absence of arrows from this

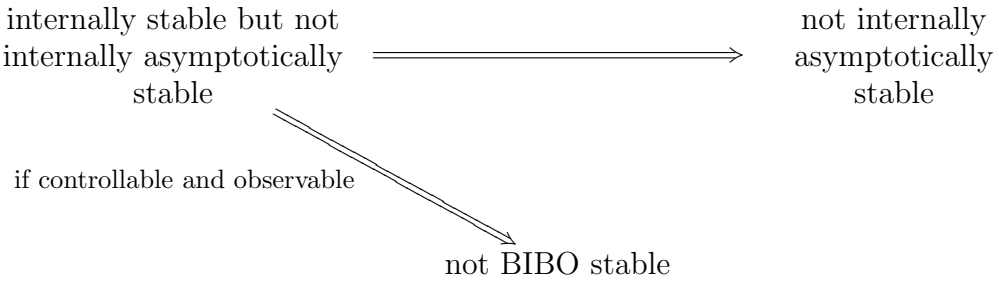


Figure 5.2 Negative implication when a system is internally stable, but not internally asymptotically stable

diagram. The best approach here is to understand that there are principles that underline when one can infer one type of stability from another. If these principles are understood, then matters are quite straightforward. A clearer resolution of the connection between BIBO stability and internal stability is obtained in Section 10.1 when we introduce the concepts of "stabilisability" and "detectability." The complete version of Figures 5.1 and 5.2 is given by Figure 10.1. Note that we have some water to put under the bridge to get there...

5.2.2 BIBO stability of SISO linear systems in input/output form

It is now clear how we may extend the above discussion of BIBO stability to systems in input/output form, at least when they are proper.

5.12 Definition A proper SISO linear system (N, D) in input/output form is ***bounded input, bounded output stable*** (***BIBO stable***) if the SISO linear system $\Sigma_{N,D}$ is BIBO stable. •

From Corollary 5.7 follows the next result giving necessary and sufficient conditions for BIBO stability of strictly proper SISO systems in input/output form.

5.13 Proposition *A proper SISO linear system (N, D) in input/output form is BIBO stable if and only if $T_{N,D}$ has no poles in $\overline{\mathbb{C}}_+$.*

The question then arises, "What about SISO linear systems in input/output form that are *not* proper?" Well, such systems can readily be shown to be not BIBO stable, no matter what the character of the denominator polynomial D. The following result shows why this is the case.

5.14 Proposition *If (N, D) is an improper SISO linear system in input/output form, then there exists an input u satisfying the properties*

(i) $|u(t)| \leq 1$ *for all* $t \geq 0$ *and*

(ii) if y satisfies $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$, then for any $M > 0$ there exists $t > 0$ so that $|y(t)| > M$.

Proof From Theorem C.6 we may write

$$\frac{N(s)}{D(s)} = R(s) + P(s)$$

where R is a strictly proper rational function and P is a polynomial of degree at least 1. Therefore, for any input u, the output y will be a sum $y = y_1 + y_2$ where

$$\hat{y}_1(s) = R(s)\hat{u}(s), \quad \hat{y}_2(s) = P(s)\hat{u}(s). \tag{5.4}$$

If R has poles in $\overline{\mathbb{C}}_+$, then the result follows in usual manner of the proof of Theorem 5.6. So we may as well suppose that R has no poles in $\overline{\mathbb{C}}_+$, so that the solution y_1 is bounded. We will show, then, that y_2 is not bounded. Let us choose $u(t) = \sin(t^2)$. Any derivative of u will involve terms polynomial in t and such terms will not be bounded as $t \to \infty$. But y_2, by (5.4), is a linear combination of derivatives of u, so the result follows. ■

5.3 Norm interpretations of BIBO stability

In this section, we offer interpretations of the stability characterisations of the previous section in terms of various norms for transfer functions and for signals. The material in this section will be familiar to those who have had a good course in signals and systems. However, it is rare that the subject be treated in the manner we do here, although its value for understanding control problems is now well established.

5.3.1 Signal norms

We begin our discussion by talking about ways to define the "size" of a signal. The development in this section often is made in a more advanced setting where the student is assumed to have some background in measure theory. However, it is possible to get across the basic ideas in the absence of this machinery, and we try to do this here.

For $p \geq 1$ and for a function $f\colon (-\infty, \infty) \to \mathbb{R}$ denote

$$\|f\|_p = \Big(\int_{-\infty}^{\infty} |f(t)|^p \, \mathrm{d}t\Big)^{1/p}$$

which we call the $\mathbf{L_p}$***-norm*** of y. Denote

$$\mathrm{L}_p(-\infty, \infty) = \big\{f\colon (-\infty, \infty) \to \mathbb{R} \;\big|\; \|f\|_p < \infty\big\}.$$

Functions in $\mathrm{L}_p(-\infty, \infty)$ are said to be $\mathbf{L_p}$***-integrable***. The case where $p = \infty$ is handled separately by defining

$$\|f\|_\infty = \sup_{\alpha \geq 0}\{|f(t)| \leq \alpha \text{ for almost every } t\}$$

as the $\mathbf{L_\infty}$***-norm*** of y. The L_∞-norm is sometimes referred to as the ***sup norm***. Here "almost every" means except on a set $T \subset (-\infty, \infty)$ having the property that

$$\int_T \mathrm{d}t = 0.$$

We denote

$$\mathrm{L}_\infty(-\infty, \infty) = \{f\colon (-\infty, \infty) \to \mathbb{R} \mid \|f\|_\infty < \infty\}$$

as the set of functions that we declare to be $\mathbf{L_\infty}$***-integrable***. Note that we are dealing here with functions defined on $(-\infty, \infty)$, whereas with control systems, one most often has functions that are defined to be zero for $t < 0$. This is still covered by what we do, and the extra generality is convenient.

Most interesting to us will be the L_p spaces $\mathrm{L}_2(-\infty, \infty)$ and $\mathrm{L}_\infty(-\infty, \infty)$. The two sets of functions certainly do not coincide, as the following collection of examples indicate.

5.15 Examples

1. The function $\cos t$ is in $\mathrm{L}_\infty(-\infty, \infty)$, but is in none of the spaces $\mathrm{L}_p(-\infty, \infty)$ for $1 \leq p < \infty$. In particular, it is not L_2-integrable.
2. The function $f(t) = \frac{1}{1+t}$ is not L_1-integrable, although it is L_2-integrable; one computes $\|f\|_2 = 1$.
3. Define
$$f(t) = \begin{cases} \sqrt{\frac{1}{t}}, & t \in (0, 1] \\ 0, & \text{otherwise.} \end{cases}$$
One then checks that $\|f\|_1 = 2$, but that f is not L_2-integrable. Also, since $\lim_{t \to 1_-} f(t) = \infty$, the function is not L_∞-integrable.
4. Define
$$f(t) = \begin{cases} \ln t, & t \in (0, 1] \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\lim_{t\to 0_+} f(t) = \infty$; thus f is not L_∞-integrable. Nonetheless, one checks that if p is an integer, $\|f\|_p = (p!)^{1/p}$, so f is L_p-integrable for integers $p \in [1, \infty)$. More generally one has $\|f\|_p = \Gamma(1+p)^{1/p}$ where the Γ-function generalises the factorial to non-integer values. •

There is another measure of signal size we shall employ that differs somewhat from the above measures in that it is not a norm. We let $f\colon (-\infty, \infty) \to \mathbb{R}$ be a function and say that f is a ***power signal*** if the limit

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f^2(t)\,\mathrm{d}t$$

exists. For a power signal f we then define

$$\mathrm{pow}(f) = \Big(\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f^2(t)\,\mathrm{d}t\Big)^{1/2},$$

which we call the ***average power*** of f. If we consider the function $f(t) = \frac{1}{(1+t)^2}$ we observe that $\mathrm{pow}(f) = 0$ even though f is nonzero. Thus pow is certainly not a norm. Nevertheless, it is a useful, and often used, measure of a signal's size.

The following result gives some relationships between the various L_p-norms and the pow operation.

5.16 Proposition *The following statements hold:*

(i) if $f \in \mathrm{L}_2(-\infty, \infty)$ *then* $\mathrm{pow}(f) = 0$*;*

(ii) if $f \in \mathrm{L}_\infty(-\infty, \infty)$ *is a power signal then* $\mathrm{pow}(f) \le \|f\|_\infty$*;*

(iii) if $f \in \mathrm{L}_1(-\infty, \infty) \cap \mathrm{L}_\infty(-\infty, \infty)$ *then* $\|f\|_2 \le \sqrt{\|f\|_\infty \|f\|_1}$.

Proof (i) For $T > 0$ we have

$$\begin{aligned} &\int_{-T}^{T} f^2(t)\,\mathrm{d}t \le \|f\|_2^2 \\ \implies \quad &\frac{1}{2T} \int_{-T}^{T} f^2(t)\,\mathrm{d}t \le \frac{1}{T}\|f\|_2^2. \end{aligned}$$

The result follows since as $T \to \infty$, the right-hand side goes to zero.

(ii) We compute

$$\begin{aligned} \mathrm{pow}(f) &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f^2(t)\,\mathrm{d}t \\ &\le \|f\|_\infty^2 \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \mathrm{d}t \\ &= \|f\|_\infty^2. \end{aligned}$$

(iii) We have

$$\begin{aligned} \|f\|_2^2 &= \int_{-\infty}^{\infty} f^2(t)\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} |f(t)||f(t)|\,\mathrm{d}t \\ &\le \|f\|_\infty \int_{-\infty}^{\infty} |f(t)|\,\mathrm{d}t \\ &= \|f\|_\infty \|f\|_1, \end{aligned}$$

as desired. ■

The relationships between the various L_p-spaces we shall care about and the pow operation are shown in Venn diagrammatic form in Figure 5.3.

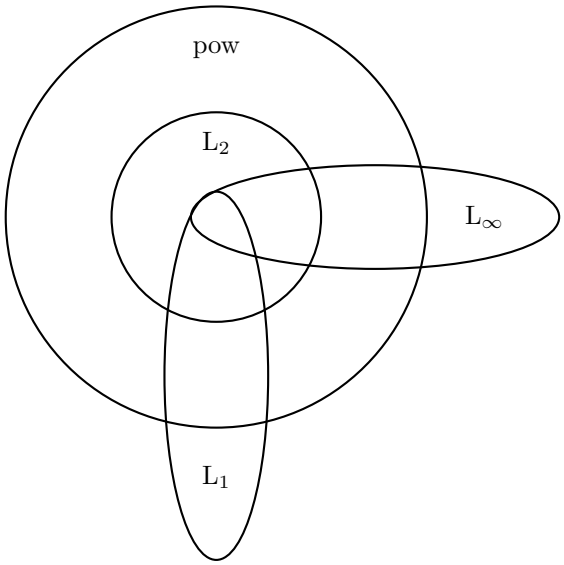


Figure 5.3 Venn diagram for relationships between L_p-spaces and pow

5.3.2 Hardy spaces and transfer function norms

For a meromorphic complex-valued function f we will be interested in measuring the "size" by f by evaluating its restriction to the imaginary axis. To this end, given a meromorphic function f, we follow the analogue of our time-domain norms and define, for $p \geq 1$, the $\mathbf{H_p}$***-norm*** of f by

$$\|f\|_p = \Big(\frac{1}{2\pi}\int_{-\infty}^{\infty} |f(i\omega)|^p \,\mathrm{d}\omega\Big)^{1/p}.$$

In like manner we define the $\mathbf{H_\infty}$***-norm*** of f by

$$\|f\|_\infty = \sup_\omega |f(i\omega)|.$$

While these definitions make sense for any meromorphic function f, we are interested in particular such functions. In particular, we denote

$$\mathrm{RL}_p = \{R \in \mathbb{R}(s) \mid \|R\|_p < \infty\}$$

for $p \in [1, \infty) \cup \{\infty\}$. Let us call the class of meromorphic functions f that are analytic in $\mathbb{C}_+$ ***Hardy functions***.[1] We then have

$$\mathrm{H}_p^+ = \{f \mid f \text{ is a Hardy function with } \|f\|_p < \infty\},$$

for $p \in [0, \infty) \cup \{\infty\}$. We also have $\mathrm{RH}_p^+ = \mathbb{R}(s) \cap \mathrm{H}_p$ as the Hardy functions with bounded H_p-norm that are real rational. In actuality, we shall only be interested in the case when

[1] After George Harold Hardy (1877-1947).

$p \in \{1, 2, \infty\}$, but the definition of the H_p-norm holds generally. One must be careful when one sees the symbol $\|\cdot\|_p$ that one understands what the argument is. In one case we mean it to measure the norm of a function of t defined on $(-\infty, \infty)$, and in another case we use it to define the norm of a complex function measured by looking at its values on the imaginary axis.

Note that with the above notation, we have the following characterisation of BIBO stability.

5.17 Proposition *A proper SISO linear system (N, D) in input/output form is BIBO stable if and only if $T_{N,D} \in \mathrm{RH}^+_\infty$.*

The following result gives straightforward characterisations of the various rational function spaces we have been talking about.

5.18 Proposition *The following statements hold:*

(i) RL_∞ *consists of those functions in* $\mathbb{R}(s)$ *that*
  (a) have no poles on $i\mathbb{R}$ *and*
  (b) are proper;

(ii) RH^+_∞ *consists of those functions in* $\mathbb{R}(s)$ *that*
  (a) have no poles in $\overline{\mathbb{C}}_+$ *and*
  (b) are proper;

(iii) RL_2 *consists of those functions in* $\mathbb{R}(s)$ *that*
  (a) have no poles on $i\mathbb{R}$ *and*
  (b) are strictly proper.

(iv) RH^+_2 *consists of those functions in* $\mathbb{R}(s)$ *that*
  (a) have no poles in $\overline{\mathbb{C}}_+$ *and*
  (b) are strictly proper.

Proof Clearly we may prove the first and second, and then the third and fourth assertions together.

(i) and (ii): This part of the proposition follows since a rational Hardy function is proper if and only if $\lim_{s\to\infty}|R(s)| < \infty$, and since $|R(i\omega)|$ is bounded for all $\omega \in \mathbb{R}$ if and only if R has no poles on $i\mathbb{R}$. The same applies for RL_∞.

(iii) and (iv) Clearly if $R \in \mathrm{RH}^+_2$ then $\lim_{s\to\infty}|R(s)| = 0$, meaning that R must be strictly proper. We also need to show that $R \in \mathrm{RH}^+_2$ implies that R has no poles on $i\mathbb{R}$. We shall do this by showing that if R has poles on $i\mathbb{R}$ then $R \notin \mathrm{RH}^+_2$. Indeed, if R has a pole at $\pm i\omega_0$ then near $i\omega_0$, R will essentially look like

$$R(s) \approx \frac{C}{(s - i\omega_0)^k}$$

for some positive integer k and some $C \in \mathbb{C}$. Let us define $\tilde{R}$ to be the function on the right hand side of this approximation, and note that

$$\begin{aligned}\int_{\omega_0-\epsilon}^{\omega_0+\epsilon} \left|\tilde{R}(i\omega)\right|^2 \mathrm{d}\omega &= \int_{\omega_0-\epsilon}^{\omega_0+\epsilon} \left|\frac{C}{(i(\omega - \omega_0))^k}\right|^2 \mathrm{d}\omega \\ &= |C|^2 \int_{-\epsilon}^{\epsilon} \left|\frac{1}{\xi^k}\right|^2 \mathrm{d}\xi \\ &= \infty.\end{aligned}$$

Thus the contribution to $\|R\|_2$ of a pole on the imaginary axis will always be unbounded.

Conversely, if R is strictly proper with no poles on the imaginary axis, then one can find a sufficiently large $M > 0$ and a sufficiently small $\tau > 0$ so that

$$\Big|\frac{M}{1 + i\tau\omega}\Big| \geq |R(i\omega)|, \quad \omega \in \mathbb{R}.$$

One then computes

$$\int_{-\infty}^{\infty} \Big|\frac{M}{1 + i\tau\omega}\Big|^2 \, \mathrm{d}\omega = \frac{M}{\sqrt{2\tau}}.$$

This implies that $\|R\|_2 \leq \frac{M}{\sqrt{2\tau}}$ and so $R \in \mathrm{RH}_2^+$.

Clearly the above argument for RH_2^+ also applies for RL_2. ■

5.3.3 Stability interpretations of norms

To characterise BIBO stability in terms of these signal norms, we consider a SISO linear system (N, D) in input/output form. We wish to flush out the input/output properties of a transfer function relative to the L_p signal norms and the pow operation. For notational convenience, let us adopt the notation $\|\cdot\|_{\mathrm{pow}} = \mathrm{pow}$ and let $\mathrm{L}_{\mathrm{pow}}(-\infty, \infty)$ denote those functions f for which $\mathrm{pow}(f)$ is defined. This *is* an abuse of notation since pow is not a norm. However, the abuse is useful for making the following definition.

5.19 Definition Let $R \in \mathbb{R}(s)$ be a proper rational function, and for $u \in \mathrm{L}_2(-\infty, \infty)$ let $y_u \colon (-\infty, \infty) \to \mathbb{R}$ be the function satisfying $\hat{y}_u(s) = R(s)\hat{u}(s)$. For $p_1, p_2 \in [1, \infty) \cup \{\infty\} \cup \{\mathrm{pow}\}$, the $\mathbf{L}_{p_1} \to \mathbf{L}_{p_2}$***-gain*** of R is defined by

$$\|R\|_{p_1 \to p_2} = \sup_{\substack{u \in \mathrm{L}_{p_1}(-\infty,\infty) \\ u \text{ not zero}}} \frac{\|y_u\|_{p_2}}{\|u\|_{p_1}}.$$

If (N, D) is SISO linear system in input/output form, then (N, D) is $\mathbf{L}_{p_1} \to \mathbf{L}_{p_2}$***-stable*** if $\|T_{N,D}\|_{p_1 \to p_2} < \infty$. •

This definition of $\mathrm{L}_{p_1} \to \mathrm{L}_{p_2}$-stability is motivated by the following obvious result.

5.20 Proposition *Let (N, D) be an $\mathrm{L}_{p_1} \to \mathrm{L}_{p_2}$ stable SISO linear system in input/output form and let $u \colon (-\infty, \infty) \to \mathbb{R}$ be an input with $y_u \colon (-\infty, \infty) \to \mathbb{R}$ the function satisfying $\hat{y}_u(s) = T_{N,D}(s)\hat{u}(s)$. If $u \in \mathrm{L}_{p_1}(-\infty, \infty)$ then*

$$\|y_u\|_{p_2} \leq \|T_{N,D}\|_{p_1 \to p_2} \|u\|_{p_1}.$$

In particular, $u \in \mathrm{L}_{p_1}(-\infty, \infty)$ implies that $y_u \in \mathrm{L}_{p_2}(-\infty, \infty)$.

Although our definitions have been made in the general context of L_p-spaces, we are primarily interested in the cases where $p_1, p_2 \in \{1, 2, \infty\}$. In particular, we would like to be able to relate the various gains for transfer functions to the Hardy space norms of the previous section. The following result gives these characterisations. The proofs, as you will see, is somewhat long and involved.

5.21 Theorem *For a proper BIBO stable SISO linear system (N, D) in input/output form, let $C \in \mathbb{R}$ be defined by $T_{N,D}(s) = T_{\tilde{N},\tilde{D}}(s) + C$ where $(\tilde{N}, \tilde{D})$ is strictly proper. Thus $C = 0$ if (N, D) is itself strictly proper. The following statements then hold:*

(i) $\|T_{N,D}\|_{2\to 2} = \|T_{N,D}\|_\infty$;
(ii) $\|T_{N,D}\|_{2\to\infty} = \|T_{N,D}\|_2$;
(iii) $\|T_{N,D}\|_{2\to\text{pow}} = 0$;
(iv) $\|T_{N,D}\|_{\infty\to 2} = \infty$;
(v) $\|T_{N,D}\|_{\infty\to\infty} \leq \|h_{\tilde{N},\tilde{D}}\|_1 + |C|$;
(vi) $\|T_{N,D}\|_{\infty\to\text{pow}} \leq \|T_{N,D}\|_\infty$;
(vii) $\|T_{N,D}\|_{\text{pow}\to 2} = \infty$;
(viii) $\|T_{N,D}\|_{\text{pow}\to\infty} = \infty$;
(ix) $\|T_{N,D}\|_{\text{pow}\to\text{pow}} = \|T_{N,D}\|_\infty$.

If (N, D) is strictly proper, then part (v) can be improved to $\|T_{N,D}\|_{\infty\to\infty} = \|h_{N,D}\|_1$.

Proof (i) By Parseval's Theorem we have $\|f\|_2 = \|\hat{f}\|_2$ for any function $f \in \mathrm{L}_2(-\infty, \infty)$. Therefore

$$\begin{aligned}
\|y_u\|_2^2 &= \|\hat{y}_u\|_2^2 \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} |\hat{y}_u(i\omega)|^2 \,\mathrm{d}\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} |T_{N,D}(i\omega)|^2 |\hat{u}(i\omega)|^2 \,\mathrm{d}\omega \\
&\leq \|T_{N,D}\|_\infty^2 \frac{1}{2\pi}\int_{-\infty}^{\infty} |\hat{u}(i\omega)|^2 \,\mathrm{d}\omega \\
&= \|T_{N,D}\|_\infty^2 \|\hat{u}\|_2^2 \\
&= \|T_{N,D}\|_\infty^2 \|u\|_2^2.
\end{aligned}$$

This shows that $\|T_{N,D}\|_{2\to 2} \leq \|T_{N,D}\|_\infty$. We shall show that this is the least upper bound. Let $\omega_0 \in \mathbb{R}_+$ be a frequency at which $\|T_{N,D}\|_\infty$ is attained. First let us suppose that ω_0 is finite. For $\epsilon > 0$ define u_ϵ to have the property

$$\hat{u}_\epsilon(i\omega) = \begin{cases} \sqrt{\pi/2\epsilon}, & |\omega - \omega_0| < \epsilon \text{ or } |\omega + \omega_0| < \epsilon \\ 0, & \text{otherwise.} \end{cases}$$

Then, by Parseval's Theorem, $\|u\|_2 = 1$. We also compute

$$\begin{aligned}
\lim_{\epsilon\to 0} \|\hat{y}_{u_\epsilon}\| &= \frac{1}{2\pi}\Big(\pi|T_{N,D}(-i\omega_0)|^2 + \pi|T_{N,D}(i\omega_0)|^2\Big) \\
&= |T_{N,D}(i\omega_0)|^2 \\
&= \|T_{N,D}\|_\infty^2.
\end{aligned}$$

If $\|T_{N,D}\|_\infty$ is not attained at a finite frequency, then we define u_ϵ so that

$$\hat{u}_\epsilon(i\omega) = \begin{cases} \sqrt{\pi/2\epsilon}, & |\omega - \frac{1}{\epsilon}| < \epsilon \text{ or } |\omega + \frac{1}{\epsilon}| < \epsilon \\ 0, & \text{otherwise.} \end{cases}$$

In this case we still have $\|u\|_2 = 1$, but now we have

$$\lim_{\epsilon\to 0} \|\hat{y}_{u_\epsilon}\| = \lim_{\omega\to\infty} |T_{N,D}(i\omega)|^2 = \|T_{N,D}\|_2.$$

In either case we have shown that $\|T_{N,D}\|_\infty$ is a least upper bound for $\|T_{N,D}\|_{2\to 2}$.

(ii) Here we employ the Cauchy-Schwartz inequality to determine

$$\begin{aligned}|y_u(t)| &= \int_{-\infty}^{\infty} h_{N,D}(t-\tau)u(\tau)\,\mathrm{d}\tau \\ &\leq \Big(\int_{-\infty}^{\infty} h_{N,D}^2(t-\tau)\,\mathrm{d}\tau\Big)^{1/2}\Big(\int_{-\infty}^{\infty} u^2(\tau)\,\mathrm{d}\tau\Big)^{1/2} \\ &= \|h_{N,D}\|_2\|u\|_2 \\ &= \|T_{N,D}\|_2\|u\|_2,\end{aligned}$$

where the last equality follows from Parseval's Theorem. Thus we have shown that $\|T_{N,D}\|_{2\to\infty} \leq \|T_{N,D}\|_2$. This is also the least upper bound since if we take

$$u(t) = \frac{h_{N,D}(-t)}{\|T_{N,D}\|_2},$$

we determine by Parseval's Theorem that $\|u\|_2 = 1$ and from our above computations that $|y(0)| = \|T_{N,D}\|_2$ which means that $\|y_u\|_\infty \geq \|T_{N,D}\|_2$, as desired.

(iii) Since y_u is L_2-integrable if u is L_2-integrable by part (i), this follows from Proposition 5.16(i).

(iv) Let $\omega \in \mathbb{R}_+$ have the property that $T_{N,D}(i\omega) \neq 0$. Take $u(t) = \sin\omega t$. By Theorem 4.1 we have

$$y_u(t) = \mathrm{Re}(T_\Sigma(i\omega))\sin\omega t + \mathrm{Im}(T_\Sigma(i\omega))\cos\omega t + \tilde{y}_h(t) + C$$

where $\lim_{t\to\infty} y_h(t) = 0$. In this case we have $\|u\|_\infty = 1$ and $\|y_u\|_2 = \infty$.

(v) We compute

$$\begin{aligned}|y(t)| &= \Big|\int_{-\infty}^{\infty} h_{\tilde{N},\tilde{D}}(t-\tau)u(\tau)\,\mathrm{d}\tau + Cu(t)\Big| \\ &= \Big|\int_{-\infty}^{\infty} h_{\tilde{N},\tilde{D}}(\tau)u(t-\tau)\,\mathrm{d}\tau + Cu(t)\Big| \\ &\leq \int_{-\infty}^{\infty} |h_{\tilde{N},\tilde{D}}(\tau)u(t-\tau)|\,\mathrm{d}\tau + |C||u(t)| \\ &\leq \|u\|_\infty\Big(\int_{-\infty}^{\infty} |h_{\tilde{N},\tilde{D}}(\tau)|\,\mathrm{d}\tau + |C|\Big) \\ &= \big(\|h_{\tilde{N},\tilde{D}}\|_1 + |C|\big)\|u\|_\infty.\end{aligned}$$

This shows that $\|T_{N,D}\|_{\infty\to\infty} \leq \|h_{\tilde{N},\tilde{D}}\|_1 + |C|$ as stated. To see that this is the least upper bound when (N, D) is strictly proper (cf. the final statement in the theorem), fix $t > 0$ and define u so that

$$u(t-\tau) = \begin{cases} +1, & h_{N,D}(\tau) \geq 0 \\ -1, & h_{N,D}(\tau) < 0. \end{cases}$$

Then we have $\|u\|_\infty = 1$ and

$$\begin{aligned}y_u(t) &= \int_{-\infty}^{\infty} h_{N,D}(\tau)u(t-\tau)\,\mathrm{d}\tau \\ &= \int_{-\infty}^{\infty} |h_{N,D}(\tau)|\,\mathrm{d}\tau \\ &= \|h_{N,D}\|_1.\end{aligned}$$

Thus $\|y_u\|_\infty \geq \|h_{N,D}\|_1$.

(vii) To carry out this part of the proof, we need a little diversion. For a power signal f define

$$\rho(f)(t) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(\tau) f(t+\tau)\, \mathrm{d}\tau$$

and note that $\rho(f)(0) = \mathrm{pow}(f)$. The limit in the definition of $\rho(f)$ may not exist for all τ, but it will exist for certain power signals. Let f be a nonzero power signal for which the limit does exist. Denote by $\sigma(f)$ the Fourier transform of $\rho(f)$:

$$\sigma(f)(\omega) = \int_{-\infty}^{\infty} \rho(f)(t) e^{-i\omega t}\, \mathrm{d}t.$$

Therefore, since $\rho(f)$ is the inverse Fourier transform of $\sigma(f)$ we have

$$\mathrm{pow}(f)^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \sigma(f)(\omega)\, \mathrm{d}\omega. \tag{5.5}$$

Now we claim that if y_u is related to u by $\hat{y}_u(s) = T_{N,D}(s)\hat{u}(s)$ where u is a power signal for which $\rho(u)$ exists, then we have

$$\sigma(y_u)(\omega) = |T_{N,D}(i\omega)|^2 \sigma(u)(\omega). \tag{5.6}$$

Indeed note that

$$y_u(t) y_u(t+\tau) = \int_{-\infty}^{\infty} h_{N,D}(\alpha) y(t) u(t+\tau-\alpha)\, \mathrm{d}\alpha,$$

so that

$$\rho(y_u)(t) = \int_{-\infty}^{\infty} h_{N,D}(\tau) \rho(y_u, u)(t-\tau)\, \mathrm{d}\tau,$$

where

$$\rho(f,g)(t) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(\tau) g(t+\tau)\, \mathrm{d}\tau.$$

In like manner we verify that

$$\rho(y_u, u)(t) = \int_{-\infty}^{\infty} h^-_{N,D}(t-\tau) \rho(u)(\tau)\, \mathrm{d}\tau,$$

where $h^-_{N,D}(t) = h_{N,D}(-t)$. Therefore we have $\rho(y_u) = h_{N,D} * h^-_{N,D} * \rho(u)$, where $*$ signifies convolution. One readily verifies that the Fourier transform of $h^-_{N,D}$ is the complex conjugate of the Fourier transform of $h_{N,D}$. Therefore

$$\sigma(y_u)(\omega) = T_{N,D}(i\omega) \bar{T}_{N,D}(i\omega) \sigma(u)(\omega),$$

which gives (5.6) as desired. Using (5.5) combined with (5.6) we then have

$$\mathrm{pow}(y_u)^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} |T_{N,D}(i\omega)|^2 \sigma(u)(\omega). \tag{5.7}$$

Provided that we choose u so that $|T_{N,D}(i\omega)|^2 \sigma(u)(\omega)$ is not identically zero, we see that $\mathrm{pow}(y_u)^2 > 0$ so that $\|y_u\| = \infty$.

(ix) By (5.7) we have $\mathrm{pow}(y_u) \leq \|T_{N,D}\|_\infty \mathrm{pow}(u)$. Therefore $\|T_{N,D}\|_{\mathrm{pow}\to\mathrm{pow}} \leq \|T_{N,D}\|_\infty$. To show that this is the least upper bound, let $\omega_0 \in \mathbb{R}_+$ be a frequency at which $\|T_{N,D}\|_\infty$ is realised, and first suppose that ω_0 is finite. Now let $u(t) = \sqrt{2}\sin\omega_0 t$. One readily computes $\rho(u)(t) = \cos\omega_0 t$, implying by (5.5) that $\mathrm{pow}(u) = 1$. Also we clearly have

$$\sigma(u)(\omega) = \pi\big(\delta(\omega-\omega_0) + \delta(\omega+\omega_0)\big),$$

An application of (5.7) then gives

$$\begin{aligned}\mathrm{pow}(y_u)^2 &= \frac{1}{2}\big(|T_{N,D}(i\omega_0)|^2 + |T_{N,D}(-i\omega_0)|^2\big)\\ &= |T_{N,D}(i\omega_0)|^2\\ &= \|T_{N,D}\|_\infty^2.\end{aligned}$$

If $\|T_{N,D}\|_\infty$ is attained only in the limit as frequency goes to infinity, then the above argument is readily modified to show that one can find a signal u so that $\mathrm{pow}(y_u)$ is arbitrarily close to $\|T_{N,D}\|_\infty$.

(vi) Let $u \in \mathrm{L}_\infty(-\infty,\infty)$ be a power signal. By Proposition 5.16(ii) we have $\mathrm{pow}(u) \leq \|u\|_\infty$. It therefore follows that

$$\begin{aligned}\|T_{N,D}\|_{\infty\to\mathrm{pow}} &= \sup_{\substack{u\in\mathrm{L}_\infty(-\infty,\infty)\\ u \text{ not zero}}} \frac{\|y_u\|_{\mathrm{pow}}}{\|u\|_\infty}\\ &\leq \sup_{\substack{u\in\mathrm{L}_\infty(-\infty,\infty)\\ u\in\mathrm{L}_{\mathrm{pow}}(-\infty,\infty)\\ u \text{ not zero}}} \frac{\|y_u\|_{\mathrm{pow}}}{\|u\|_\infty}\\ &\leq \sup_{\substack{u\in\mathrm{L}_\infty(-\infty,\infty)\\ u\in\mathrm{L}_{\mathrm{pow}}(-\infty,\infty)\\ u \text{ not zero}}} \frac{\|y_u\|_{\mathrm{pow}}}{\|u\|_{\mathrm{pow}}}.\end{aligned}$$

During the course of the proof of part (ix) we showed that there exists a power signal u with $\mathrm{pow}(u) = 1$ with the property that $\mathrm{pow}(y_u) = \|T_{N,D}\|_\infty$. Therefore, this part of the theorem follows.

(viii) For $k \geq 1$ define

$$u_k(t) = \begin{cases} k, & t \in (k, k+\frac{1}{k^3})\\ 0, & \text{otherwise,}\end{cases}$$

and define an input u by

$$u(t) = \sum_{k=1}^{\infty} u_k(t).$$

For $T \geq 1$ let $k(T)$ be the largest integer less than T. One computes

$$\int_{-T}^{T} u^2(t)\,\mathrm{d}t = \begin{cases} \sum_{k=1}^{k(T)-1}\frac{1}{k}, & T \leq k(T)-1\\ \sum_{k=1}^{k(T)}\frac{1}{k}, & T \geq k(T)\\ \big(\sum_{k=1}^{k(T)-1}\frac{1}{k}\big) + t\frac{1}{k(T)}, & t \in [k(T)-1, k(T)].\end{cases}$$

Thus we have

$$\lim_{T\to 0}\frac{1}{2T}\int_{-T}^{T}u^2(t)\,\mathrm{d}t = \lim_{k\to\infty}\frac{1}{2N}\int_{-N}^{N}u^2(t)\,\mathrm{d}t$$
$$= \lim_{N\to\infty}\sum_{k=1}^{N}\frac{1}{k}.$$

Since

$$\sum_{k=1}^{N}\frac{1}{k} < \int_1^N \frac{1}{t}\,\mathrm{d}t = \ln N$$

we have

$$\lim_{T\to 0}\frac{1}{2T}\int_{-T}^{T}u^2(t)\,\mathrm{d}t = \lim_{N\to\infty}\frac{\ln N}{N} = 0.$$

Thus $u \in \mathrm{L}_{\mathrm{pow}}$. ■ Finish

Note that our notion of BIBO stability exactly coincides with $\mathrm{L}_\infty \to \mathrm{L}_\infty$ stability. The following result summarises this along with our other notions of stability.

5.22 Theorem *Let (N, D) be a proper SISO linear system in input/output form. The following statements are equivalent:*

(i) *(N, D) is BIBO stable;*

(ii) *$T_{N,D} \in \mathrm{RH}_\infty^+$;*

(iii) *(N, D) is $\mathrm{L}_2 \to \mathrm{L}_2$-stable;*

(iv) *(N, D) is $\mathrm{L}_\infty \to \mathrm{L}_\infty$-stable;*

(v) *(N, D) is $\mathrm{L}_{\mathrm{pow}} \to \mathrm{L}_{\mathrm{pow}}$-stable.*

Furthermore, if any of the preceding three conditions hold then

$$\|T_{N,D}\|_{2,2} = \|T_{N,D}\|_{\mathrm{pow}\to\mathrm{pow}} = \|T_{N,D}\|_\infty.$$

Proof We shall only prove those parts of the theorem not covered by Theorem 5.21, or other previous results.

(iii) $\Longrightarrow$ (ii) We suppose that $T_{N,D} \notin \mathrm{RH}_\infty^+$ so that D has a root in $\overline{\mathbb{C}}_+$. Let $a < 0$ have the property that all roots of D lie to the right of $\{s \in \mathbb{C} \mid \mathrm{Re}(s) = a\}$. Thus if $u(t) = e^{at}1(t)$ then $u \in \mathrm{L}_2(-\infty, \infty)$. Let $p \in \overline{\mathbb{C}}_+$ be a root for D of multiplicity k. Then

$$\hat{y}_u(s) = R_1(s) + \sum_{j=1}^{k}\frac{R_{2,j}(s)}{(s-p)^k},$$

where $R_1, R_{2,1}, \ldots, R_{2,k}$ are rational functions analytic at p. Taking inverse Laplace transforms gives

$$y_u(t) = y_1(t) + \sum_{j=1}^{k} t^j e^{\mathrm{Re}(p)t}\big(a_j\cos(\mathrm{Im}(p)t) + b_j\sin(\mathrm{Im}(p)t)\big)$$

with $a_j^2 + b_j^2 \neq 0$, $j = 1, \ldots, k$. In particular, since $\mathrm{Re}(p) \geq 0$, $y_u \notin \mathrm{L}_2(-\infty, \infty)$.

(v) $\Longrightarrow$ (ii) ■ Finish

Note that the theorem can be interpreted as saying that a system is BIBO stable if and only if the energy/power of the output corresponding to a finite energy/power input is also finite (here one thinks of the L_2-norm of a signal as a measure of its energy and of pow as its power). At the moment, it seems like this additional characterisation of BIBO stability in terms of $L_2 \to L_2$ and $L_{pow} \to L_{pow}$-stability is perhaps pointless. But the fact of the matter is that this is far from the truth. As we shall see in Section 8.5, the use of the L_2-norm to characterise stability has valuable implications for quantitative performance measures, and their achievement through "H_∞ methods." This is an approach given a thorough treatment by Doyle, Francis, and Tannenbaum [1990] and Morris [2000].

5.4 Liapunov methods

Liapunov methods for stability are particularly useful in the context of stability for nonlinear differential equations and control systems. However, even for linear systems where there are more "concrete" stability characterisations, Liapunov stability theory is useful as it gives a collection of results that are useful, for example, in optimal control for such systems. An application along these lines is the subject of Section 14.3.2. These techniques were pioneered by Aleksandr Mikhailovich Liapunov (1857–1918); see [Liapunov 1893].

5.4.1 Background and terminology

The Liapunov method for determining stability has a general character that we will present briefly in order that we may understand the linear results in a larger context. Let us consider a vector differential equation

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)) \tag{5.8}$$

and an equilibrium point $\boldsymbol{x}_0$ for $\boldsymbol{f}$; thus $\boldsymbol{f}(\boldsymbol{x}_0) = \boldsymbol{0}$. We wish to determine conditions for stability and asymptotic stability of the equilibrium point. First we should provide definitions for these notions of stability as they are a little different from their linear counterparts.

5.23 Definition The differential equation (5.8) is:

(i) ***stable*** if there exists $M > 0$ so that for every $0 < \delta \leq M$ there exists $\epsilon < 0$ so that if $\|\boldsymbol{x}(0) - \boldsymbol{x}_0\| < \epsilon$ then $\|\boldsymbol{x}(t) - \boldsymbol{x}_0\| < \delta$ for every $t > 0$;

(ii) ***asymptotically stable*** if there exists $M > 0$ so that the inequality $\|\boldsymbol{x}(0) - \boldsymbol{x}_0\| < M$ implies that $\lim_{t\to\infty} \|\boldsymbol{x}(t) - \boldsymbol{x}_0\| = 0$. •

Our definition differs from the definitions of stability for linear systems in that it is only local. We do not require that *all* solutions be bounded in order that $\boldsymbol{x}_0$ be stable, only those whose initial conditions are sufficiently close to $\boldsymbol{x}_0$ (and similarly for asymptotic stability).

The Liapunov idea for determining stability is to find a function V that has a local minimum at $\boldsymbol{x}_0$ and whose time derivative along solutions of the differential equation (5.8) is nonpositive. To be precise about this, let us make a definition.

5.24 Definition A function $V : \mathbb{R}^n \to \mathbb{R}$ is a ***Liapunov function*** for the equilibrium point $\boldsymbol{x}_0$ of (5.8) if

(i) $V(\boldsymbol{x}_0) = 0$,

(ii) $V(\boldsymbol{x}) \geq 0$, and

(iii) there exists $M > 0$ so that if $\|\boldsymbol{x}(0) - \boldsymbol{x}_0\| < M$ then $\frac{\mathrm{d}}{\mathrm{d}t} V(\boldsymbol{x}(t)) \leq 0$.

If the nonstrict inequalities in parts (ii) and (iii) are strict, then V is a ***proper Liapunov function***. •

We will not prove the following theorem as we shall prove it in the cases we care about in the next section. Readers interested in the proof may refer to, for example, [Khalil 2001].

5.25 Theorem *Consider the differential equation (5.8) with $\boldsymbol{x}_0$ an equilibrium point. The following statements hold:*

(i) $\boldsymbol{x}_0$ is stable if there is a Liapunov function for $\boldsymbol{x}_0$;

(ii) $\boldsymbol{x}_0$ is asymptotically stable if there is a proper Liapunov function for $\boldsymbol{x}_0$.

Although we do not prove this theorem, it should nonetheless seem reasonable, particularly the second part. Indeed, since in this case we have $\frac{\mathrm{d}}{\mathrm{d}t}V(\boldsymbol{x}(t)) < 0$ and since $\boldsymbol{x}_0$ is a strict local minimum for V, it stands to reason that all solutions should be tending towards this strict local minimum as $t \to \infty$.

Of course, we are interested in linear differential equations of the form

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t).$$

Our interest is in Liapunov functions of a special sort. We shall consider Liapunov functions that are quadratic in $\boldsymbol{x}$. To define such a function, let $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ be symmetric and let $V(\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{P}\boldsymbol{x}$. We then compute

$$\begin{aligned}\frac{\mathrm{d}V(\boldsymbol{x}(t))}{\mathrm{d}t} &= \dot{\boldsymbol{x}}^t(t)\boldsymbol{P}\boldsymbol{x}(t) + \boldsymbol{x}^t(t)\boldsymbol{P}\dot{\boldsymbol{x}}(t)\\ &= \boldsymbol{x}^t(t)(\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A})\boldsymbol{x}(t).\end{aligned}$$

Note that the matrix $\boldsymbol{Q} = -\boldsymbol{A}^t\boldsymbol{P} - \boldsymbol{P}\boldsymbol{A}$ is itself symmetric. Now, to apply Theorem 5.25 we need to be able to characterise when the functions $\boldsymbol{x}^t\boldsymbol{P}\boldsymbol{x}$ and $\boldsymbol{x}^t\boldsymbol{Q}\boldsymbol{x}$ is nonnegative. This we do with the following definition.

5.26 Definition Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be symmetric.

(i) $\boldsymbol{M}$ is ***positive-definite*** (written $\boldsymbol{M} > 0$) if $\boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$.

(ii) $\boldsymbol{M}$ is ***negative-definite*** (written $\boldsymbol{M} < 0$) if $\boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x} < 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$.

(iii) $\boldsymbol{M}$ is ***positive-semidefinite*** (written $\boldsymbol{M} \geq 0$) if $\boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

(iv) $\boldsymbol{M}$ is ***negative-semidefinite*** (written $\boldsymbol{M} \leq 0$) if $\boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x} \leq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$. •

The matter of determining when a matrix is positive-(semi)definite or negative-(semi)definite is quite a simple matter in principle when one remembers that a symmetric matrix is guaranteed to have real eigenvalues. With this in mind, we have the following result whose proof is a simple exercise.

5.27 Proposition *For $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be symmetric the following statements hold:*

(i) $\boldsymbol{M}$ is positive-definite if and only if $\operatorname{spec}(\boldsymbol{M}) \subset \mathbb{C}_+ \cap \mathbb{R}$;

(ii) $\boldsymbol{M}$ is negative-definite if and only if $\operatorname{spec}(\boldsymbol{M}) \subset \mathbb{C}_- \cap \mathbb{R}$;

(iii) $\boldsymbol{M}$ is positive-semidefinite if and only if $\operatorname{spec}(\boldsymbol{M}) \subset \overline{\mathbb{C}}_+ \cap \mathbb{R}$;

(iv) $\boldsymbol{M}$ is negative-semidefinite if and only if $\operatorname{spec}(\boldsymbol{M}) \subset \overline{\mathbb{C}}_- \cap \mathbb{R}$.

Another characterisation of positive-definiteness involves the principal minors of $\boldsymbol{M}$. The following result is not entirely trivial, and a proof may be found in [Gantmacher 1959a].

5.28 Theorem *A symmetric $n \times n$ matrix $\boldsymbol{M}$ is positive-definite if and only if all principal minors of $\boldsymbol{M}$ are positive.*

Along these lines, the following result from linear algebra will be helpful to us in the next section.

5.29 Proposition *If $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is positive-definite then there exists $\delta, \epsilon > 0$ so that for every $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ we have*

$$\epsilon \boldsymbol{x}^t\boldsymbol{x} < \boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x} < \delta \boldsymbol{x}^t\boldsymbol{x}.$$

Proof Let $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ be a matrix for which $\boldsymbol{D} = \boldsymbol{T}\boldsymbol{M}\boldsymbol{T}^{-1}$ is diagonal. Recall that $\boldsymbol{T}$ can be chosen so that it is orthogonal, i.e., so that its rows and columns are orthonormal bases for $\mathbb{R}^n$. It follows that $\boldsymbol{T}^{-1} = \boldsymbol{T}^t$. Let us also suppose that the diagonal elements $d_1, \dots, d_n$ of $\boldsymbol{D}$ are ordered so that $d_1 \leq d_2 \leq \cdots \leq d_n$. Let us define $\epsilon = \frac{1}{2}d_1$ and $\delta = 2d_n$. Since for $\boldsymbol{x} = (x_1, \dots, x_n)$ we have

$$\boldsymbol{x}^t\boldsymbol{D}\boldsymbol{x} = \sum_{i=1}^{n} d_i x_i^2,$$

it follows that

$$\epsilon \boldsymbol{x}^t\boldsymbol{x} < \boldsymbol{x}^t\boldsymbol{D}\boldsymbol{x} < \delta \boldsymbol{x}^t\boldsymbol{x}$$

for every $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$. Therefore, since

$$\boldsymbol{x}^t\boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^t\boldsymbol{T}^t\boldsymbol{D}\boldsymbol{T}\boldsymbol{x} = (\boldsymbol{T}\boldsymbol{x})^t\boldsymbol{D}(\boldsymbol{T}\boldsymbol{x}),$$

the result follows. ■

With this background and notation, we are ready to proceed with the results concerning Liapunov functions for linear differential equations.

5.4.2 Liapunov functions for linear systems

The reader will wish to recall from Remark 2.18 our discussion of observability for MIMO systems, as we will put this to use in this section. A ***Liapunov triple*** is a triple $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ of $n \times n$ real matrices with $\boldsymbol{P}$ and $\boldsymbol{Q}$ symmetric and satisfying

$$\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{Q}.$$

We may now state our first result.

5.30 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system and let $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ be a Liapunov triple. The following statements hold:*

(i) if $\boldsymbol{P}$ is positive-definite and $\boldsymbol{Q}$ is positive-semidefinite, then Σ is internally stable;

(ii) if $\boldsymbol{P}$ is positive-definite, $\boldsymbol{Q}$ is positive-semidefinite, and $(\boldsymbol{A}, \boldsymbol{Q})$ is observable, then Σ is internally asymptotically stable;

(iii) if $\boldsymbol{P}$ is not positive-semidefinite, $\boldsymbol{Q}$ is positive-semi-definite, and $(\boldsymbol{A}, \boldsymbol{Q})$ is observable, then Σ is internally unstable.

Proof (i) As in Proposition 5.29, let $\epsilon, \delta > 0$ have the property that

$$\epsilon \boldsymbol{x}^t\boldsymbol{x} < \boldsymbol{x}^t\boldsymbol{P}\boldsymbol{x} < \delta \boldsymbol{x}^t\boldsymbol{x}$$

for $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$. Let $V(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{P}\boldsymbol{x}$. We compute

$$\frac{\mathrm{d}V(\boldsymbol{x}(t))}{\mathrm{d}t} = \boldsymbol{x}^t(\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A})\boldsymbol{x} = -\boldsymbol{x}^t\boldsymbol{Q}\boldsymbol{x},$$

since $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple. As $\boldsymbol{Q}$ is positive-semidefinite, this implies that

$$V(\boldsymbol{x}(t)) - V(\boldsymbol{x}(0)) = \int_0^t \frac{\mathrm{d}V(\boldsymbol{x}(t))}{\mathrm{d}t}\,\mathrm{d}t \leq 0$$

for all $t \geq 0$. Thus, for $t \geq 0$,

$$\begin{aligned} & \boldsymbol{x}^t(t)\boldsymbol{P}\boldsymbol{x}(t) \leq \boldsymbol{x}^t(0)\boldsymbol{P}\boldsymbol{x}(0) \\ \implies \quad & \epsilon\boldsymbol{x}^t(t)\boldsymbol{x}(t) < \delta\boldsymbol{x}^t(0)\boldsymbol{x}(0) \\ \implies \quad & \boldsymbol{x}^t(t)\boldsymbol{x}(t) < \frac{\delta}{\epsilon}\boldsymbol{x}^t(0)\boldsymbol{x}(0) \\ \implies \quad & \|\boldsymbol{x}(t)\| < \sqrt{\frac{\delta}{\epsilon}}\|\boldsymbol{x}(0)\|. \end{aligned}$$

Thus $\|\boldsymbol{x}(t)\|$ is bounded for all $t \geq 0$, and for linear systems, this implies internal stability.

(ii) We suppose that $\boldsymbol{P}$ is positive-definite, $\boldsymbol{Q}$ is positive-semidefinite, $(\boldsymbol{A}, \boldsymbol{Q})$ is observable, and that Σ is not internally asymptotically stable. By (i) we know Σ is stable, so it must be the case that $\boldsymbol{A}$ has at least one eigenvalue on the imaginary axis, and therefore a nontrivial periodic solution $\boldsymbol{x}(t)$. From our characterisation of the matrix exponential in Section B.2 we know that this periodic solution evolves in a two-dimensional subspace that we shall denote by L. What's more, every solution of $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$ with initial condition in L is periodic and remains in L. This implies that L is $\boldsymbol{A}$-invariant. Indeed, if $\boldsymbol{x} \in L$ then

$$\boldsymbol{A}\boldsymbol{x} = \lim_{t\to 0} \frac{e^{\boldsymbol{A}t}\boldsymbol{x} - \boldsymbol{x}}{t} \in L$$

since $\boldsymbol{x}, e^{\boldsymbol{A}t}\boldsymbol{x} \in L$. We also claim that the subspace L is in $\ker(\boldsymbol{Q})$. To see this, suppose that the solutions on L have period T. If $V(\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{P}\boldsymbol{x}$, then for any solution $\boldsymbol{x}(t)$ in L we have

$$0 = V(\boldsymbol{x}(T)) - V(\boldsymbol{x}(0)) = \int_0^T \frac{\mathrm{d}V(\boldsymbol{x}(t))}{\mathrm{d}t}\,\mathrm{d}t = -\int_0^T \boldsymbol{x}^t(t)\boldsymbol{Q}\boldsymbol{x}(t)\,\mathrm{d}t.$$

Since $\boldsymbol{Q}$ is positive-semidefinite this implies that $\boldsymbol{x}^t(t)\boldsymbol{Q}\boldsymbol{x}(t) = 0$. Thus $L \subset \ker(\boldsymbol{Q})$, as claimed. Thus, with our initial assumptions, we have shown the existence of an nontrivial $\boldsymbol{A}$-invariant subspace of $\ker(\boldsymbol{Q})$. This is a contradiction, however, since $(\boldsymbol{A}, \boldsymbol{Q})$ is observable. It follows, therefore, that Σ is internally asymptotically stable.

(iii) Since $\boldsymbol{Q}$ is positive-semidefinite and $(\boldsymbol{A}, \boldsymbol{Q})$ is observable, the argument from (ii) shows that there are no nontrivial periodic solutions to $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$. Thus this part of the theorem will follow if we can show that Σ is not internally asymptotically stable. By hypothesis, there exists $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ so that $V(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}}^t\boldsymbol{P}\bar{\boldsymbol{x}} < 0$. Let $\boldsymbol{x}(t)$ be the solution of $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{x}(0) = \bar{\boldsymbol{x}}$. As in the proof of (i) we have $V(\boldsymbol{x}(t)) \leq V(\bar{\boldsymbol{x}}) < 0$ for all $t \geq 0$ since $\boldsymbol{Q}$ is positive-semidefinite. If we denote

$$r = \inf\{\|\boldsymbol{x}\| \mid V(\boldsymbol{x}) \leq V(\bar{\boldsymbol{x}})\},$$

then we have shown that $\|x\|(t) > r$ for all $t \geq 0$. This prohibits internal asymptotic stability, and in this case, internal stability. ■

5.31 Example **(Example 5.4 cont'd)** We again look at the 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix},$$

letting $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ for some $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{D}$. For this example, there are various cases to consider, and we look at them separately in view of Theorem 5.30. In the following discussion, the reader should compare the conclusions with those of Example 5.4.

1. $a = 0$ and $b = 0$: In this case, we know the system is internally unstable. However, it turns out to be impossible to find a symmetric $\boldsymbol{P}$ and a positive-semidefinite $\boldsymbol{Q}$ so that $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple, and so that $(\boldsymbol{A}, \boldsymbol{Q})$ is observable (cf. Exercise E5.16). Thus we *cannot* use part (iii) of Theorem 5.30 to assert internal instability. We are off to a bad start! But things start to look better.
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 $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ are a Liapunov triple. Since $\boldsymbol{P}$ is positive-definite and $\boldsymbol{Q}$ is positive-semidefinite, internal stability follows from part (i) of Theorem 5.30. Note that $(\boldsymbol{A}, \boldsymbol{Q})$ is not observable, so internal asymptotic stability cannot be concluded from part (ii).
3. $a = 0$ 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 $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ are a Liapunov triple. Since $\boldsymbol{P}$ is not positive-semidefinite (its eigenvalues are $\{\pm\frac{1}{2}\}$) and since $\boldsymbol{Q}$ is positive-definite and $(\boldsymbol{A}, \boldsymbol{Q})$ is observable ($\boldsymbol{Q}$ is invertible), it follows from part (iii) of Theorem 5.30 that the system is internally unstable.
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 $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple. 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. We also compute

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{Q}) = \begin{bmatrix} 0 & 0 \\ 0 & 2a \\ 0 & 0 \\ 0 & -2a^2 \end{bmatrix},$$

verifying that $(\boldsymbol{A}, \boldsymbol{Q})$ is not observable. Thus from part (i) of Theorem 5.30 we conclude that Σ is internally stable, but we cannot conclude internal asymptotic stability from (ii).
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 $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple. We 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. Since $\boldsymbol{P}$ is positive-definite, we may conclude from part (ii) of Theorem 5.30 that Σ is internally asymptotically stable.

6. $a > 0$ and $b < 0$: Again we use

$$\boldsymbol{P} = \begin{bmatrix} b & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 2a \end{bmatrix}.$$

Now, since $\boldsymbol{P}$ is not positive-semidefinite, from part (iii) of Theorem 5.30, we conclude that Σ is internally unstable.

7. $a < 0$ and $b = 0$: This case is much like case 1 in that the system is internally unstable, but we cannot find a symmetric $\boldsymbol{P}$ and a positive-semidefinite $\boldsymbol{Q}$ so that $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple, and so that $(\boldsymbol{A}, \boldsymbol{Q})$ is observable (again see Exercise E5.16).

8. $a < 0$ 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 $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple. 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, from part (iii) of Theorem 5.30 we conclude that Σ is internally unstable.

9. $a < 0$ and $b < 0$: 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 tells us that Σ is internally unstable. •

Note that in two of the nine cases in the preceding example, it was not possible to apply Theorem 5.30 to conclude internal instability of a system. This points out something of a weakness of the Liapunov approach, as compared to Theorem 5.2 which captures all possible cases of internal stability and instability. Nevertheless, the Liapunov characterisation of stability can be a useful one in practice. It is used by us in Chapters 14 and 15.

While Theorem 5.30 tells us how we certain Liapunov triples imply certain stability properties, often one wishes for a converse to such results. Thus one starts with a system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ that is stable in some way, and one wishes to ascertain the character of the corresponding Liapunov triples. While the utility of such an exercise is not immediately obvious, it will come up in Section 14.3.2 when characterising solutions of an optimal control problem.

5.32 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system with $\boldsymbol{A}$ Hurwitz. The following statements hold:*

(i) for any symmetric $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ there exists a unique symmetric $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ so that $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple;

(ii) if $\boldsymbol{Q}$ is positive-semidefinite with $\boldsymbol{P}$ the unique symmetric matrix for which $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple, then $\boldsymbol{P}$ is positive-semidefinite;

(iii) if $\boldsymbol{Q}$ is positive-semidefinite with $\boldsymbol{P}$ the unique symmetric matrix for which $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple, then $\boldsymbol{P}$ is positive-definite if and only if $(\boldsymbol{A}, \boldsymbol{Q})$ is observable.

Proof (i) We claim that if we define

$$\boldsymbol{P} = \int_0^\infty e^{\boldsymbol{A}^t t} \boldsymbol{Q} e^{\boldsymbol{A}t} \, \mathrm{d}t \tag{5.9}$$

then $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple. First note that since $\boldsymbol{A}$ is Hurwitz, the integral does indeed converge. We also have

$$\begin{aligned}
\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} &= \boldsymbol{A}^t\Big(\int_0^\infty e^{\boldsymbol{A}^t t} \boldsymbol{Q} e^{\boldsymbol{A}t} \, \mathrm{d}t\Big) + \Big(\int_0^\infty e^{\boldsymbol{A}^t t} \boldsymbol{Q} e^{\boldsymbol{A}t} \, \mathrm{d}t\Big)\boldsymbol{A} \\
&= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\big(e^{\boldsymbol{A}^t t} \boldsymbol{Q} e^{\boldsymbol{A}t}\big) \, \mathrm{d}t \\
&= e^{\boldsymbol{A}^t t} \boldsymbol{Q} e^{\boldsymbol{A}t}\big|_0^\infty = -\boldsymbol{Q},
\end{aligned}$$

as desired. We now show that $\boldsymbol{P}$ as defined is the *only* symmetric matrix for which $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple. Suppose that $\tilde{\boldsymbol{P}}$ also has the property that $(\boldsymbol{A}, \tilde{\boldsymbol{P}}, \boldsymbol{Q})$ is a Liapunov triple, and let $\boldsymbol{\Delta} = \tilde{\boldsymbol{P}} - \boldsymbol{P}$. Then one sees that $\boldsymbol{A}^t\boldsymbol{\Delta} + \boldsymbol{\Delta}\boldsymbol{A} = \boldsymbol{0}_{n\times n}$. If we let

$$\boldsymbol{\Lambda}(t) = e^{\boldsymbol{A}^t t} \boldsymbol{\Delta} e^{\boldsymbol{A}t},$$

then

$$\frac{\mathrm{d}\boldsymbol{\Lambda}(t)}{\mathrm{d}t} = e^{\boldsymbol{A}^t t}\big(\boldsymbol{A}^t\boldsymbol{\Delta} + \boldsymbol{\Delta}\boldsymbol{A}\big)e^{\boldsymbol{A}t} = \boldsymbol{0}_{n\times n}.$$

Therefore $\boldsymbol{\Lambda}(t)$ is constant, and since $\boldsymbol{\Lambda}(0) = \boldsymbol{\Delta}$, it follows that $\boldsymbol{\Lambda}(t) = \boldsymbol{\Delta}$ for all t. However, since $\boldsymbol{A}$ is Hurwitz, it also follows that $\lim_{t\to\infty} \boldsymbol{\Lambda}(t) = \boldsymbol{0}_{n\times n}$. Thus $\boldsymbol{\Delta} = \boldsymbol{0}_{n\times n}$, so that $\tilde{\boldsymbol{P}} = \boldsymbol{P}$.

(ii) If $\boldsymbol{P}$ is defined by (5.9) we have

$$\boldsymbol{x}^t\boldsymbol{P}\boldsymbol{x} = \int_0^\infty \big(e^{\boldsymbol{A}t}\boldsymbol{x}\big)^t \boldsymbol{Q}\big(e^{\boldsymbol{A}t}\boldsymbol{x}\big) \, \mathrm{d}t.$$

Therefore, if $\boldsymbol{Q}$ is positive-semidefinite, it follows that $\boldsymbol{P}$ is positive-semidefinite.

(iii) Here we employ a lemma.

1 Lemma *If $\boldsymbol{Q}$ is positive-semidefinite then $(\boldsymbol{A}, \boldsymbol{Q})$ is observable if and only if the matrix $\boldsymbol{P}$ defined by (5.9) is invertible.*

Proof First suppose that $(\boldsymbol{A}, \boldsymbol{Q})$ is observable and let $\boldsymbol{x} \in \ker(\boldsymbol{P})$. Then

$$\int_0^\infty \big(e^{\boldsymbol{A}t}\boldsymbol{x}\big)^t \boldsymbol{Q}\big(e^{\boldsymbol{A}t}\boldsymbol{x}\big) \, \mathrm{d}t = 0.$$

Since $\boldsymbol{Q}$ is positive-semidefinite, this implies that $e^{\boldsymbol{A}t}\boldsymbol{x} \in \ker(\boldsymbol{Q})$ for all t. Differentiating this inclusion with respect to t k times in succession gives $\boldsymbol{A}^k e^{\boldsymbol{A}t}\boldsymbol{x} \in \ker(\boldsymbol{Q})$ for any $k > 0$. Evaluating at $t = 0$ shows that $\boldsymbol{x}$ is in the kernel of the matrix

$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{Q}) = \left[\begin{array}{c} \boldsymbol{Q} \\ \hline \boldsymbol{QA} \\ \hline \vdots \\ \hline \boldsymbol{QA}^{n-1} \end{array}\right].$$

Since $(\boldsymbol{A}, \boldsymbol{Q})$ is observable, this implies that $\boldsymbol{x} = \boldsymbol{0}$. Thus we have shown that $\ker(\boldsymbol{P}) = \{\boldsymbol{0}\}$, or equivalently that $\boldsymbol{P}$ is invertible.

Now suppose that $\boldsymbol{P}$ is invertible. Then the expression

$$\int_0^\infty \big(e^{\boldsymbol{A}t}\boldsymbol{x}\big)^t \boldsymbol{Q}\big(e^{\boldsymbol{A}t}\boldsymbol{x}\big)\,\mathrm{d}t$$

is zero if and only if $\boldsymbol{x} = \boldsymbol{0}$. Since $\boldsymbol{Q}$ is positive-semidefinite, this means that the expression

$$\big(e^{\boldsymbol{A}t}\boldsymbol{x}\big)^t \boldsymbol{Q}\big(e^{\boldsymbol{A}t}\boldsymbol{x}\big)$$

is zero if and only if $\boldsymbol{x} = \boldsymbol{0}$. Since $e^{\boldsymbol{A}t}$ is invertible, this implies that $\boldsymbol{Q}$ must be positive-definite, and in particular, invertible. In this case, $(\boldsymbol{A}, \boldsymbol{Q})$ is clearly observable. ▼

With the lemma at hand, the remainder of the proof is straightforward. Indeed, from part (ii) we know that $\boldsymbol{P}$ is positive-semidefinite. The lemma now says that $\boldsymbol{P}$ is positive-definite if and only if $(\boldsymbol{A}, \boldsymbol{Q})$ is observable, as desired. ■

5.33 Example **(Example 5.4 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 flush out some aspects of Theorem 5.32.

1. $a > 0$ and $b > 0$: This is exactly the case when $\boldsymbol{A}$ is Hurwitz, so that part (i) of Theorem 5.32 implies that for any symmetric $\boldsymbol{Q}$ there is a unique symmetric $\boldsymbol{P}$ so that $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple. As we saw in the proof of Theorem 5.32, one can determine $\boldsymbol{P}$ with the formula

$$\boldsymbol{P} = \int_0^\infty e^{\boldsymbol{A}^t t}\boldsymbol{Q}e^{\boldsymbol{A}t}\,\mathrm{d}t. \tag{5.10}$$

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 $e^{\boldsymbol{A}t}$. 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

$$e^{\boldsymbol{t}} = e^{-t}\begin{bmatrix} \cos t + \sin t & \sin t \\ -2\sin t & \cos t - \sin t \end{bmatrix}$$

In this case one can directly apply (5.10) 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

$$e^{\boldsymbol{A}t} = e^{-t}\begin{bmatrix} 1+t & t \\ -t & 1-t \end{bmatrix}.$$

Again, a direct computation using (5.10) 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 observable. Therefore, part (iii) of Theorem 5.32 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

$$\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{Q}. \tag{5.11}$$

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 E5.16). This demonstrates that when $\boldsymbol{A}$ is not Hurwitz, part (i) of Theorem 5.32 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 $\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = \boldsymbol{0}_{2\times 2}$. Thus if $\boldsymbol{P}$ is any solution to (5.11) 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 5.30, if

$$\boldsymbol{P} = \begin{bmatrix} a^2 & a \\ a & 2 \end{bmatrix},$$

then $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple. What we have shown i that $(\boldsymbol{A}, \boldsymbol{P} + \boldsymbol{P}_0, \boldsymbol{Q})$ is also a Liapunov triple. Thus part (i) of Theorem 5.32 can fail in the matter of uniqueness when $\boldsymbol{A}$ is not Hurwitz. •

5.5 Identifying polynomials with roots in $\mathbb{C}_-$

From our discussion of Section 5.2 we see that it is very important that T_Σ have poles only in the negative half-plane. However, checking that such a condition holds may not be so easy. One way to do this is to establish conditions on the coefficients of the denominator polynomial of T_Σ (after making pole/zero cancellations, of course). In this section, we present three methods for doing exactly this. We also look at a test for the poles lying 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***.

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[2] (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.

5.5.1 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 $\left[\begin{smallmatrix} a_{11} & a_{1,k+1} \\ a_{21} & a_{2,k+1} \end{smallmatrix}\right]$. 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 $P \in \mathbb{R}[s]$. Define two polynomials $P_+, P_- \in \mathbb{R}[s]$ as the even and odd part of P. To be clear about this, if

$$P(s) = p_0 + p_1 s + p_2 s^2 + p_3 s^3 + \cdots + p_{n-1} s^{n-1} + p_n s^n,$$

then

$$P_+(s) = p_0 + p_2 s + p_4 s^2 + \dots, \quad P_-(s) = p_1 + p_3 s + p_5 s^2 + \dots.$$

Note then that $P(s) = P_+(s^2) + sP_-(s^2)$. 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

[2]Maxwell, of course, is better known for his famous equations of electromagnetism.

with the coefficients of lowest powers of s, and increasing to higher powers of s. Thus the first three rows of $R(P)$ are

$$\begin{array}{cccc} p_0 & p_2 & \cdots & p_{2k} & \cdots \\ p_1 & p_3 & \cdots & p_{2k+1} & \cdots \\ p_1p_2 - p_0p_3 & p_1p_4 - p_0p_5 & \cdots & p_1p_{2k+2} - p_0p_{2k+3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

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.

5.34 Theorem (**Routh** [1877]) *A polynomial*

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \in \mathbb{R}[s]$$

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

$$P_2(s) = s^{-1}\big(P_1(0)P_0(s) - P_0(0)P_1(s)\big).$$

Note that the constant coefficient of $P_1(0)P_0(s) - P_0(0)P_1(s)$ is zero, so this does indeed define P_2 as a polynomial. Now inductively define

$$P_k(s) = s^{-1}\big(P_{k-1}(0)P_{k-2}(s) - P_{k-2}(0)P_{k-1}(s)\big)$$

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 P_k with the constant coefficient in the first column. Thus the hypothesis of the theorem is equivalent to the condition that $P_0(0), P_1(0), \ldots, P_n(0)$ 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 $k = 0, 1, 2$. 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

$$P_{k-1}(s) = p_{k-1,0} + p_{k-1,1}s + \cdots$$

and

$$P_k = p_{k,0} + p_{k,1}s + \cdots,$$

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 $P(s) = P_0(s^2) + sP_1(s^2)$ and define $Q \in \mathbb{R}[s]$ by $Q(s) = P_1(s^2) + sP_2(s^2)$. One may readily verify that $\deg(Q) \leq n-1$. Indeed, in the proof of Theorem 5.36, 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* $p_n > 0$*;*

(ii) Q *is Hurwitz,* $q_{n-1} > 0$*, and* $P(0) > 0$*.*

Proof We have already noted that $P(s) = P_0(s^2) + sP_1(s^2)$. We may also compute

$$Q(s) = P_1(s^2) + s^{-1}\big(P_1(0)P_0(s^2) - P_0(0)P_1(s^2)\big). \tag{5.12}$$

For $\lambda \in [0,1]$ define $Q_\lambda(s) = (1-\lambda)P(s) + \lambda Q(s)$, and compute

$$Q_\lambda(s) = \big((1-\lambda) + s^{-1}\lambda P_1(0)\big)P_0(s^2) + \big((1-\lambda)s + \lambda - s^{-1}\lambda P_0(0)\big)P_1(s^2).$$

The polynomials $P_0(s^2)$ and $P_1(s^2)$ 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 $i\omega_0$ is a nonzero imaginary root then we must have

$$\big((1-\lambda) - i\omega_0^{-1}\lambda P_1(0)\big)P_0(-\omega_0^2) + \big((1-\lambda)i\omega_0 + \lambda + i\omega_0^{-1}\lambda P_0(0)\big)P_1(-\omega_0^2) = 0.$$

Balancing real and imaginary parts of this equation gives

$$\begin{aligned} &(1-\lambda)P_0(-\omega_0^2) + \lambda P_1(-\omega_0^2) = 0 \\ &\lambda\omega_0^{-1}\big(P_0(0)P_1(-\omega_0^2) - P_1(0)P_0(-\omega_0^2)\big) + \omega_0(1-\lambda)P_1(-\omega_0^2). \end{aligned} \tag{5.13}$$

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 $P(0) > 0$ and $Q(0) > 0$, all imaginary axis roots $i\omega_0$ 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, $Z_0 \subset \mathbb{C}_-$. Our previous computations then show that $Z_\lambda \cap i\mathbb{R} = \emptyset$ 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 E5.18 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 E5.18 and $p_n > 0$.

(ii) $\Longrightarrow$ (i) Let us adopt the notation $N(\lambda)$ and Z_λ from the previous part of the proof. Since Q is Hurwitz, $Z_1 \subset \mathbb{C}_-$. 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 $P(s) = \tilde{P}(s)(s - z_0)$ where $\tilde{P}$ is Hurwitz. By Exercise E5.18 it follows that all coefficients of $\tilde{P}$ are positive. If we write

$$\tilde{P} = \tilde{p}_{n-1}s^{n-1} + \tilde{p}_{n-2}s^{n-2} + \cdots + \tilde{p}_1 s + \tilde{p}_0,$$

then

$$P(s) = \tilde{p}_{n-1}s^n + (\tilde{p}_{n-2} - z_0\tilde{p}_{n-1})s^{n-1} + \cdots + (\tilde{p}_0 - z_0\tilde{p}_1)s - \tilde{p}_0 z_0.$$

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

$$Q_k(s) = P_k(s^2) + sP_{k+1}(s^2), \quad k = 1, \ldots, n,$$

are Hurwitz and that $\deg(Q_k) = n - k$, $k = 1, \ldots, n$. What's more, the coefficient of s^{n-k} is positive in Q_k. Now, by Exercise E5.18 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 s being positive, from Exercise E5.18 it follows that the coefficient of s 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, $k = 1, \ldots, n$, must also be Hurwitz, with $\deg(Q_k) = n - k$ where the coefficient of s^{n-k} in Q_k is positive. In particular, by Exercise E5.18, all coefficients of Q_{k_0-1} are positive. However, since $Q_{k_0-1}(s) = P_{k_0-1}(s^2) + sP_{k_0}(s^2)$ it follows that the coefficient of s 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.

5.35 Example Let us apply the criteria to the simplest nontrivial example possible: $P(s) = s^2 + as + b$. 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's 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. •

5.5.2 The Hurwitz criterion

The method we discuss in this section is work of Adolf Hurwitz (1859–1919). The key ingredient in the Hurwitz construction is a matrix formed from the coefficients of a polynomial

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \in \mathbb{R}[s].$$

We denote the ***Hurwitz matrix*** by $\boldsymbol{H}(P) \in \mathbb{R}^{n\times n}$ 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 $\boldsymbol{H}(P)_k \in \mathbb{R}^{k\times k}$, $k = 1, \ldots, n$, to be the matrix of elements $\boldsymbol{H}(P)_{ij}$, $i, j = 1, \ldots, k$. 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.

5.36 Theorem (**Hurwitz** [1895]) *A polynomial*

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \in \mathbb{R}[s]$$

is Hurwitz if and only if the n ***Hurwitz determinants*** *$\Delta_1, \ldots, \Delta_n$ are positive.*

Proof Let us begin by resuming with the notation from the proof of Theorem 5.34. In particular, we recall the definition of $Q(s) = P_1(s^2) + sP_2(s^2)$. 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

$$Q(s) = p_1 + (p_1p_2 - p_0p_3)s + p_3s^2 + (p_1p_4 - p_0p_5)s^3 + \cdots.$$

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 $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ 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

$$\boldsymbol{H}(P)\boldsymbol{T} = \begin{bmatrix} & & & \vdots \\ & \boldsymbol{H}(Q) & & p_4 \\ & & & p_2 \\ 0 & \cdots & 0 & p_0 \end{bmatrix}. \tag{5.14}$$

We now let $\Delta_1, \ldots, \Delta_n$ be the determinants defined above and let $\tilde{\Delta}_1, \ldots, \tilde{\Delta}_{n-1}$ be the similar determinants corresponding to $\boldsymbol{H}(Q)$. A straightforward computation using (5.14) gives the following relationships between the Δ's and the $\tilde{\Delta}$'s:

$$\begin{aligned} \Delta_1 &= p_1 \\ \Delta_{k+1} &= \begin{cases} p_1^{-\lfloor \frac{k}{2} \rfloor}\tilde{\Delta}_k, & k \text{ even} \\ p_1^{-\lceil \frac{k}{2} \rceil}\tilde{\Delta}_k, & k \text{ odd} \end{cases}, \qquad k = 1, \ldots, n-1, \end{aligned} \tag{5.15}$$

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 E5.18, it follows that $p_1 > 0$ so that $\Delta_1 > 0$. By Lemma 2 of Theorem 5.34 it also follows that Q is Hurwitz. Thus $\tilde{\Delta}_1 > 0$. A trivial induction argument on $n = \deg(P)$ then shows that $\Delta_2, \ldots, \Delta_n > 0$.

Now suppose that one of $\Delta_1, \ldots, \Delta_n$ is nonpositive and that P is Hurwitz. Since Q is then Hurwitz by Lemma 2 of Theorem 5.34, 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.

5.37 Example **(Example 5.35 cont'd)** Let us apply the criteria to our simple example of $P(s) = s^2 + as + b$. 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 5.35. •

5.5.3 The Hermite criterion

We next look at a manner of determining whether a polynomial is Hurwitz which makes contact with the Liapunov methods of Section 5.4. This method is due to Charles Hermite (1822–1901) [see Hermite 1854]. 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

$$\boldsymbol{P}(P) = \begin{bmatrix} p_1 p_2 & 0 \\ 0 & p_0 p_1 \end{bmatrix},$$

for $n = 3$ we have

$$\boldsymbol{P}(P) = \begin{bmatrix} p_2 p_3 & 0 & p_0 p_3 \\ 0 & p_1 p_2 - p_0 p_3 & 0 \\ p_0 p_3 & 0 & p_0 p_1 \end{bmatrix},$$

and for $n = 4$ we have

$$\boldsymbol{P}(P) = \begin{bmatrix} p_3 p_4 & 0 & p_1 p_4 & 0 \\ 0 & p_2 p_3 - p_1 p_4 & 0 & p_0 p_3 \\ p_1 p_4 & 0 & p_1 p_2 - p_0 p_3 & 0 \\ 0 & p_0 p_3 & 0 & p_0 p_1 \end{bmatrix}.$$

The following theorem gives necessary and sufficient conditions for P to be Hurwitz based on its Hermite matrix. The slick proof using Liapunov methods comes from the paper of Parks [1962].

5.38 Theorem *A polynomial*

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \in \mathbb{R}[s]$$

is Hurwitz if and only if $\boldsymbol{P}(P)$ *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 5.30(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 5.32(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.

5.39 Example **(Example 5.35 cont'd)** 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 5.38, 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 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

$$\boldsymbol{C}(P) = \begin{bmatrix} p_1p_2 \end{bmatrix}, \quad \boldsymbol{D}(P) = \begin{bmatrix} p_0p_1 \end{bmatrix},$$

for $n = 3$ we have

$$\boldsymbol{C}(P) = \begin{bmatrix} p_2p_3 & p_0p_3 \\ p_0p_3 & p_0p_1 \end{bmatrix}, \quad \boldsymbol{D}(P) = \begin{bmatrix} p_1p_2 - p_0p_3 \end{bmatrix},$$

and for $n = 4$ we have

$$\boldsymbol{C}(P) = \begin{bmatrix} p_3p_4 & p_1p_4 \\ p_1p_4 & p_1p_2 - p_0p_3 \end{bmatrix}, \quad \boldsymbol{D}(P) = \begin{bmatrix} p_2p_3 - p_1p_4 & p_0p_3 \\ p_0p_3 & p_0p_1 \end{bmatrix}.$$

Let us record a useful property of the matrices $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$, noting that they are symmetric.

5.40 Lemma *$\boldsymbol{P}(P)$ is positive-definite if and only if both $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$ 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

$$\boldsymbol{x}^t \boldsymbol{P}(P) \boldsymbol{x} = \boldsymbol{x}_{\text{odd}}^t \boldsymbol{C}(P) \boldsymbol{x}_{\text{odd}} + \boldsymbol{x}_{\text{even}}^t \boldsymbol{D}(P) \boldsymbol{x}_{\text{even}}. \tag{5.16}$$

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 (5.16) 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. Our proof follows that of Anderson [1972].

5.41 Theorem *A polynomial*

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \in \mathbb{R}[s]$$

is Hurwitz if and only if any one of the following conditions holds:

(i) $p_{2k} > 0$, $k \in \{0, 1, \dots, \lfloor \frac{n-1}{2} \rfloor\}$ *and* $\boldsymbol{C}(P)$ *is positive-definite;*

(ii) $p_{2k} > 0$, $k \in \{0, 1, \dots, \lfloor \frac{n-1}{2} \rfloor\}$ *and* $\boldsymbol{D}(P)$ *is positive-definite;*

(iii) $p_0 > 0$, $p_{2k+1} > 0$, $k \in \{0, 1, \dots, \lfloor \frac{n-2}{2} \rfloor\}$ *and* $\boldsymbol{C}(P)$ *is positive-definite;*

(iv) $p_0 > 0$, $p_{2k+1} > 0$, $k \in \{0, 1, \dots, \lfloor \frac{n-2}{2} \rfloor\}$ *and* $\boldsymbol{D}(P)$ *is positive-definite.*

Proof First suppose that P is Hurwitz. Then all coefficients are positive (see Exercise E5.18) and $\boldsymbol{P}(P)$ is positive-definite by Theorem 5.38. This implies that $\boldsymbol{C}(P)$ and $\boldsymbol{D}(P)$ are positive-definite by Lemma 5.40, 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

$$\boldsymbol{C}(P)^{1/2} \boldsymbol{A}_1(P) \boldsymbol{C}(P)^{-1/2} = -\boldsymbol{C}(P)^{-1/2} \boldsymbol{D}(P) \boldsymbol{C}(P)^{-1/2}. \tag{5.17}$$

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 (5.17) it now follows that $\boldsymbol{D}(P)$ is positive-definite, and so P is Hurwitz by Lemma 5.40 and Theorem 5.38.

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 E5.20). Also, the coefficients for P^{-1} are obtained by reversing 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

$$\boldsymbol{C}(P)\boldsymbol{A}_2(P) = -\begin{bmatrix} \boldsymbol{D}(P) & \boldsymbol{0} \\ \boldsymbol{0}^t & 0 \end{bmatrix}. \tag{5.18}$$

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

$$s^{(n+1)/2} + \frac{p_{n-2}}{p_n} s^{(n+1)/2-1} + \cdots + \frac{p_1}{p_n} s.$$

This polynomial clearly has a zero root. However, since (iii) holds, for positive real values of s 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 (5.18) 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 5.40 and Theorem 5.38.

Case 4: We assume (i) and that n is odd. As in Case 2, define $P^{-1}(s) = s^n P(\frac{1}{s})$. 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 E5.20.

Case 5: We assume (ii) and that n is odd. For $\epsilon > 0$ define $P_\epsilon \in \mathbb{R}[s]$ by $P_\epsilon(s) = (s+\epsilon)P(s)$. Thus the degree of P_ϵ is now even. Indeed,

$$P_\epsilon(s) = p_n s^{n+1} + (p_{n-1} + \epsilon p_n)s^n + \cdots + (p_0 + \epsilon p_1)s + \epsilon p_0.$$

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 $\boldsymbol{D}(P_\epsilon)$ 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 $P^{-1}(s) = s^n P(\frac{1}{s})$ 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 E5.20.

Case 7: We assume (iii) and that n is even. As with Case 5, we define $P_\epsilon(s) = (s+\epsilon)P(s)$ 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 5.40 and Theorem 5.36. Thus P is itself also Hurwitz.

Case 8: We assume (iv) and that n is even. Taking $P^{-1}(s) = s^n P(\frac{1}{s})$ 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. ■

5.5.4 The Liénard-Chipart criterion

Although less well-known than the criterion of Routh and Hurwitz, the test we give due to Liénard and Chipart [1914][3] has the advantage of delivering fewer determinantal inequalities to test. This results from their being a dependence on some of the Hurwitz determinants. This is given thorough discussion by Gantmacher [1959b]. Here we state the result, and give a proof due to Anderson [1972] that is more elementary than that of Gantmacher.

[3]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.

5.42 Theorem *A polynomial*

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \in \mathbb{R}[s]$$

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 Theorems 5.28 and 5.41 and after one checks 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 $\Delta_2, \Delta_4, \ldots$. This observation is made by a computation which we omit, and appears to be first been noticed by Fujiwara [1915]. ■

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 $n = 2$.

5.43 Example **(Example 5.35 cont'd)** We consider the polynomial $P(s) = s^2 + as + b$. Recall that the Hurwitz determinants were computed in Example 5.37:

$$\Delta_1 = a, \quad \Delta_2 = ab.$$

Let us write down the four conditions of Theorem 5.42:

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 E5.22 to see how the test can be put to good use. •

5.5.5 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

$$P(s) = p_n s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0, \tag{5.19}$$

one may know that the coefficients satisfy inequalities of the form

$$p_i^{\min} \leq p_i \leq p_i^{\max}, \quad i = 0, 1, \ldots, n. \tag{5.20}$$

In this case, the following remarkable theorem of Kharitonov [1978] gives a simple test for the stability of the polynomial for all possible values for the coefficients. 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].

5.44 Theorem *Given a polynomial of the form (5.19) with the coefficients satisfying the inequalities (5.20), 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 E5.18, 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 5.4 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.

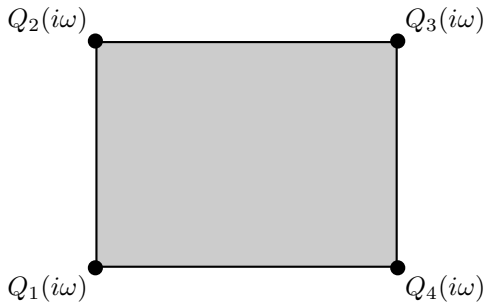


Figure 5.4 $R(\omega)$

2 Lemma *If $P \in \mathbb{R}[s]$ is monic and Hurwitz with $\deg(P) \geq 1$, then $\measuredangle P(i\omega)$ 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

$$\measuredangle P(i\omega) = \sum_{j=1}^{n} \measuredangle(i\omega + |\sigma_j| - i\omega_j) = \sum_{j=1}^{n} \arctan\Big(\frac{\omega - \omega_j}{|\sigma_j|}\Big).$$

Since $|\sigma_j| > 0$, each term in the sum is continuous and strictly increasing, and thus so too is $\measuredangle P(i\omega)$. ▼

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

$$\lambda Q_1 + (1 - \lambda)P_{\boldsymbol{p}_0} \tag{5.21}$$

is of the form $P_{\boldsymbol{p}_\lambda}$ for some allowable $\boldsymbol{p}_\lambda$. Indeed, the equation (5.21) 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 $P_{\boldsymbol{p}}$ are Hurwitz for allowable $\boldsymbol{p}$. ■

5.45 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.
3. 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.
4. A proof of Kharitonov's theorem, using Liapunov methods (see Section 5.4), is given by Mansour and Anderson [1993]. •

Let us apply the Kharitonov test in the simplest case when $n = 2$.

5.46 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 5.44 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 5.44. •

5.6 Summary

The matter of stability is, of course, of essential importance. What we have done in this chapter is quite simple, so let us outline the major facts.

1. It should be understood that internal stability is a notion relevant only to SISO linear systems. The difference between stability and asymptotic stability should be understood.
2. The conditions for internal stability are generally simple. The only subtleties occur when there are repeated eigenvalues on the imaginary axis. All of this needs to be understood.
3. BIBO stability is really the stability type of most importance in this book. One should understand when it happens. One should also know how, when it does not happen, to produce an unbounded output with a bounded input.

4. Norm characterisations if BIBO stability provide additional insight, and offer a clarifying language with which to organise BIBO stability. Furthermore, some of the technical results concerning such matters will be useful in discussions of performance in Section 9.3 and of robustness in Chapter 15.
5. One should be able to apply the Hurwitz and Routh criteria freely.
6. The Liapunov method offer a different sort of characterisation of internal stability. One should be able to apply the theorems presented.

Exercises

E5.1 Consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{D})$ of Example 5.4, i.e., with

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix}.$$

By explicitly computing a basis of solutions to $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$ in each case, verify by direct calculation the conclusions of Example 5.4. If you wish, you may choose specific values of the parameters a and b for each of the eight cases of Example 6.4. Make sure that you cover sub-possibilities in each case that might arise from eigenvalues being real or complex.

E5.2 Determine the internal stability of the linearised pendulum/cart system of Exercise E1.5 for each of the following cases:

(a) the equilibrium point $(0, 0)$;
(b) the equilibrium point $(0, \pi)$.

E5.3 For the double pendulum system of Exercise E1.6, determine the internal stability for the linearised system about the following equilibria:

(a) the equilibrium point $(0, 0, 0, 0)$;
(b) the equilibrium point $(0, \pi, 0, 0)$;
(c) the equilibrium point $(\pi, 0, 0, 0)$;
(d) the equilibrium point $(\pi, \pi, 0, 0)$.

E5.4 For the coupled tank system of Exercise E1.11 determine the internal stability of the linearisation.

E5.5 Determine the internal stability of the coupled mass system of Exercise E1.4, both with and without damping. You may suppose that the mass and spring constant are positive and that the damping factor is nonnegative.

E5.6 Consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ defined by

$$\boldsymbol{A} = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

for $\sigma \in \mathbb{R}$ and $\omega > 0$.

(a) For which values of the parameters σ and ω is Σ spectrally stable?
(b) For which values of the parameters σ and ω is Σ internally stable? Internally asymptotically stable?
(c) For zero input, describe the qualitative behaviour of the states of the system when the parameters σ and ω are chosen so that the system is internally stable but not internally asymptotically stable.
(d) For zero input, describe the qualitative behaviour of the states of the system when the parameters σ and ω are chosen so that the system is internally unstable.
(e) For which values of the parameters σ and ω is Σ BIBO stable?
(f) When the system is not BIBO stable, determine a bounded input that produces an unbounded output.

E5.7 Consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$ defined by

$$\boldsymbol{A} = \begin{bmatrix} \sigma & \omega & 1 & 0 \\ -\omega & \sigma & 0 & 1 \\ 0 & 0 & \sigma & \omega \\ 0 & 0 & -\omega & \sigma \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

where $\sigma \in \mathbb{R}$ and $\omega > 0$.

(a) For which values of the parameters σ and ω is Σ spectrally stable?

(b) For which values of the parameters σ and ω is Σ internally stable? Internally asymptotically stable?

(c) For zero input, describe the qualitative behaviour of the states of the system when the parameters σ and ω are chosen so that the system is internally stable but not internally asymptotically stable.

(d) For zero input, describe the qualitative behaviour of the states of the system when the parameters σ and ω are chosen so that the system is internally unstable.

(e) For which values of the parameters σ and ω is Σ BIBO stable?

(f) When the system is not BIBO stable, determine a bounded input that produces an unbounded output.

E5.8 Show that (N, D) is BIBO stable if and only if for every $u \in \mathrm{L}_\infty[0, \infty)$, the function y satisfying $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)u(t)$ also lies in $\mathrm{L}_\infty[0, \infty)$.

E5.9 Determine whether the pendulum/cart system of Exercises E1.5 and E2.4 is BIBO stable in each of the following linearisations:

(a) the equilibrium point $(0, 0)$ with cart position as output;

(b) the equilibrium point $(0, 0)$ with cart velocity as output;

(c) the equilibrium point $(0, 0)$ with pendulum angle as output;

(d) the equilibrium point $(0, 0)$ with pendulum angular velocity as output;

(e) the equilibrium point $(0, \pi)$ with cart position as output;

(f) the equilibrium point $(0, \pi)$ with cart velocity as output;

(g) the equilibrium point $(0, \pi)$ with pendulum angle as output;

(h) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output.

E5.10 Determine whether the double pendulum system of Exercises E1.6 and E2.5 is BIBO stable in each of the following cases:

(a) the equilibrium point $(0, 0, 0, 0)$ with the pendubot input;

(b) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;

(c) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;

(d) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;

(e) the equilibrium point $(0, 0, 0, 0)$ with the acrobot input;

(f) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;

(g) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;

(h) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input.

In each case, use the angle of the second link as output.

E5.11 Consider the coupled tank system of Exercises E1.11 and E2.6. Determine the BIBO stability of the linearisations in the following cases:

(a) the output is the level in tank 1;

(b) the output is the level in tank 2;
(c) the output is the difference in the levels,

E5.12 Consider the coupled mass system of Exercise E1.4 and with inputs as described in Exercise E2.19. For this problem, leave α as an arbitrary parameter. We consider a damping force of a very special form. We ask that the damping force on each mass be given by $-d(\dot{x}_1 + \dot{x}_2)$. The mass and spring constant may be supposed positive, and the damping constant is nonnegative.

(a) Represent the system as a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{D})$—note that in Exercises E1.4 and E2.19, everything except the matrix $\boldsymbol{A}$ has already been determined.
(b) Determine for which values of α, mass, spring constant, and damping constant the system is BIBO stable.
(c) Are there any parameter values for which the system is BIBO stable, but for which you might not be confident with the system's state behaviour? Explain your answer.

E5.13 Let (N, D) be a SISO linear system in input/output form. In this exercise, if $u\colon [0, \infty) \to \mathbb{R}$ is an input, y_u will be the output defined so that $\hat{y}_u(s) = T_{N,D}(s)\hat{u}(s)$.

(a) For $\epsilon > 0$ define an input u_ϵ by

$$u_\epsilon(t) = \begin{cases} \frac{1}{\epsilon}, & t \in [0, \epsilon] \\ 0, & \text{otherwise.} \end{cases}$$

Determine
(i) $\lim_{\epsilon \to 0} \|y_{u_\epsilon}\|_2$;
(ii) $\lim_{\epsilon \to 0} \|y_{u_\epsilon}\|_\infty$;
(iii) $\lim_{\epsilon \to 0} \operatorname{pow}(y_{u_\epsilon})$.

(b) If $u(t) = \sin(\omega t)$ determine
(i) $\|y_u\|_2$;
(ii) $\|y_u\|_\infty$;
(iii) $\operatorname{pow}(y_u)$.

E5.14 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system.

(a) Show that if $\boldsymbol{A} + \boldsymbol{A}^t$ is negative-semidefinite then Σ is internally stable.
(b) Show that if $\boldsymbol{A} + \boldsymbol{A}^t$ is negative-definite then Σ is internally asymptotically stable.

E5.15 Let $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ be a Liapunov triple for which $\boldsymbol{P}$ and $\boldsymbol{Q}$ are positive-definite. Show that $\boldsymbol{A}$ is Hurwitz.

E5.16 Let

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & a \end{bmatrix}$$

for $a \geq 0$. Show that if $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{Q})$ is a Liapunov triple for which $\boldsymbol{Q}$ is positive-semidefinite, then $(\boldsymbol{A}, \boldsymbol{Q})$ is not observable.

E5.17 Consider the polynomial $P(s) = s^3 + as^2 + bs + c$.

(a) Use the Routh criteria to determine conditions on the coefficients a, b, and c that ensure that the polynomial P is Hurwitz.

(b) Use the Hurwitz criteria to determine conditions on the coefficients a, b, and c that ensure that the polynomial P is Hurwitz.
(c) Verify that the conditions on the coefficients from parts (a) and (b) are equivalent.
(d) Give an example of a polynomial of the form of P that is Hurwitz.
(e) Give an example of a polynomial of the form of P for which all coefficients are strictly positive, but that is not Hurwitz.

E5.18 A useful necessary condition for a polynomial to have all roots in $\mathbb{C}_-$ is given by the following theorem.

Theorem *If the polynomial*

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \in \mathbb{R}[s]$$

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.

The Routh/Hurwitz method gives a means of determining whether the roots of a polynomial are stable, but gives no indication of "how stable" they are. In the following exercise, you will examine conditions for a polynomial to be stable, and with some margin for error.

E5.19 Let $P(s) = s^2 + as + b$, and for $\delta > 0$ denote

$$R_\delta = \{s \in \mathbb{C} \mid \operatorname{Re}(s) < -\delta\}.$$

Thus R_δ consists of those points lying a distance at least δ to the left of the imaginary axis.

(a) Using the Routh criterion as a basis, derive necessary and sufficient conditions for all roots of P to lie in R_δ.
Hint: *The polynomial* $\tilde{\mathrm{P}}(\mathrm{s}) = \mathrm{P}(\mathrm{s} + \delta)$ *must be Hurwitz.*
(b) Again using the Routh criterion as a basis, state and prove necessary and sufficient conditions for the roots of a general polynomial to lie in R_δ.

Note that one can do this for any of the methods we have provided for characterising Hurwitz polynomials.

E5.20 Consider a polynomial

$$P(s) = p_n s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \in \mathbb{R}[s]$$

with $p_0, p_n \neq 0$, and define $P^{-1} \in \mathbb{R}[s]$ by $P^{-1}(s) = s^n P(\frac{1}{s})$.

(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.

E5.21 For the following two polynomials,

(a) $P(s) = s^3 + as^2 + bs + c$,
(b) $P(s) = s^4 + as^3 + bs^2 + cs + d$,

do the following:

1. Using principal minors to test positive-definiteness, write the conditions of the Hermite criterion, Theorem 5.38, for P to be Hurwitz.

2. Again using principal minors to test positive-definiteness, write the four conditions of the reduced Hermite criterion, Theorem 5.41, for P to be Hurwitz, and ascertain which is the least restrictive.

E5.22 For the following two polynomials,

(a) $P(s) = s^3 + as^2 + bs + c$,

(b) $P(s) = s^4 + as^3 + bs^2 + cs + d$,

write down the four conditions of the Liénard-Chipart criterion, Theorem 5.42, and determine which is the least restrictive.

E5.23 Consider a general degree three polynomial

$$P(s) = s^3 + as^2 + bs + c,$$

where the coefficients satisfy

$$(a, b, c) \in [a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}] \times [c_{\min}, c_{\max}]. \tag{E5.1}$$

Use Kharitonov's test, Theorem 5.44, to give conditions on the bounds for the intervals for a, b, and c so that P is Hurwitz for all coefficients satisfying (E5.1).

Chapter 6

Interconnections and feedback

We now begin to enter into the more "design" oriented parts of the course. The concept of feedback is central to much of control, and we will be employing the analysis tools developed in the time-domain, the s-domain, and the frequency domain to develop ways of evaluating and designing feedback control systems. The value of feedback is at the same time obvious and mysterious. It is clear that it ought to be employed, but it often has effects that are subtle. Somehow the most basic feedback scheme is the PID controller that we discuss in Section 6.5. Here we can get an idea of how feedback can effect a closed-loop system.

You may wish to take a look at the DC motor system we talked about in Section 1.2 in order to see a very concrete display of the disadvantages of open-loop control, and how this can be repaired to advantage with a closed-loop scheme.

Contents

6.1 Signal flow graphs

We have seen in our discussion of block diagrams in Section 3.1 that one might be interested in connecting a bunch of transfer functions like the one depicted in Figure 3.7. In this section we indicate how one does this in a systematic and general way. This will enable us to provide useful results on stability of such interconnected systems in Section 6.2.3. The advantage of doing this systematically is that we can provide results that will hold for *any* block diagram setup, not just a few standard ones.

The signal flow graph was first studied systematically in control by Mason(1953, 1953). Many of the issues surrounding signal flow graphs are presented nicely in the paper of Lynch [1961]. A general discussion of applications of graph theory may be found in the book of Chen [1976]. Here the signal flow graph can be seen as a certain type of graph, and its properties are revealed in this context. Our presentation will often follow that of Zadeh and Desoer [1979].

6.1.1 Definitions and examples

The notion of a signal flow graph can be seen as a modification of the concept of a block diagram The idea is to introduce nodes to represent the signals in a system, and then connect the nodes with branches, and assign to each branch a rational function that performs the duty of a block in a block diagram. For example, Figure 3.1 would simply appear as shown in Figure 6.1. Before we proceed to further illustrations of how to construct signal flow

$$x_1 \xrightarrow{R} x_2$$

Figure 6.1 A simple signal flow graph

graphs along the lines of what we did with block diagrams, let's say just what we are talking about. Part of the point of signal flow graphs is that we can do this in a precise way.

6.1 Definition Denote $\boldsymbol{n} = \{1, 2, \dots, n\}$ and let $\mathcal{I} \subset \boldsymbol{n} \times \boldsymbol{n}$.

(i) A ***signal flow graph with interconnections*** $\mathcal{I}$ is a pair $(\mathcal{S}, \mathcal{G})$ where $\mathcal{S}$ is a collection $\{x_1, \dots, x_n\}$ of ***nodes*** or ***signals***, and

$$\mathcal{G} = \{G_{ij} \in \mathbb{R}(s) \mid (i, j) \in \mathcal{I}\}$$

is a collection of rational functions that we call ***branches*** or ***gains***. The branch G_{ij} ***originates*** from the node x_j and ***terminates*** at the node x_i.

(ii) A node x_i is a ***sink*** if no branches originate from x_i.

(iii) A node x_i is a ***source*** if no branches terminate at x_i.

(iv) A source x_i is an ***input*** if only one branch originates from x_i, and the gain of this branch is 1.

(v) A sink x_i is an ***output*** if only one branch terminates at x_i and this branch has weight 1. •

For example, for the simple signal flow graph of Figure 6.1, we have $\boldsymbol{n} = \{1, 2\}$, $\mathcal{I} = \{(1, 2)\}$, $\mathcal{S} = \{x_1, x_2\}$, and $\mathcal{G} = \{G_{21} = R\}$. The node x_1 is a source and the node x_2 is a sink. In this case note that we do not have the branch from x_2 to x_1. This is why we define $\mathcal{I}$ as we do—we will almost never want all n^2 possible interconnections, and those that we *do* want

are specified by $\mathcal{I}$. Note that we assume that at most one branch can connect two given nodes. Thus, if we have a situation like that in Figure 6.2, we will replace this with the

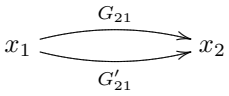


Figure 6.2 Two branches connecting the same nodes

situation in Figure 6.3.

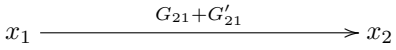


Figure 6.3 Branches added to give a single branch

If you think this obtuse and abstract, you are right. But if you work at it, you will see why we make the definitions as we do. Perhaps a few more examples will make things clearer.

6.2 Examples

1. The signal flow graph corresponding to the series block diagram of Figure 3.2 is shown in Figure 6.4. Note that here we have $\boldsymbol{n} = \{1, 2, 3\}$, $\mathcal{I} = \{(1, 2), (2, 3)\}$, $\mathcal{S} = \{x_1, x_2, x_3\}$, and

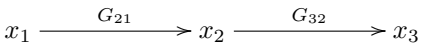


Figure 6.4 The signal flow graph for a series interconnection

$\mathcal{G} = \{G_{21}, G_{32}\}$. The node x_1 is a source and x_3 is a sink. There are a possible $n^2 = 9$ interconnections, and we only have two of them realised in this graph. The transfer function from x_1 to x_2 is simply G_{21}; that is, $x_2 = G_{21}x_1$. We also read off $x_3 = G_{32}x_2$, and so $x_3 = G_{21}G_{32}x_3$.

2. We now look at the representation of a parallel interconnection. The signal flow graph is shown in Figure 6.5, and we have $\boldsymbol{n} = \{1, 2, 3, 4\}$, $\mathcal{I} = \{(1, 2), (1, 3), (2, 4), (3, 4)\}$,

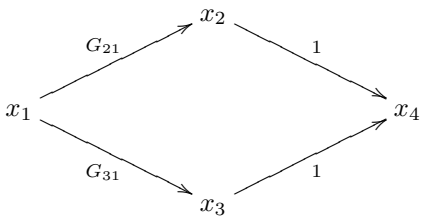


Figure 6.5 The signal flow graph for a parallel interconnection

$\mathcal{S} = \{x_1, x_2, x_3, x_4\}$, and $\mathcal{G} = \{G_{21}, G_{31}, G_{42} = 1, G_{43} = 1\}$. One sees that x_1 is a source and x_4 is a sink. From the graph we read the relations $x_2 = G_{21}x_1$, $x_3 = G_{31}x_1$, and $x_4 = x_2 + x_3$. This gives $x_4 = (G_{21} + G_{31})x_1$.

3. The preceding two signal flow graphs are very simple in some sense. To obtain the transfer function from the signal flow graph is a matter of looking at the graph and applying the obvious rules. Let us now look at a "feedback loop" as a signal flow graph. The graph is in Figure 6.6, and we have $\boldsymbol{n} = \{1, 2, 3, 4, 5\}$, $\mathcal{I} = \{(1, 2), (2, 3), (3, 4), (4, 2), (4, 5)\}$,

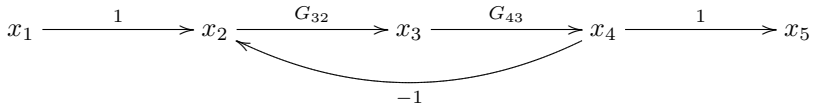


Figure 6.6 The signal flow graph for a negative feedback loop

$\mathcal{S} = \{x_1, x_2, x_3, x_4, x_5\}$, and $\mathcal{G} = \{G_{21} = 1, G_{32}, G_{43}, G_{24} = -1, G_{54} = 1\}$. Clearly, x_1 is an input and x_5 is an output. From the graph we read the relationships $x_2 = x_1 - x_4$, $x_3 = G_{32}x_2$, $x_4 = G_{43}x_3$, and $x_5 = x_4$. This gives, in the usual manner,

$$\begin{aligned} & x_5 = x_4 = G_{43}x_3 = G_{43}G_{32}x_2 = G_{43}G_{32}(x_1 - x_4) \\ \Longrightarrow \quad & x_4 = x_5 = \frac{G_{43}G_{32}}{1 + G_{43}G_{32}}x_1. \end{aligned}$$

•

4. Suppose we wish to extract more information from the negative feedback loop of the previous example. In Figure 6.7 we depict a situation where we have added an input

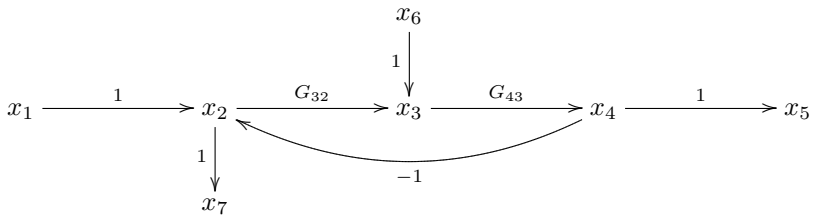


Figure 6.7 Signal flow graph for negative feedback loop with extra structure

signal x_6 to the graph, and tapped an output signal $x_7 = x_2$. Thinking about the control situation, one might wish to think of x_6 as a disturbance to the system, and of x_7 as being the error signal (this will be seen in a better context in Sections 6.3, 8.3, and 8.4). In doing so we have added the input x_6 to the existing input x_1 and the output x_7 to the existing output x_5. Let us see how the two outputs get expressed as functions of the two inputs. We have the relations $x_2 = x_1 - x_4$, $x_3 = G_{32}x_2 + x_6$, $x_4 = G_{43}x_3$, $x_5 = x_4$, and $x_7 = x_2$. We combine these to get

$$\begin{aligned} & x_5 = x_4 = G_{43}x_3 = G_{32}G_{43}x_2 + G_{43}x_6 = G_{32}G_{43}x_1 - G_{32}G_{43}x_4 + G_{43}x_6 \\ \Longrightarrow \quad & x_5 = x_4 = \frac{G_{32}G_{43}}{1 + G_{32}G_{43}}x_1 + \frac{G_{43}}{1 + G_{32}G_{43}}x_6 \\ & x_7 = x_2 = x_1 - x_4 = x_1 - G_{43}x_3 = x_1 - G_{43}G_{32}x_2 - G_{43}x_6 \\ \Longrightarrow \quad & x_7 = x_2 = \frac{1}{1 + G_{32}G_{43}}x_1 - \frac{G_{43}}{1 + G_{32}G_{43}}x_6. \end{aligned}$$

We see that we essentially obtain a system of two linear equations that expresses the input/output relations for the graph. •

One can consider *any* node x_j to effectively be an output by adding to the signal flow graph a node x_{n+1} and a branch $G_{n+1,j}$ with gain equal to 1. One can also add an input to any node x_j by adding a node x_{n+1} and a branch $G_{j,n+1}$ whose gain is 1.

6.1.2 Signal flow graphs and systems of equations

A signal flow graph is also a representation of a set of linear equations whose coefficients belong to any field. That is to say, we consider a set of linear equations where the coefficients are anything that can be added, multiplied, and divided. The particular field that is of interest to us is $\mathbb{R}(s)$. What's more, we consider a very particular type of linear equation; one of the form

$$\begin{aligned}(1-G_{11})x_1 - G_{12}x_2 - \cdots - G_{1n}x_n &= u_1\\ -G_{21}x_1 + (1-G_{22})x_2 - \cdots - G_{2n}x_n &= u_2\\ &\vdots\\ -G_{n1}x_1 - G_{n2}x_2 - \cdots + (1-G_{nn})x_n &= u_n,\end{aligned} \tag{6.1}$$

where $G_{ij} \in \mathbb{R}(s)$, $i,j = 1,\dots,n$. The reason for using this type of equation will become clear shortly. However, we note that corresponding to the equations (6.1) is a natural signal flow graph. The following construction indicates how this is determined.

6.3 From linear equation to signal flow graph Given a set of linear equations of the form (6.1), perform the following steps:

(i) place a node for each variable $x_1,\dots,x_n$;

(ii) place a node for each input $u_1,\dots,u_n$;

(iii) for each nonzero G_{ij}, $i,j = 1,\dots,n$, draw a branch originating from node j and terminating in node i, having gain G_{ij};

(iv) for $i = 1,\dots,n$, draw a branch of gain 1 from u_i to x_i.

The result is a signal flow graph with nodes $\{x_1,\dots,x_n, x_{n+1}=u_1,\dots,x_{2n}=u_n\}$ and gains $\mathcal{G} = \{G_{ij} \mid i,j = 1,\dots,n\} \cup \{G_{1,n+1}=1,\dots,G_{n,2n}=1\}$. •

For example, in Figure 6.8 we show how this is done when $n = 2$. One can readily verify

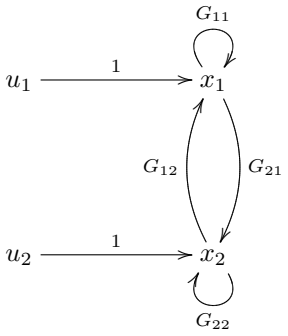


Figure 6.8 A general 2 node signal flow graph

that a "balance" at each node will yield the equations (6.1).

This establishes a graph for each set of equations of the form (6.1). It is also true that one can go from a graph to a set of equations. To state how to do this, we provide the following recipe.

6.4 From signal flow graph to linear equation Given a signal flow graph $(\mathcal{S}, \mathcal{G})$, to any source x_i that is not an input, add a node $x_{j_i} = u_i$ to $\mathcal{S}$ and a branch G_{ij_i} to $\mathcal{G}$. After doing this for each such source, we arrive at a new signal flow graph $(\mathcal{S}', \mathcal{G}')$. For each node x_i in $(\mathcal{S}', \mathcal{G}')$ that is not an input perform the following steps:

(i) let $x_{j_1}, \dots, x_{j_k}$ be the collection of nodes for which there are branches connecting them with x_i;

(ii) form the product of $x_{j_1}, \dots, x_{j_k}$ with the respective branch gains;

(iii) set the sum of these products equal to x_i.

The result is an equation defining the node balance at node x_i. Some of the inputs may be zero. •

Thus we establish a $1-1$ correspondence between signal flow graphs and linear equations of the form (6.1) (with some inputs and gains possibly zero). Let us denote by $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ the matrix of rational functions that serves as the coefficient matrix in (6.1). Thus

$$\boldsymbol{G}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 - G_{11} & -G_{12} & \cdots & -G_{1n} \\ -G_{21} & 1 - G_{22} & \cdots & -G_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -G_{n1} & -G_{n2} & \cdots & 1 - G_{nn} \end{bmatrix}. \tag{6.2}$$

We call $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ the ***structure matrix*** for $(\mathcal{S}, \mathcal{G})$. Note that $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ is a matrix whose components are rational functions. We denote the collection of $n \times n$ matrices with rational function components by $\mathbb{R}(s)^{n\times n}$. Again we note that for a given signal flow graph, of course, many of the terms in this matrix might be zero. The objective of making the connection between signal flow graphs and linear equations is that the matrix formulation puts at our disposal all of the tools from linear algebra. Indeed, one could simply use the form (6.1) to study signal flow graphs. However, this would sweep under the carpet the special structure of the equations that results from their being derived as node equations of a signal flow graph. One of the main objectives of this section is to deal with this aspect of the signal flow graph. But for now, let us look at our signal flow graphs of Example 6.2 as systems of equations.

6.5 Examples **(Example 6.2 cont'd)** We shall simply write down the matrix $\boldsymbol{G}$ that appears in (6.1). In each case, we shall consider, as prescribed by the above procedure, the system with an input attached to each source that is not itself an input. Thus we ensure that the matrix $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ represents all states of the system.

1. For the series interconnection we work with the signal flow graph of Figure 6.9, and we

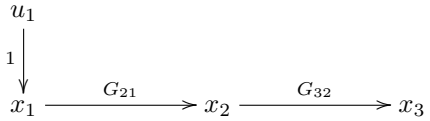


Figure 6.9 Series signal flow graph with input added

determine

$$\boldsymbol{G}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 \\ -G_{21} & 1 & 0 \\ 0 & -G_{32} & 1 \end{bmatrix}.$$

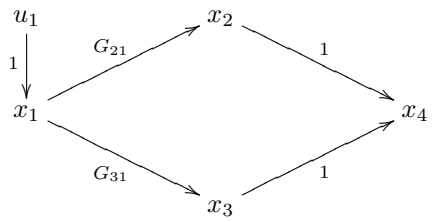


Figure 6.10 Parallel signal flow graph with input added

2. For the parallel interconnection we work with the signal flow graph of Figure 6.10, and we determine

$$\boldsymbol{G}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -G_{21} & 1 & 0 & 0 \\ -G_{31} & 0 & 1 & 0 \\ 0 & -G_{42} & -G_{43} & 1 \end{bmatrix}.$$

3. Finally, for the negative feedback interconnection we work with the signal flow graph of Figure 6.11, and we determine

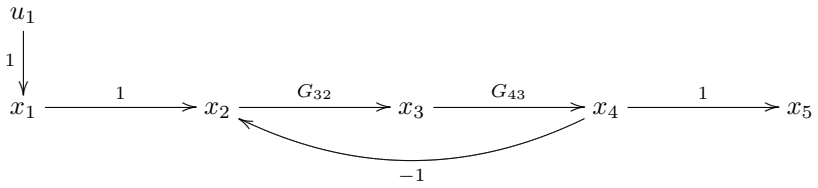


Figure 6.11 Negative feedback signal flow graph with input added

$$\boldsymbol{G}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -G_{21} & 1 & 0 & -G_{24} & 0 \\ 0 & -G_{32} & 1 & 0 & 0 \\ 0 & 0 & -G_{43} & 1 & 0 \\ 0 & 0 & 0 & -G_{54} & 1 \end{bmatrix}.$$

Note that we may take some of the gains to be 1 when they appear that way in Example 6.2. •

6.1.3 Subgraphs, paths, and loops

In the constructions of the next section, we need to have a good understanding of paths through a signal flow graph. To do this, we make some definitions.

6.6 Definition Let $(\mathcal{S}, \mathcal{G})$ be a signal flow graph with interconnections $\mathcal{I} \subset \boldsymbol{n} \times \boldsymbol{n}$.

(i) A ***subgraph*** of $(\mathcal{S}, \mathcal{G})$ is a pair $(\mathcal{S}', \mathcal{G}')$ where $\mathcal{S}' \subset \mathcal{S}$ and where $\mathcal{G}' \subset \mathcal{G}$ satisfies $G_{ij} \in \mathcal{G}'$ implies that $x_i, x_j \in \mathcal{S}'$. If $\mathcal{S}' = \{x_{i_1}, \dots, x_{i_{n'}}\}$, then we say $(\mathcal{S}', \mathcal{G}')$ has ***interconnections*** $\{i_1, \dots, i_{n'}\}$. We also denote by $\mathcal{I}' \subset \mathcal{I}$ the subset defined by $\mathcal{G}'$. Thus $(i', j') \in \mathcal{I}'$ if and only if $G_{i'j'} \in \mathcal{G}'$.

(ii) A subgraph $(\mathcal{S}', \mathcal{G}')$ is ***connected*** if for any pair $x_i, x_j \in \mathcal{S}'$ there exists nodes $x_{i_0} = x_i, x_{i_1}, \dots, x_{i_k} = x_j \in \mathcal{S}'$ and gains $G_1, \dots, G_k \in \mathcal{G}'$ so that $G_\ell \in \{G_{i_\ell, i_{\ell+1}}, G_{i_{\ell+1}, i_\ell}\}$, $\ell \in \{0, \dots, k-1\}$.

(iii) A ***path*** in $(\mathcal{S}, \mathcal{G})$ is a sequence $\{x_{i_1}, \dots, x_{i_k}\}$ of nodes with the property that $(i_j, i_{j+1}) \in \mathcal{I}$ for each $j \in \{1, \dots, k-1\}$. That is to say, there is a branch connecting the jth element in the sequence with the $(j+1)$st element in the sequence.

(iv) A path $\{x_{i_1}, \dots, x_{i_k}\}$ is ***simple*** if the set $\{(i_{j+1}, i_j) \mid j \in \{1, \dots, k-1\}\}$ is distinct.

(v) A path $\{x_{i_1}, \dots, x_{i_k}\}$ is ***directed*** if the vertices are distinct. We denote the set of all directed paths in $(\mathcal{S}, \mathcal{G})$ by $\text{Path}(\mathcal{S}, \mathcal{G})$.

(vi) A ***forward path*** is a directed path from an input node of $(\mathcal{S}, \mathcal{G})$ to a node of $(\mathcal{S}, \mathcal{G})$. The set of forward paths from an input x_i to a node x_j is denoted $\text{Path}_{ji}(\mathcal{S}, \mathcal{G})$.

(vii) The number of branches in a directed path is the ***length*** of the directed path.

(viii) A ***loop*** is a simple path $\{x_{i_1}, \dots, x_{i_k}\}$ with the property that $x_{i_1} = x_{i_k}$. The set of loops in $(\mathcal{S}, \mathcal{G})$ we denote by $\text{Loop}(\mathcal{S}, \mathcal{G})$.

(ix) The product of the gains in a directed path is the ***gain*** of the directed path.

(x) The ***loop gain*** of a loop L is the product of the gains comprising the loop, and is denoted G_L.

(xi) A finite collection of loops is ***nontouching*** if they have no nodes or branch gains in common. •

Perhaps a few words of clarification about the less obvious parts of the definition are in order.

1. The notion of a connected subgraph has a simplicity that belies its formal definition. It merely means that it is possible to go from any node to any other node, provided one ignores the orientation of the branches.
2. Note that the nodes in a loop must be distinct. That a loop cannot follow a path that goes through any node more than once.
3. It is easy to be misled into thinking that any directed path is a forward path. This is not necessarily true since for a forward path originates at an input *for* $(\mathcal{S}, \mathcal{G})$.

Again, this looks pretty formidable, but is in fact quite simple. We can illustrate this easily by looking at the examples we have given of signal flow graphs.

6.7 Examples

1. For the very simple signal flow graph of Figure 6.1 there is but one path of length 1, and it is $\{x_1, x_2\}$. There are no loops.
2. For the series signal flow graph of Figure 6.4 we have two paths of length 1,
$$\{x_1, x_2\}, \quad \{x_2, x_3\},$$
and one path of length 2,
$$\{x_1, x_2, x_3\}.$$
Again, there are no loops.
3. The parallel signal flow graph of Figure 6.5, forgetting for the moment that some of the gains are predetermined to be 1, has the paths
$$\{x_1, x_2\}, \quad \{x_2, x_4\}, \quad \{x_1, x_3\}, \quad \{x_3, x_4\},$$
of length 1, and paths
$$\{x_1, x_2, x_4\}, \quad \{x_1, x_3, x_4\}$$
of length 2. There are no loops in this graph either.

4. The negative feedback graph of Figure 6.7 has an infinite number of paths. Let us list the basic ones.
   (a) length 1:
   $$\{x_1, x_2\}, \quad \{x_2, x_3\}, \quad \{x_3, x_4\}, \quad \{x_4, x_5\}, \quad \{x_3, x_6\}, \quad \{x_2, x_7\}.$$
   (b) length 2:
   $$\{x_1, x_2, x_3\}, \quad \{x_2, x_3, x_4\}, \quad \{x_3, x_4, x_5\}, \quad \{x_3, x_4, x_6\}, \quad \{x_2, x_4, x_7\}.$$
   (c) length 3:
   $$\{x_1, x_2, x_3, x_4\}, \quad \{x_2, x_3, x_5, x_5\}, \quad \{x_3, x_4, x_5, x_6\}.$$
   (d) length 4:
   $$\{x_1, x_2, x_3, x_4, x_5\}.$$
   Some of these paths may be concatenated to get paths of any length desired. The reason for this is that there is a loop given by
   $$\{x_2, x_3, x_4\}.$$ •

6.1.4 Cofactors and the determinants

Next we wish to define, and state some properties of, some quantities associated with the matrix $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$. These quantities we will put to use in the next section in proving an important theorem in the subject of signal flow graphs: Mason's Rule. This is a rule for determining the transfer function between any input and any output of a signal flow graph. But this is getting ahead of ourselves. We have a lot of work to do in the interim.

Let us proceed with the development. For $k \geq 1$ we denote

$$\text{Loop}_k(\mathcal{S}, \mathcal{G}) = \{(L_{j_1}, \dots, L_{j_k}) \in (\text{Loop}(\mathcal{S}, \mathcal{G}))^k \mid L_1, \dots, L_k \text{ are nontouching}\}.$$

That is, $\text{Loop}_k(\mathcal{S}, \mathcal{G})$ consists of those k-tuples of loops, none of which touch the others. Note that $\text{Loop}_1(\mathcal{S}, \mathcal{G}) = \text{Loop}(\mathcal{S}, \mathcal{G})$ and that for k sufficiently large (and not very large at all in most examples we shall see), $\text{Loop}_k(\mathcal{S}, \mathcal{G}) = \emptyset$. The ***determinant*** of a signal flow graph $(\mathcal{S}, \mathcal{G})$ is defined to be

$$\Delta_{\mathcal{S},\mathcal{G}} = 1 + \sum_{k \geq 1} \left(\frac{(-1)^k}{k!} \sum_{\substack{(L_1,\dots,L_k) \in \\ \text{Loop}_k(\mathcal{G},\mathcal{S})}} (G_{L_1} \cdots G_{L_k}) \right). \tag{6.3}$$

It turns out that $\Delta_{\mathcal{S},\mathcal{G}}$ is exactly the determinant of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$.

6.8 Proposition $\Delta_{\mathcal{S},\mathcal{G}} = \det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}$.

Proof The proof is involved, but not difficult. We accomplish it with a series of lemmas. First we note that the definition of the determinant gives an expression of the form

$$\det \boldsymbol{G}_{\mathcal{S},\mathcal{G}} = 1 + \sum_{\alpha} G_\alpha, \tag{6.4}$$

where G_α is a product of branch gains. We denote by $(\mathcal{S}_\alpha, \mathcal{G}_\alpha)$ the subgraph of $(\mathcal{S}, \mathcal{G})$ comprised of the nodes and branches that are involved in a typical term G_α.

We explore some properties of the subgraph $(\mathcal{S}_G, \mathcal{G}_G)$.

1 Lemma *For each node $x_i \in \mathcal{S}_\alpha$, there is at most one branch in $\mathcal{G}_\alpha$ terminating at x_i, and at most one branch in $\mathcal{G}_\alpha$ originating from x_i.*

Proof Let $\mathcal{G} = \{G_{\alpha 1}, \dots, G_{\alpha k}\}$ be the gains that form G_α. By the definition of the determinant, for each row of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ there is at most one $j \in \{1, \dots, k\}$ so that $G_{\alpha j}$ is an element of that row. A similar statement holds for each column of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$. However, from these statements exactly follows the lemma. ▼

2 Lemma *G_α is a product of loop gains of nontouching loops.*

Proof Suppose that $(\mathcal{S}_\alpha, \mathcal{G}_\alpha)$ consists of a collection $(\mathcal{S}_{\alpha 1}, \mathcal{G}_{\alpha 1}), \dots, (\mathcal{S}_{\alpha \ell}, \mathcal{G}_{\alpha \ell})$ of connected subgraphs. We shall show that each of these connected subgraphs is a loop. It then follows from Lemma 1 that the loops will be nontouching. We proceed by contradiction, supposing that one of the connected components, say $(\mathcal{S}_{\alpha 1}, \mathcal{G}_{\alpha 1})$, is not a loop. From Lemma 1 it follows that $(\mathcal{S}_{\alpha 1}, \mathcal{G}_{\alpha 1})$ is a directed path between two nodes; suppose that these nodes are x_i and x_j. Since no branches terminate at x_i, there are no elements from row i of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ in G_α. Since no branches originate from x_j, there are no elements from column j of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ in G_α. It therefore follows that the 1's in position (i,i) and (j,j) appear in the expression for G_α. However, since $\mathcal{S}_{\alpha 1}, \mathcal{G}_{\alpha 1})$ is a directed path between x_i and x_j it follows that there are some $k_i, k_j \in \boldsymbol{n}$ so that $G_{k_i i}$ and G_{j,k_j} appear in G_α. This is in contradiction to Lemma 1. ▼

3 Lemma *If G_α is the product of the loop gains from k loops, then it has sign $(-1)^k$ in the expression for* $\det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}$.

Proof First note that the determinant of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ is unaffected by the numbering of the nodes. Indeed, if one interchanges i and j in a given numbering scheme, then in $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$, both the ith and jth columns and the ith and jth rows are swapped. Since each of these operations gives a change of sign in the determinant, the determinant itself is the same.

Suppose that $(\mathcal{S}_\alpha, \mathcal{G}_\alpha)$ is comprised of k nontouching loops of lengths $\ell_1, \dots, \ell_k$. Denote the nodes in these loops by

$$\{x_1, \dots, x_{\ell_1}, x_1\},\ \{x_{\ell_1+1}, \dots, x_{\ell_2}, x_{\ell+1}\},\ \dots,\ \{x_{\ell_{k-1}+1}, \dots, x_{\ell_k}, x_{\ell_{k-1}+1}\}.$$

According to the definition (A.1) of the determinant, to determine the sign of the contribution of G_α, we should determine the sign of the permutation

$$\sigma = \begin{pmatrix} 1 & \cdots & \ell_1 - 1 & \ell_1 & \ell_1 + 1 & \cdots & \ell_2 - 1 & \ell_2 & \cdots & \ell_{k-1} + 1 & \cdots & \ell_k - 1 & \ell_k \\ 2 & \cdots & \ell_1 & 1 & \ell_1 + 2 & \cdots & \ell_2 & \ell_1 + 1 & \cdots & \ell_{k-1} & \cdots & \ell_k & \ell_{k-1} + 1 \end{pmatrix}.$$

By $\ell_1 - 1$ transpositions we may shift the 1 in the ℓ_1st position to the 1st position. By $\ell_2 - 1$ transpositions, the $\ell_1 + 1$ in the $(\ell_1 + \ell_2)$th position can be shifted to the $(\ell_1 + 1)$st position. Proceeding in this manner, we see that the sign of the permutation is given by

$$\operatorname{sgn}(\sigma) = \prod_{j=1}^{k} (-1)^{\ell_j - 1} \tag{6.5}$$

Since each of the gains occurs with sign -1 in $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$, they will contribute the sign

$$\operatorname{sgn}(\sigma) \prod_{j=1}^{k} (-1)^{\ell_j}$$

to G_α. However, by (6.5), we have

$$\operatorname{sgn}(\sigma)\prod_{j=1}^{k}(-1)^{\ell_j} = (-1)^k,$$

and so the lemma follows. ▼

The three previous lemmas show that the terms G_α in the expression (6.4) for $\det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ have the form of the terms in $\Delta_{\mathcal{S},\mathcal{G}}$. It remains to show that every term in $\Delta_{\mathcal{G},\mathcal{G}}$ appears in $\det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}$. Thus suppose that $(\mathcal{S}_{\alpha 1},\mathcal{G}_{\alpha 1}),\dots,(\mathcal{S}_{\alpha\ell},\mathcal{G}_{\alpha\ell})$ is a collection of nontouching loops. Since these loops are nontouching, it follows that in any given row or column, there can reside at most one gain from the set $\mathcal{G}_{\alpha 1}\cup\dots\mathcal{G}_{\alpha\ell}$. Therefore, an expansion of the determinant will contain the product of the gains from the set $\mathcal{G}_{\alpha 1}\cup\dots\mathcal{G}_{\alpha\ell}$. Furthermore, they must have the same sign in $\det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ as in $\Delta_{\mathcal{S},\mathcal{G}}$ by Lemma 3. This concludes the proof. ■

Let us see how to apply this in two examples, one of which we have seen, and one of which is complicated enough to make use of the new terminology we have introduced.

6.9 Examples

1. Let us first look at the signal flow graph depicted in Figure 6.7. For clarity it helps to ignore the fact that $G_{24} = -1$. This graph has a single loop

$$L = \{x_2, x_3, x_4, x_2\}.$$

This means that $\operatorname{Loop}_1(\mathcal{S},\mathcal{G}) = \{L\}$ and $\operatorname{Loop}_k(\mathcal{S},\mathcal{G}) = \emptyset$ for $k \geq 2$. The gain of the loop is simply the product $G_{32}G_{43}G_{24}$. Therefore, in this case, the determinant as per (6.3) is

$$\Delta_{\mathcal{S},\mathcal{G}} = 1 - G_{32}G_{43}G_{24}.$$

If we now remember that $G_{24} = -1$ we get the term $1 + G_{32}G_{43}$ which appears in the denominators in Example 6.2–4. This, of course, is no coincidence.

2. Next we consider a new signal flow graph, namely the one depicted in Figure 6.12. This

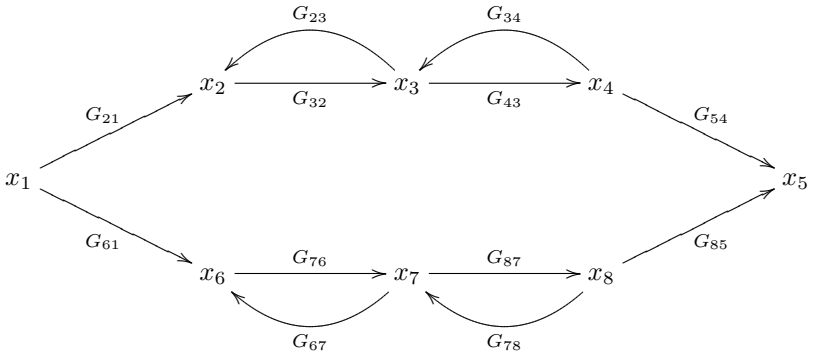


Figure 6.12 A signal flow graph with multiple loops

graph has four loops:

$$L_1 = \{x_2, x_3, x_2\},\quad L_2 = \{x_3, x_4, x_3\},\quad L_3 = \{x_6, x_7, x_6\},\quad L_4 = \{x_7, x_8, x_7\}.$$

We have $\operatorname{Loop}_1(\mathcal{S},\mathcal{G}) = \{L_1, L_2, L_3, L_4\}$ and

$$\operatorname{Loop}_2(\mathcal{S},\mathcal{G}) = \{(L_1, L_3), (L_1, L_4), (L_2, L_3), (L_2, L_4), (L_3, L_1), (L_4, L_1), (L_3, L_1), (L_4, L_2)\}.$$

An application of (6.3) gives

$$\Delta_{\mathcal{S},\mathcal{G}} = 1 - (G_{L_1} + G_{L_2} + G_{L_3} + G_{L_4}) + (G_{L_1}G_{L_3} + G_{L_1}G_{L_4} + G_{L_2}G_{L_3} + G_{L_2}G_{L_4}).$$

One may, of course, substitute into this expression the various gains given in terms of the branch gains. •

Now we turn to the cofactor of a path. For $P \in \mathrm{Path}(\mathcal{S},\mathcal{G})$ let $\mathcal{G}_P$ denote the branches of $\mathcal{G}$ with those comprising P removed, and those branches having a node in common with P removed. We note that $(\mathcal{S},\mathcal{G}_P)$ is itself a signal flow graph. If $(\mathcal{S},\mathcal{G})$ is connected, then $(\mathcal{S},\mathcal{G}_P)$ may not be. For $P \in \mathrm{Path}(\mathcal{S},\mathcal{G})$ the ***cofactor*** of $\mathcal{P}$ is defined by $\mathrm{Cof}_P(\mathcal{S},\mathcal{G}) = \Delta_{\mathcal{S},\mathcal{G}_P}$.

Let us illustrate this for the examples whose determinants we have computed.

6.10 Examples

1. For the signal flow graph depicted in Figure 6.7 let us consider various paths. Again we do not pay attention to specific values assign to branch gains.
   (a) $P_1 = \{x_1, x_2, x_3, x_4, x_5\}$: If we remove this path, the graph has no loops and so $\mathrm{Cof}_{P_1}(\mathcal{S},\mathcal{G}) = 1$.
   (b) $P_2 = \{x_1, x_2, x_7\}$: Removing this leaves intact the existing loop, and so the determinant of the graph remains unchanged. Therefore we have $\mathrm{Cof}_{P_2}(\mathcal{S},\mathcal{G}) = \Delta_{\mathcal{S},\mathcal{G}} = 1 - G_{32}G_{43}G_{24}$.
   (c) $P_3 = \{x_3, x_4, x_5, x_6\}$: Removal of this path leaves a signal flow graph with no loops so we must have $\mathrm{Cof}_{P_3}(\mathcal{S},\mathcal{G}) = 1$.
   (d) $P_4 = \{x_2, x_3, x_4, x_6, x_7\}$: Again, if P_4 is removed, we are left with no loops and this then gives $\mathrm{Cof}_{P_4}(\mathcal{S},\mathcal{G}) = 1$.
2. Next we look at the signal flow graph of Figure 6.12. We consider two paths.
   (a) $P_1 = \{x_1, x_2, x_3, x_4, x_5\}$: Removing this loop leaves two loops remaining: L_3 and L_4. The determinant of the resulting graph is, by (6.3),

$$\mathrm{Cof}_{P_1}(\mathcal{S},\mathcal{G}) = 1 - (G_{L_3} + G_{L_4}).$$

   (b) $P_2 = \{x_1, x_6, x_7, x_8, x_5\}$: This situation is rather like that for the path P_1 and we determine that

$$\mathrm{Cof}_{P_2}(\mathcal{S},\mathcal{G}) = 1 - (G_{L_1} + G_{L_2}).$$

Into these expressions for the cofactors, one may substitute the branch gains. •

6.1.5 Mason's Rule

With the notion of cofactor and determinant clearly explicated, we can state Mason's Rule for finding the transfer function between an input to a signal flow graph and any node in the graph. In this section we suppose that we are working with a signal flow graph $(\mathcal{S},\mathcal{G})$ with interconnections $\mathcal{I} \subset \boldsymbol{n} \times \boldsymbol{n}$. In order to simplify matters and make extra hypotheses for everything we do, let us make a blanket assumption in this section.

The following result gives an easy expression for the transfer function between the input u_i at x_i and any other node. The following result can be found in the papers of Mason (1953, 1953). It is not actually given a rigorous proof there, but one is given by Zadeh and Desoer [1979]. We follow the latter proof.

6.11 Theorem **(Mason's Rule)** *Let $(\mathcal{S},\mathcal{G})$ be a signal flow graph. For $i,j\in\{1,\dots,n\}$ we have*

$$x_j=\sum_{P\in\mathrm{Path}_{ji}(\mathcal{S},\mathcal{G})}\frac{G_P\mathrm{Cof}_P(\mathcal{S},\mathcal{G})}{\Delta_{\mathcal{S},\mathcal{G}}}u_i.$$

Proof Without loss of generality, suppose that $i=1$ and that node 1 is an input. Denote $\boldsymbol{x}_1=(x_1,0,\dots,0)$ and let $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)$ be the matrix $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ with the jth column replaced with $\boldsymbol{x}_1$. Cramer's Rule then says that

$$x_j=\frac{\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)}{\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}}.$$

From Proposition 6.8 the theorem will follow if we can show that

$$\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)=\sum_{P\in\mathrm{Path}_{j1}(\mathcal{S},\mathcal{G})}G_P\mathrm{Cof}_P(\mathcal{S},\mathcal{G}).$$

Let $(\mathcal{S}_j,\mathcal{G}_j)$ be the subgraph of $(\mathcal{S},\mathcal{G})$ corresponding to the matrix $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)$. One can readily ascertain that one may arrive at $(\mathcal{S}_j,\mathcal{G}_j)$ from $(\mathcal{S},\mathcal{G})$ by performing the following operations:

1. remove all branches originating from x_j;
2. add a branch with gain x_1 originating from x_j and terminating at x_1.

Since the only nonzero term in the jth column is the term x_1 appearing in the first row, an expansion of the determinant about the jth column shows that $\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)$ has the form

$$\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)=x_1\prod_\alpha G_\alpha.$$

By Proposition 6.8 we know that $\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)$ is comprised of a sum of products of loop gains for nontouching loops. Thus in each of the products must appear a loop gain of the form

$$f_1G_{k_21}G_{k_3k_2}\cdots G_{jk_{\ell-1}}.$$

It follows that $\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)$ is a sum of products of terms falling into three types:

1. x_1;
2. the gain of a forward path P from x_1 to x_j;
3. a product of loop gains for loops that share no nodes or branches with P.

Let us fix such a term corresponding to a given forward path P_α. Now let us renumber the nodes so that we have

$$\boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1,j)=\begin{bmatrix}1&0&\cdots&0&x_1&\\-G_{k_21}&1&\cdots&0&0&\\\vdots&\vdots&\ddots&\vdots&\vdots&\boldsymbol{0}_{\ell,n-\ell}\\0&0&\cdots&1&0&\\0&0&\cdots&-G_{jk_{\ell-1}}&0&\\&&\boldsymbol{0}_{n-\ell,\ell}&&&\tilde{\boldsymbol{G}}\end{bmatrix}\tag{6.6}$$

for some matrix $\tilde{\boldsymbol{G}} \in \mathbb{R}^{(n-\ell)\times(n-\ell)}$. Therefore we have

$$\begin{aligned}\det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1, j) &= (-1)^{\ell-1} x_1(-G_{k_2 1})(-G_{k_3 k_2}) \cdots (-G_{j k_{\ell-1}}) \det \tilde{\boldsymbol{G}} \\ &= x_1 G_{k_2 1} G_{k_3 k_2} \cdots G_{j k_{\ell-1}} \det \tilde{\boldsymbol{G}}.\end{aligned}$$

Now note the form of the matrix in (6.6) shows that the signal flow graph corresponding to $\tilde{\boldsymbol{G}}$ is obtained from $(\mathcal{S}, \mathcal{G})$ by removing all branches associated with the forward path P_α. From this it follows that all products in $\det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1, j)$ are of the form $G_P \mathrm{Cof}_P(\mathcal{S}, \mathcal{G})$ for some forward path P.

It remains to show that any such expression will appear as a product in the expression for $\det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}(\boldsymbol{x}_1, j)$. This may be shown as follows. Let P be a forward path comprised of gains $G_{k_2 1}, G_{k_3 k_2}, \ldots, G_{j,k_{\ell-1}}$. The structure of $(\mathcal{S}_j, \mathcal{G}_j)$ implies that by adding the gain x_1, we have the gain for a loop in $(\mathcal{S}_j, \mathcal{G}_j)$. Now, as we saw in the proof of Proposition 6.8, this term, multiplied by the loop gains of nontouching loops, is ensured to appear in the determinant of $(\mathcal{S}_j, \mathcal{G}_j)$. ■

If a signal flow graph has multiple inputs $u_1, \ldots, u_k$, then one can apply Mason's rule for each input, and the resulting expression for a non-input node x_j is then of the form

$$x_j = \sum_{i=1}^{k} T_{ji} u_i,$$

where

$$T_{ji} = \sum_{P \in \mathrm{Path}_{ji}(\mathcal{S},\mathcal{G})} \frac{G_P \mathrm{Cof}_P(\mathcal{S}, \mathcal{G})}{\Delta_{\mathcal{S},\mathcal{G}}}$$

is the ***graph transmittance*** from the input u_i to the node x_j. Note that it is possible that for a given i and j, $\mathrm{Path}_{ji}(\mathcal{S}, \mathcal{G})$ will be empty. In this case, we take the graph transmittance to be zero.

The following result gives a useful interpretation of the set of all graph transmittances.

6.12 Corollary *The matrix $\boldsymbol{T}_{\mathcal{S},\mathcal{G}} \in \mathbb{R}(s)^{n\times n}$ whose (i,j)th component is the graph transmittance T_{ij} is the inverse of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$.*

Proof As we saw in the proof of Theorem 6.11, the numerator term in the expression for T_{ji} is the determinant of the matrix obtained by replacing the jth column of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ by the ith standard basis vector $\boldsymbol{e}_i$. By Cramer's Rule we infer that T_{ji} is the jth component of the solution vector $\boldsymbol{t}_i$ for the linear equation $\boldsymbol{G}_{\mathcal{S},\mathcal{G}} \boldsymbol{t}_i = \boldsymbol{e}_i$. This means that if we define

$$\boldsymbol{T}_{\mathcal{S},\mathcal{G}} = \left[\begin{array}{c|c|c|c} \boldsymbol{t}_1 & \cdots & \boldsymbol{t}_n \end{array} \right],$$

then we have

$$\boldsymbol{G}_{\mathcal{S},\mathcal{G}} \boldsymbol{T}_{\mathcal{S},\mathcal{G}} = \left[\begin{array}{c|c|c|c} \boldsymbol{e}_1 & \cdots & \boldsymbol{e}_n \end{array} \right] = \boldsymbol{I}_n.$$

From the definition of inverse, the result follows. ■

Let us show how the above developments all come together to allow an easy determination of the various transfer functions for the two examples we are considering.

6.13 Examples **(Example 6.9 cont'd)** In each example, we will number paths as we did in Example 6.9.

1. For the signal flow graph of Figure 6.7 there are two sinks, x_1 and x_6, and two sources, x_5 and x_7. To each of the sources, let us attach an input so that we are properly in the setup of Theorem 6.11. Recalling our labelling of paths from Example 6.9, we have

$$\mathrm{Path}_{51}(\mathcal{S},\mathcal{G}) = \{P_1\}, \quad \mathrm{Path}_{71}(\mathcal{S},\mathcal{G}) = \{P_2\},$$
$$\mathrm{Path}_{56}(\mathcal{S},\mathcal{G}) = \{P_3\}, \quad \mathrm{Path}_{76}(\mathcal{S},\mathcal{G}) = \{P_4\}.$$

   These are, of course, the paths whose cofactors we computed in Example 6.9. Now we can compute, using Mason's Rule, the coefficients in expressions of our two sinks x_5 and x_7 involving the two sources x_1 and x_6. We have

$$T_{51} = \frac{G_{21}G_{32}G_{43}G_{54}}{1 - G_{32}G_{43}G_{24}}$$
$$T_{71} = \frac{G_{21}G_{72}}{1 - G_{32}G_{43}G_{24}}$$
$$T_{56} = \frac{G_{36}G_{43}G_{54}}{1 - G_{32}G_{43}G_{24}}$$
$$T_{76} = \frac{G_{36}G_{43}G_{24}G_{72}}{1 - G_{32}G_{43}G_{24}}$$

   and so we thus obtain the explicit expressions

$$x_5 = T_{51}x_1 + T_{56}x_6, \quad x_7 = T_{71}x_1 + T_{76}x_6.$$

   Let's see how this checks out when $G_{21} = G_{54} = G_{72} = G_{36} = -G_{24} = 1$. In this case we obtain

$$x_5 = \frac{G_{32}G_{43}}{1 + G_{32}G_{43}}x_1 + \frac{G_{43}}{1 + G_{32}G_{43}}x_6, \quad x_7 = \frac{1}{1 + G_{32}G_{43}}x_1 - \frac{G_{43}}{1 + G_{32}G_{43}}x_6$$

   as we did when we performed the calculations "by hand" back in Example 6.2.
2. We shall compute the transfer function from x_1 to x_5. We have $\mathrm{Path}_{51}(\mathcal{S},\mathcal{G}) = \{P_1, P_2\}$. By Mason's Rule, and using the determinant and cofactors we have already computed, we have

$$x_5 = \frac{G_{21}G_{32}G_{43}G_{54}\big(1 - (G_{L_3} + G_{L_4})\big) + G_{61}G_{76}G_{87}G_{58}\big(1 - (G_{L_1} + G_{L_2})\big)}{\Delta_{\mathcal{S},\mathcal{G}}}x_1.$$

   As always, we may substitute into this the values for the branch gains to get an horrific formula for the transfer function. But just imagine trying to do this "by hand"! •

We have provided in this section a systematic way of deriving the transfer function between various inputs and outputs in a signal flow graph. What's more, we have identified an important piece of structure in any such transfer function: the determinant of the graph. We shall put this to use when studying stability of interconnected systems in Section 6.2.3.

6.1.6 Sensitivity, return difference, and loop transmittance

Up to this point, the discussion has been centred around the various transfer functions appearing in a signal flow graph. Let us now look at other interesting objects, sensitivity, return difference, and loop transmittance. These will turn out to be interesting in the special context of single-loop systems in Section 6.3. The ideas we discuss in this section are presented also in the books of Truxal [1955] and Horowitz [1963].

First we consider the notion of sensitivity. Let us first give a precise definition.

6.14 Definition Let $(\mathcal{S},\mathcal{G})$ be a signal flow graph with interconnections $\mathcal{I}\subset \boldsymbol{n}\times\boldsymbol{n}$, let $i,j\in\boldsymbol{n}$, and let $(\ell,k)\in\mathcal{I}$. Let T_{ji} be the graph transmittance from node i to node j and let $G_{\ell k}$ be the branch gain from node k to node ℓ. The ***sensitivity of T_{ij} to $G_{k\ell}$*** is

$$S^{ji}_{\ell k}=\frac{\partial(\ln T_{ji})}{\partial(\ln G_{\ell k})},$$

where T_{ji} is regarded as a function of the scalars G_{sr}, $(s,r)\in\mathcal{I}$. •

Let us try to gather some intuition concerning this definition. If f is a scalar function of a scalar variable x then note that

$$\begin{aligned}\frac{\mathrm{d}(\ln f(x))}{\mathrm{d}(\ln x)}&=\frac{\mathrm{d}(\ln f(e^{\ln x}))}{\mathrm{d}(\ln x)}\\&=\frac{1}{f(e^{\ln x})}\frac{\mathrm{d}(f(e^{\ln x}))}{\mathrm{d}(\ln x)}\\&=\frac{1}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}(\ln x)}\\&=\frac{x}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x}\\&=\lim_{\Delta x\to 0}\frac{f(x+\Delta x)/f(x)}{(x+\Delta x)/x}.\end{aligned}$$

The punchline is that $\frac{\mathrm{d}(\ln f(x))}{\mathrm{d}(\ln x)}$, evaluated at a particular x_0, gives the rate of f, normalised by $f(x_0)$, with respect to x, normalised by x_0. Thus one might say that

$$\frac{\mathrm{d}(\ln f(x))}{\mathrm{d}(\ln x)}=\frac{\mathrm{d}(\%\text{ change in }f)}{\mathrm{d}(\%\text{ change in }x)}.$$

In any event, $S^{ji}_{\ell k}$ measures the dependence of T_{ji} on $G_{\ell k}$ in some sense.

Let us now give a formula for $S^{ji}_{\ell k}$ in terms of graph determinants.

6.15 Proposition *Let $(\mathcal{S},\mathcal{G})$ be a signal flow graph with interconnections $\mathcal{I}\subset\boldsymbol{n}\times\boldsymbol{n}$, let $i,j\in\boldsymbol{n}$, and let $(\ell,k)\in\mathcal{I}$. Let $\mathcal{G}_{(\ell,k)}=\mathcal{G}\setminus\{G_{\ell k}\}$. We then have*

$$S^{ji}_{\ell k}=\frac{\Delta_{\mathcal{S},\mathcal{G}_{(\ell,k)}}}{\Delta_{\mathcal{S},\mathcal{G}}}-\frac{\sum_{P'\in\mathrm{Path}_{ji}(\mathcal{S},\mathcal{G}_{(\ell,k)})}G_{P'}\mathrm{Cof}_{P'}(\mathcal{S},\mathcal{G}_{(\ell,k)})}{\sum_{P\in\mathrm{Path}_{ji}(\mathcal{S},\mathcal{G})}G_P\mathrm{Cof}_P(\mathcal{S},\mathcal{G})}.\tag{6.7}$$

Proof In the proof we shall use the formula $\frac{\mathrm{d}(\ln f(x))}{\mathrm{d}(\ln x)}=\frac{x}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x}$ derived above.

We have $T_{ji}=\frac{F_{ji}}{\Delta_{\mathcal{S},\mathcal{G}}}$ where

$$F_{ji}=\sum_{P\in\mathrm{Path}_{ji}(\mathcal{S},\mathcal{G})}G_P\mathrm{Cof}_P(\mathcal{S},\mathcal{G}).$$

Therefore

$$\frac{\partial(\ln T_{ij})}{\partial(\ln G_{\ell k})}=\frac{\partial(\ln F_{ji})}{\partial(\ln G_{\ell k})}-\frac{\partial(\ln\Delta_{\mathcal{S},\mathcal{G}})}{\partial(\ln G_{\ell k})}.$$

Now we note that each of the expressions F_{ji} and $\Delta_{\mathcal{S},\mathcal{G}}$ is a sum of terms, each of which is either independent of $G_{\ell k}$ or depends linearly on $G_{\ell k}$. Therefore we have

$$G_{\ell k}\frac{\partial F_{ji}}{\partial G_{\ell k}}=(F_{ji})_{\ell k},\quad G_{\ell k}\frac{\partial\Delta_{\mathcal{S},\mathcal{G}}}{\partial G_{\ell k}}=(\Delta_{\mathcal{S},\mathcal{G}})_{\ell k},$$

where $(F_{ji})_{\ell k}$ and $(\Delta_{\mathcal{S},\mathcal{G}})_{\ell k}$ are the terms in F_{ji} and $\Delta_{\mathcal{S},\mathcal{G}}$ that involve $G_{\ell k}$. Therefore,

$$\begin{aligned}\frac{\partial(\ln F_{ji})}{\partial(\ln G_{\ell k})} &= \frac{(F_{ji})_{\ell k}}{F_{ji}} = \frac{F_{ji} - \tilde{F}_{ji}^{\ell k}}{F_{ji}}\\ \frac{\partial(\ln \Delta_{\mathcal{S},\mathcal{G}})}{\partial(\ln G_{\ell k})} &= \frac{(\Delta_{\mathcal{S},\mathcal{G}})_{\ell k}}{\Delta_{\mathcal{S},\mathcal{G}}} = \frac{\Delta_{\mathcal{S},\mathcal{G}} - \tilde{\Delta}_{\mathcal{S},\mathcal{G}}^{\ell k}}{\Delta_{\mathcal{S},\mathcal{G}}},\end{aligned}$$

thus defining $\tilde{F}_{ji}^{\ell k}$ and $\tilde{\Delta}_{\mathcal{S},\mathcal{G}}^{\ell k}$. A moments thought tests the veracity of the formulae

$$\begin{aligned}\tilde{F}_{ji}^{\ell k} &= \sum_{P' \in \text{Path}_{ji}(\mathcal{S},\mathcal{G}_{(\ell,k)})} G_{P'} \text{Cof}_{P'}(\mathcal{S}, \mathcal{G}_{(\ell,k)})\\ \tilde{\Delta}_{\mathcal{S},\mathcal{G}}^{\ell k} &= \Delta_{\mathcal{S},\mathcal{G}_{(\ell,k)}},\end{aligned}$$

giving the result. ■

It is much easier to say in words what the symbols in the result mean. The signal flow graph $(\mathcal{S}, \mathcal{G}_{(\ell,k)})$ is that obtained by removing the branch from node k to node ℓ. Thus the numerators,

$$\Delta_{\mathcal{S},\mathcal{G}_{(\ell,k)}} \quad \text{and} \quad \sum_{P' \in \text{Path}_{ji}(\mathcal{S},\mathcal{G}_{(\ell,k)})} G_{P'} \text{Cof}_{P'}(\mathcal{S}, \mathcal{G}_{(\ell,k)}),$$

in each of the terms on the right-hand side of (6.7) are simply the numerator and denominator for the transfer function from node i to node j in the graph $(\mathcal{S}, \mathcal{G}_{(\ell,k)})$ as given by Mason's rule. Thus these can often be obtained pretty easily. Let us see how this works in an example.

6.16 Example We work with the single-loop signal flow graph of Figure 6.11, reproduced in Figure 6.13. Let us determine the sensitivity of the transfer function T_{51} to the gain G_{43}. It is

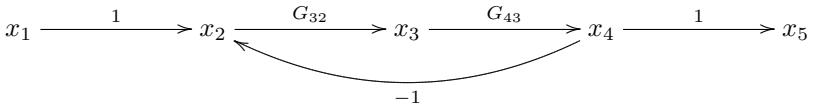


Figure 6.13 The signal flow graph for a negative feedback loop

easy to see that

$$\begin{aligned}\Delta_{\mathcal{S},\mathcal{G}_{(\ell,k)}} &= 1\\ \Delta_{\mathcal{S},\mathcal{G}} &= 1 + G_{32}G_{43}\\ \sum_{P' \in \text{Path}_{51}(\mathcal{S},\mathcal{G}_{(4,3)})} & G_{P'} \text{Cof}_{P'}(\mathcal{S}, \mathcal{G}_{(4,3)}) = 0\\ \sum_{P \in \text{Path}_{ji}(\mathcal{S},\mathcal{G})} & G_P \text{Cof}_P(\mathcal{S}, \mathcal{G}) = G_{32}G_{43}.\end{aligned}$$

From this we see that

$$S_{43}^{51} = \frac{1}{1 + G_{32}G_{43}} + \frac{0}{G_{32}G_{43}} = \frac{1}{1 + G_{32}G_{43}}.$$

We shall see this sensitivity function again in Section 6.3, and it will be an important object when we come consider design issues throughout the remainder of the text. •

Now we consider the closely related concepts of return difference and loop transmittance. First we look at loop transmittance. The loop transmittance is defined relative to a certain branch in a signal flow graph. The idea is that we cut the branch, in doing so creating a new signal flow graph with two new nodes, one a sink and one a source. The loop transmittance is the transmittance in this new graph from the newly created source to the newly created sink. Let us make this precise.

6.17 Definition Let $(\mathcal{S}, \mathcal{G})$ be a signal flow graph with interconnections $\mathcal{I} \subset \boldsymbol{n} \times \boldsymbol{n}$, and let $(j, i) \in \mathcal{I}$. Define a new signal flow graph $(\mathcal{S}_{ji}, \mathcal{G}_{ji})$ with nodes $\{x_1, \dots, x_n, x_{n+1}, x_{n+2}\}$ and with branches

$$\mathcal{G}_{ij} = (\mathcal{G} \setminus \{G_{ji}) \cup \{G_{n+1,i} = 1\} \cup \{G_{j,n+2} = G_{ji}\}.$$

(i) The ***loop transmittance*** through G_{ji} is the transmittance from x_{n+2} to x_{n+1} in the graph $(\mathcal{S}_{ji}, \mathcal{G}_{ji})$, and is denoted L_{ji}.

(ii) The ***return difference*** through G_{ji} is given by $R_{ji} = 1 - L_{ji}$. •

The return difference should be thought of as the difference between a unit signal transmitted from node x_{n+2} and the signal that results at x_{n+1}.

As usual, we want to give a characterisation of the loop transmittance in terms of determinants and related notions.

6.18 Proposition *Let $(\mathcal{S}, \mathcal{G})$ be a signal flow graph with interconnections $\mathcal{I} \subset \boldsymbol{n} \times \boldsymbol{n}$ and let $(j, i) \in \mathcal{I}$. Let $\mathcal{G}_{(j,i)} = \mathcal{G} \setminus \{G_{ji}\}$. We then have*

$$R_{ji} = \frac{\Delta_{\mathcal{S},\mathcal{G}}}{\Delta_{\mathcal{S},\mathcal{G}_{(j,i)}}}, \quad L_{ji} = 1 - \frac{\Delta_{\mathcal{S},\mathcal{G}}}{\Delta_{\mathcal{S},\mathcal{G}_{(j,i)}}}.$$

Proof If there is no loop containing the branch G_{ji} then, by definition, $L_{ji} = 0$ since no forward path connects node x_{n+2} with node x_{n+1} in the graph $(\mathcal{S}_{ji}, \mathcal{G}_{ji})$. This is consistent with the proposition since in this case $\Delta_{\mathcal{S},\mathcal{G}} = \Delta_{\mathcal{S},\mathcal{G}_{ji}}$. Thus the result holds if G_{ji} is not part of a loop. If it is part of loops $L_1, \dots, L_k$ in $(\mathcal{S}, \mathcal{G})$, then these loops will be broken when the graph $(\mathcal{S}_{ji}, \mathcal{G}_{ji})$ is formed. What's more, all elements of $\mathrm{Path}_{n+2,n+1}(\mathcal{S}_{ji}, \mathcal{G}_{ji})$ can be arrived at as follows:

1. Take a loop $L \in \{L_1, \dots, L_k\}$, ordered so that $L = (\{x_i, x_j, \dots\}, \{G_{ji}, \dots\})$.
2. Define a path $P = \{x_{n+2}, x_j, \dots, x_i, x_{n+1}\}$.

Thus the forward paths from x_{n+2} to x_{n+1} in $(\mathcal{S}_{ji}, \mathcal{G}_{ji})$ are in $1-1$-correspondence with the loops containing G_{ji} in $(\mathcal{S}, \mathcal{G})$, and which start with the branch G_{ji}. If $P \in \mathrm{Path}_{n+2,n+1}(\mathcal{S}_{ji}, \mathcal{G}_{ji})$ then we denote by $L_P \in \mathrm{Loop}(\mathcal{S}, \mathcal{G})$ the loop associated to it. The gain of $P \in \mathrm{Path}_{n+2,n+1}(\mathcal{S}_{ji}, \mathcal{G}_{ji})$ is clearly exactly the gain of L_P. Furthermore, the cofactor of a path $P \in \mathrm{Path}_{n+2,n+1}(\mathcal{S}_{ji}, \mathcal{G}_{ji})$ consists exactly of those terms in $\Delta_{\mathcal{S},\mathcal{G}}$ not involving the loop $L \in \mathrm{Loop}(\mathcal{S}, \mathcal{G})$ giving rise to P. This fact can be employed to verify the equality

$$\Delta_{\mathcal{S},\mathcal{G}} = \Delta_{\mathcal{S}_{ji},\mathcal{G}_{ji}} - \sum_{P \in \mathrm{Path}_{n+2,n+1}(\mathcal{S}_{ji},\mathcal{G}_{ji})} G_{L_P} \mathrm{Cof}_P(\mathcal{S}_{ji}, \mathcal{G}_{ji}).$$

This then means exactly that

$$\sum_{P \in \mathrm{Path}_{n+2,n+1}(\mathcal{S}_{ji},\mathcal{G}_{ji})} G_P \mathrm{Cof}_P(\mathcal{S}_{ji}, \mathcal{G}_{ji}) = \Delta_{\mathcal{S}_{ji},\mathcal{G}_{ji}} - \Delta_{\mathcal{S},\mathcal{G}}.$$

We also clearly have $\Delta_{\mathcal{S}_{ji},\mathcal{G}_{ji}} = \Delta_{\mathcal{S},\mathcal{G}_{(j,i)}}$ since the loops in the graphs $(\mathcal{S}_{ji},\mathcal{G}_{ji})$ and $(\mathcal{S},\mathcal{G}_{(j,i)})$ agree. Thus

$$L_{ji} = \frac{\sum_{P\in\text{Path}_{n+2,n+1}(\mathcal{S}_{ji},\mathcal{G}_{ji})} G_P\text{Cof}_P(\mathcal{S}_{ji},\mathcal{G}_{ji})}{\Delta_{\mathcal{S}_{ji},\mathcal{G}_{ji}}} = \frac{\Delta_{\mathcal{S},\mathcal{G}_{(j,i)}} - \Delta_{\mathcal{S},\mathcal{G}}}{\Delta_{\mathcal{S},\mathcal{G}_{(j,i)}}},$$

giving the result. ■

This is, as always, easily exhibited in an example.

6.19 Example **(Example 6.16 cont'd)** We consider again the negative feedback loop of Figure 6.13. We had computed in Example 6.16 that

$$\Delta_{\mathcal{S},\mathcal{G}_{(4,3)}} = 1, \quad \Delta_{\mathcal{S},\mathcal{G}} = 1 + G_{32}G_{43},$$

which gives

$$R_{43} = 1 + G_{32}G_{43}, \quad L_{43} = -G_{32}G_{43}.$$

We see how should interpret the loop transmittance in this example. • finish

6.2 Interconnected SISO linear systems

The discussion of the previous section had a fairly general nature. Although this had the advantage of allowing us to get a handle on the essential features of a signal flow graph, let us bring things back into the realm of control systems. We do this in this section by introducing the notion of an interconnected SISO linear system, and discussing the properties of such things in a their general context. This treatment does not seem to appear in the current control literature, oddly enough.

6.2.1 Definitions and basic properties

First of all, let us define what we mean by an interconnected SISO linear system.

6.20 Definition An ***interconnected SISO linear system*** is a connected signal flow graph $(\mathcal{S},\mathcal{G})$ with one source (the ***input***) and one sink (the ***output***). If the nodes for the system are $\{x_1,\dots,x_n\}$,it is always assumed the source is x_1 and the sink is x_n. We assume that the single source is an actual input (i.e., that there is one branch originating from x_1 and that the branch has gain 1). This can always be done without loss of generality by adding the necessary node and branch if needed.

An interconnected SISO linear system $(\mathcal{S},\mathcal{G})$ is ***proper*** (resp. ***strictly proper***) if all gains in $\mathcal{G}$ are proper (resp. strictly proper). •

For example, the signal flow graphs of Figures 6.1, 6.4, 6.5, 6.6, and 6.12 are interconnected SISO linear systems, while that of Figure 6.7 is not.

We will want to determine the transfer function between any two signals in the signal flow graph in order to discuss stability. Of course, Mason's Rule makes this a comparatively simple chore. Since our system does not necessarily have any inputs, in order to talk about the transfer function from any signal to any other signal, we should introduce an input an arbitrary node. We do this in the obvious way, as follows, supposing $(\mathcal{S},\mathcal{G})$ to have n nodes. For $i\in\{1,\dots,n\}$ define the ***ith-appended system*** to be the signal flow graph $(\mathcal{S}_i,\mathcal{G}_i)$ with $\mathcal{S}_i = \mathcal{S}\cup\{x_{n+1}=u_i\}$ and $\mathcal{G}_i = \mathcal{G}\cup\{G_{i,n+1}=1\}$. Thus we simply "tack on" another node with a branch of gain 1 going to node i. This renders the new node u_i an input, and

the transfer function from node i to node j is then the graph transmittance T_{ji}, this being determined by Mason's Rule. In particular, we define the ***transfer function*** of $(\mathcal{S}, \mathcal{G})$ by

$$T_{\mathcal{S},\mathcal{G}} = T_{n1} \in \mathbb{R}(s).$$

If $(N_{\mathcal{S},\mathcal{G}}, D_{\mathcal{S},\mathcal{G}})$ denotes the c.f.r. of $T_{\mathcal{S},\mathcal{G}}$, then we have reduced an interconnected SISO linear system to a SISO linear system in input/output form.

Let us compute the transfer function for the examples we have been using.

6.21 Examples We simply go through the steps and produce the transfer function after making the necessary simplifications. In each case, the gain G_{ij} of a branch is represented by its c.f.r. (N_{ij}, D_{ij})).

1. For the series interconnection of Figure 6.4 we ascertain that the transfer function is

$$T_{\mathcal{S},\mathcal{G}} = \frac{N_{21}N_{32}}{D_{21}D_{32}}.$$

2. For the parallel signal flow graph we have loops,

$$T_{\mathcal{S},\mathcal{G}} = \frac{N_{21}N_{42}D_{31}D_{43} + N_{31}N_{43}D_{21}D_{42}}{D_{21}D_{42}D_{31}D_{43}}.$$

3. Next we turn to the feedback loop of Figure 6.6. In Example 6.13–1 we computed the transfer function from x_1 to x_5 to be

$$T_{\mathcal{S},\mathcal{G}} = \frac{\frac{N_{21}N_{32}N_{43}N_{54}}{D_{21}D_{32}D_{43}D_{54}}}{1 - \frac{N_{32}N_{43}N_{24}}{D_{32}D_{43}D_{24}}}.$$

Simplification gives

$$T_{\mathcal{S},\mathcal{G}} = \frac{N_{21}N_{32}N_{43}N_{54}D_{24}}{D_{32}D_{43}D_{21}D_{24}D_{54} - N_{32}N_{43}N_{24}D_{21}D_{54}}.$$

In the case when $G_{21} = G_{54} = -G_{24} = 1$ we get

$$\frac{N_{32}N_{43}}{D_{32}D_{43} + N_{32}N_{43}}. \tag{6.8}$$

4. For the four loop signal flow graph of Figure 6.12 we determine the transfer function to be

$$T_{\mathcal{S},\mathcal{G}} = \frac{N_{61}N_{76}N_{87}D_{67}D_{78}}{D_{61}(D_{67}D_{76}(D_{78}D_{87} - N_{78}N_{87}) - N_{67}N_{76}D_{78}D_{87})}.$$ •

Of course, in simply writing the transfer function for an interconnected SISO linear system we have eliminated a great deal of the structure of the signal flow graph. As when one thinks of a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ as being merely an input/output system (see Section 2.3), one must take care if using only the transfer function to talk about an interconnected SISO linear system. In order to see how this goes, we need to set up the appropriate structure for an interconnected SISO linear system.

The first order of business is to set up equations of motion for a SISO linear system. We give n sets of equations of motion, depending on where the input for the systems appears. Of course, one can consider all inputs acting together by linearity. Note that the assumption that x_1 is an input, and the only source, ensures that the structure matrices for all of the n appended systems are actually the same. Now we give a procedure for going from the structure matrix $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ of rational functions to a polynomial matrix $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}$ and a polynomial vector $\boldsymbol{b}^i_{\mathcal{S},\mathcal{G}}$, corresponding to the ith-appended system $(\mathcal{S}_i, \mathcal{G}_i)$.

6.22 Procedure for constructing $(\mathbf{A}_{\mathcal{S},\mathcal{G}}, \mathbf{b}^{\mathsf{i}}_{\mathcal{S},\mathcal{G}})$ from $\mathbf{G}_{\mathcal{S},\mathcal{G}}$ Given $\boldsymbol{G}_{\mathcal{S},\mathcal{G}} \in \mathbb{R}(s)^{n\times n}$, do the following:

(i) let (N_{ij}, D_{ij}) be the c.f.r. of each branch gain $G_{ij} \in \mathcal{G}$;

(ii) for each $i \in \{1, \dots, n\}$, let $G_{ij_1}, \dots, G_{ij_\ell}$ be the nonzero gains appearing in the ith row of $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$;

(iii) multiply row i by the denominators $D_{ij_1}, \dots, D_{ij_\ell}$;

(iv) after doing this for each row, denote the resulting matrix by $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}$;

(v) define $\boldsymbol{b}^i_{\mathcal{S},\mathcal{G}}$ to be the n-vector whose ith component is $D_{ij_1} \dots D_{ij_\ell}$, the rest of the components of $\boldsymbol{b}^i_{\mathcal{S},\mathcal{G}}$ being zero. •

The ***equations of motion*** for the ith-appended system $(\mathcal{S}_i, \mathcal{G}_i)$ are then the differential equations

$$\boldsymbol{A}_{\mathcal{S},\mathcal{G}}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{b}^i_{\mathcal{S},\mathcal{G}}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)u(t), \tag{6.9}$$

where $\boldsymbol{x} = (x_1, \dots, x_n)$ are the signals for $(\mathcal{S}, \mathcal{G})$ and where $u = u_i$ is the input at node i. The procedure above systematises what one would do naturally in writing the equations of motion obtain by doing a "node balance." In this case, one would clear the denominators so that all expressions would be polynomial, and so represent differential equations in the time-domain. Let us introduce the notation $\mathbb{R}[s]^{n\times n}$ for an $n \times n$ matrix with components in $\mathbb{R}[s]$. Thus, for example, $\boldsymbol{A}_{\mathcal{S},\mathcal{G}} \in \mathbb{R}[s]^{n\times n}$. Let us also define $\boldsymbol{B}_{\mathcal{S},\mathcal{G}} \in \mathbb{R}(s)$ by

$$\boldsymbol{B}_{\mathcal{S},\mathcal{G}} = \left[\ \boldsymbol{b}^1_{\mathcal{S},\mathcal{G}} \ \middle|\ \cdots \ \middle|\ \boldsymbol{b}^n_{\mathcal{S},\mathcal{G}}\ \right],$$

as a convenient way to catalogue the input vectors $\boldsymbol{b}^1_{\mathcal{S},\mathcal{G}}, \dots, \boldsymbol{b}^n_{\mathcal{S},\mathcal{G}}$.

6.23 Remark Clearly, the equations of motion for the ith appended system are exactly equivalent to the equations $\boldsymbol{G}_{\mathcal{S},\mathcal{G}} = \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the ith standard basis vector. From this it follows that

$$\boldsymbol{A}^{-1}_{\mathcal{S},\mathcal{G}}\boldsymbol{B}_{\mathcal{S},\mathcal{G}} = \boldsymbol{G}^{-1}_{\mathcal{S},\mathcal{G}}.$$ •

Let us perform these operations for the examples we have been toting around.

6.24 Examples **(Example 6.5 cont'd)** We shall simply produce $(\boldsymbol{A}_{\mathcal{S},\mathcal{G}}, \boldsymbol{B}_{\mathcal{S},\mathcal{G}})$ by applying Procedure 6.22.

1. We determine

$$\boldsymbol{A}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 \\ -N_{21} & D_{21} & 0 \\ 0 & -N_{32} & D_{32} \end{bmatrix}, \quad \boldsymbol{B}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & D_{21} & 0 \\ 0 & 0 & D_{32} \end{bmatrix}.$$

2. We determine

$$\boldsymbol{A}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -N_{21} & D_{21} & 0 & 0 \\ -N_{31} & 0 & D_{31} & 0 \\ 0 & -N_{42}D_{43} & -N_{43}D_{43} & D_{42}D_{43} \end{bmatrix}, \quad \boldsymbol{B}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & D_{21} & 0 & 0 \\ 0 & 0 & D_{31} & 0 \\ 0 & 0 & 0 & D_{42}D_{43} \end{bmatrix}.$$

3. We determine

$$\boldsymbol{A}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -N_{21}D_{24} & D_{21}D_{24} & 0 & -N_{24}D_{21} & 0 \\ 0 & -N_{32} & D_{32} & 0 & 0 \\ 0 & 0 & -N_{43} & D_{43} & 0 \\ 0 & 0 & 0 & -N_{54} & D_{54} \end{bmatrix}$$

$$\boldsymbol{B}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & D_{21}D_{24} & 0 & 0 & 0 \\ 0 & 0 & D_{32} & 0 & 0 \\ 0 & 0 & 0 & D_{43} & 0 \\ 0 & 0 & 0 & 0 & D_{54} \end{bmatrix}.$$

4. Now let us also look at the four-loop example first introduced in Example 6.9 and shown in Figure 6.12. We have not yet defined $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ so let us do so:

$$\boldsymbol{G}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -G_{21} & 1 & -G_{23} & 0 & 0 & 0 & 0 & 0 \\ 0 & -G_{32} & 1 & -G_{34} & 0 & 0 & 0 & 0 \\ 0 & 0 & -G_{43} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -G_{54} & 1 & 0 & 0 & -G_{58} \\ -G_{61} & 0 & 0 & 0 & 0 & 1 & -G_{67} & 0 \\ 0 & 0 & 0 & 0 & 0 & -G_{76} & 1 & -G_{78} \\ 0 & 0 & 0 & 0 & 0 & 0 & -G_{87} & 1 \end{bmatrix}.$$

We then have

$$\boldsymbol{A}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -N_{21}D_{23} & D_{21}D_{23} & -N_{23}D_{21} & 0 & 0 & 0 & 0 & 0 \\ 0 & -N_{32}D_{34} & D_{32}D_{34} & -N_{34}D_{32} & 0 & 0 & 0 & 0 \\ 0 & 0 & -N_{43} & D_{43} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -N_{54}D_{58} & D_{54}D_{58} & 0 & 0 & -N_{58}D_{54} \\ -N_{61}D_{67} & 0 & 0 & 0 & 0 & D_{61}D_{67} & -N_{67}D_{61} & 0 \\ 0 & 0 & 0 & 0 & 0 & -N_{76}D_{78} & D_{76}D_{78} & -N_{78}D_{76} \\ 0 & 0 & 0 & 0 & 0 & 0 & -N_{87} & D_{87} \end{bmatrix}$$

$$\boldsymbol{B}_{\mathcal{S},\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & D_{21}D_{23} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & D_{32}D_{34} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & D_{43} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & D_{54}D_{58} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & D_{61}D_{67} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & D_{76}D_{78} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & D_{87} \end{bmatrix}.$$

This procedure for coming up with $(\boldsymbol{A}_{\mathcal{S},\mathcal{G}}, \boldsymbol{B}_{\mathcal{S},\mathcal{G}})$ is clearly simple enough, although perhaps tedious, in any given example. •

Now that we have the equations of motion (6.9) for an interconnected SISO linear system, we can proceed to analyse these equations.

6.2.2 Well-posedness

The next matter we deal with is a new one for us, the matter of well-posedness. That there is something to talk about is best illustrated by an example.

6.25 Example We consider the signal flow graph of Figure 6.14 where we take $R_L(s) = \frac{1+s-s^2}{s^2+s+1}$.

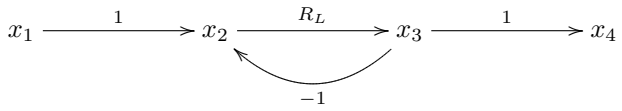


Figure 6.14 Unity gain feedback loop

Thus we have a perfectly well behaved collection of branch gains, and one can compute the characteristic polynomial (see next section) to be $P_{\mathcal{S},\mathcal{G}}(s) = 2(s+2)$, which is Hurwitz. This indicates that everything is pleasant. However, we compute the transfer function of the system to be

$$T_{\mathcal{S},\mathcal{G}}(s) = \frac{1-s}{2}.$$

This is problematic: an interconnection of proper branch gains has given rise to an improper transfer function. Such cases are undesirable as improper transfer functions are certainly not desirable (cf. Proposition 5.14). •

Clearly one would like all graph transmittances to be proper rational functions, and this leads to the following definition.

6.26 Definition An interconnected SISO linear system $(\mathcal{S},\mathcal{G})$ is ***well-posed*** if for each $i \in \{1,\dots,n\}$ the graph transmittance $T_{ji} \in \mathbb{R}(s)$ is strictly proper for each $j \in \{1,\dots,n\}$. •

Thus well-posedness is the requirement that n^2 rational functions be proper. One would like to derive simpler conditions for well-posedness. Starting down this road, the following result gives an interpretation of well-posedness in terms of the determinant $\Delta_{\mathcal{S},\mathcal{G}}$.

6.27 Proposition *A proper interconnected SISO linear system $(\mathcal{S},\mathcal{G})$ is well-posed if and only if $\lim_{s\to\infty} \Delta_{\mathcal{S},\mathcal{G}}(s) \neq 0$.*

Proof Suppose that $\lim_{s\to\infty} \Delta_{\mathcal{S},\mathcal{G}}(s) = L \neq 0$. The graph transmittance T_{ji} is given by

$$T_{ji} = \sum_{P\in \mathrm{Path}_{ji}(\mathcal{S},\mathcal{G})} \frac{G_P \mathrm{Cof}_P(\mathcal{S},\mathcal{G})}{\Delta_{\mathcal{S},\mathcal{G}}}.$$

Since $(\mathcal{S},\mathcal{G})$ is proper, it follows that $G_P\mathrm{Cof}_P(\mathcal{S},\mathcal{G})$ is proper so that $\lim_{s\to\infty} G_P(s)\mathrm{Cof}_P(\mathcal{S},\mathcal{G})(s)$ is finite. Therefore,

$$\begin{aligned}\lim_{s\to\infty} T_{ji}(s) &= \sum_{P\in \mathrm{Path}_{ji}(\mathcal{S},\mathcal{G})} \frac{\lim_{s\to\infty} G_P(s)\mathrm{Cof}_P(\mathcal{S},\mathcal{G})(s)}{\lim_{s\to\infty} \Delta_{\mathcal{S},\mathcal{G}}(s)} \\ &= \frac{1}{L} \sum_{P\in \mathrm{Path}_{ji}(\mathcal{S},\mathcal{G})} \lim_{s\to\infty} G_P(s)\mathrm{Cof}_P(\mathcal{S},\mathcal{G})(s)\end{aligned}$$

is finite. Thus T_{ji} is proper.

Conversely, suppose that the matrix $\boldsymbol{T}_{\mathcal{S},\mathcal{G}} = \boldsymbol{G}_{\mathcal{S},\mathcal{G}}^{-1}$ of transmittances consists of proper rational functions. Since the determinant of $\boldsymbol{T}_{\mathcal{S},\mathcal{G}}$ consists of sums of products of these proper rational functions, it follows that $\det \boldsymbol{T}_{\mathcal{S},\mathcal{G}}$ is itself a proper rational function. Therefore

$$\lim_{s\to\infty} \det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}^{-1}(s) \neq \infty.$$

From this we infer that

$$\lim_{s\to\infty} \det \boldsymbol{G}_{\mathcal{S},\mathcal{G}}(s) \neq 0,$$

so proving the result. ■

This gives the following sufficient condition for well-posedness, one that is satisfied in many examples.

6.28 Corollary *Let $(\mathcal{S},\mathcal{G})$ be a proper interconnected SISO linear system. If each loop in $(\mathcal{S},\mathcal{G})$ contains a branch whose gain is strictly proper, then $(\mathcal{S},\mathcal{G})$ is well-posed.*

Proof Recall that the determinant is given by

$$\Delta_{\mathcal{S},\mathcal{G}} = 1 + \sum_{\alpha} G_\alpha$$

where G_α is a product of loop gains for nontouching loops. If $(\mathcal{S},\mathcal{G})$ is proper, and each loop contains a strictly proper branch gain, then we have

$$\lim_{s\to\infty} \Delta_{\mathcal{S},\mathcal{G}}(s) = 1,$$

implying well-posedness. ■

This indicates that for many physical systems, whose branches will be comprised of strictly proper rational functions, one can expect well-posedness to be "typical." The following example reexamines our introductory example in light of our better understanding of well-posedness.

6.29 Example **(Example 6.25 cont'd)** Let us still use the signal flow graph of Figure 6.14, but now take

$$R_L(s) = \frac{1 + 2 + as^2}{s^2 + s + 1},$$

where $a \in \mathbb{R}$ is unspecified. We compute

$$\Delta_{\mathcal{S},\mathcal{G}}(s) = 1 + R_L(s) = \frac{(1+a)s^2 + 2s + 2}{s^2 + s + 1}.$$

We see that $\lim_{s\to\infty} \Delta_{\mathcal{S},\mathcal{G}}(s) = 0$ if and only if $a = -1$. Thus, even though this system does not satisfy the sufficient conditions of well-posedness in Corollary 6.28, it is only in the very special case when $a = -1$ that the system is not well-posed. •

Well-posedness in a general context is discussed by Willems [1971], and for general (even nonlinear) interconnected systems by Vidyasagar [1981]. When talking about well-posedness for systems outside the rational function context we use here, one no longer has access to simple notions like properness to characterise what it might mean for a system to be well-posed. Thus, for general systems, one uses a more basic idea connected with existence and uniqueness of solutions. This is explored in Exercise E6.8.

6.2.3 Stability for interconnected systems

In Chapter 5 we looked at certain types of stability: internal stability for SISO linear systems, and BIBO stability for input/output systems. In this chapter, we are concerned with interconnections of systems in input/output form, and the introduction of such interconnections gives rise to stability concerns that simply do not arise when there are no interconnections. The difficulty here is that the interconnections make possible some undesirable behaviour that is simply not captured by the transfer function $T_{\mathcal{S},\mathcal{G}}$ for the system.

A motivating example and definitions

The following example makes this clear the difficulties one can encounter due to the introduction of even simple interconnections.

6.30 Example We consider the simple block diagram configuration of Figure 6.15. Thus $R_1(s) =$

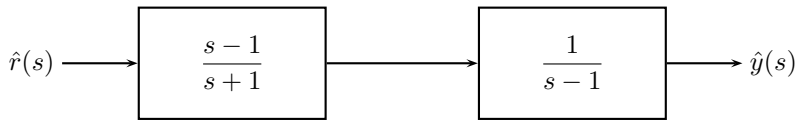


Figure 6.15 Trouble waiting to happen

$\frac{s-1}{s+1}$ and $R_2(s) = \frac{1}{s-1}$. As we saw in Section 3.1, $\hat{y} = R_1 R_2 \hat{r}$. This gives

$$\frac{\hat{y}(s)}{\hat{r}(s)} = \frac{1}{s+1}.$$

By Proposition 5.13 we see that this input/output transfer function is BIBO stable, and so on these grounds we'd wash our hands of the stability question and walk away. In doing so, we'd be too hasty. To see why this is so, suppose that the system admits some noise $\hat{n}$ as in Figure 6.16. The transfer function between $\hat{n}$ and $\hat{y}$ is then

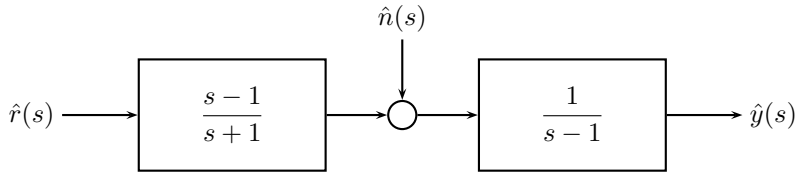


Figure 6.16 Trouble happening

$$\frac{\hat{y}(s)}{\hat{n}(s)} = \frac{1}{s-1}$$

which by Proposition 5.13 is not BIBO stable. So any slight perturbations in the signal as it goes from the R_1 block to the R_2 block will potentially be dangerously magnified in the output. •

We now address the difficulties of the above example with a notion of stability that makes sense for interconnected systems. The following definition provides notions of stability that are relevant for interconnected SISO linear systems.

6.31 Definition An interconnected SISO linear system $(\mathcal{S}, \mathcal{G})$ with nodes $\{x_1, \dots, x_n\}$ is

(i) ***internally stable*** if

$$\limsup_{t \to \infty} \|\boldsymbol{x}(t)\| < \infty$$

for every solution $\boldsymbol{x}(t)$ of $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = \boldsymbol{0}$;

(ii) ***internally asymptotically stable*** if

$$\lim_{t\to\infty} \|\boldsymbol{x}(t)\| = 0$$

for every solution $\boldsymbol{x}(t)$ of $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$;

(iii) ***internally unstable*** if it is not internally stable;

(iv) ***BIBO stable*** if there exists a constant $K > 0$ so that the conditions (1) $\boldsymbol{x}(0) = \boldsymbol{0}$ and (2) $|u(t)| \leq 1$, $t \geq 0$ imply that $x_n(t) \leq K$ where $u(t)$ and $\boldsymbol{x}(t)$ satisfy (6.9) with $i = 1$;

(v) ***interconnected bounded input, bounded output stable*** (***IBIBO stable***) if for each $i \in \{1, \dots, n\}$, the graph transmittance T_{ji} is BIBO stable for $j \in \{1, \dots, n\}$. •

For interconnected systems, we have all the notions of stability for "normal" systems, plus we have this new notion of IBIBO stability that deals with the input/output stability of the interconnection. This new type of stability formalises the procedure of adding noise to each signal, and "tapping" the output of each signal to see how the system reacts internally, apart from just looking at how the given input and output nodes act. Thus, if noise added to a node gets unstably magnified in the signal at some other node, our definition of IBIBO stability will capture this. Note that internal stability for these systems does not follow in quite the same way as for SISO linear systems since the equations for the two systems are fundamentally different: $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$ versus $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$. Indeed, one can readily see that the latter is a special case of the former (see Exercise E6.3). A systematic investigation of equations of the form $\boldsymbol{P}(\frac{\mathrm{d}}{\mathrm{d}t})\boldsymbol{x}(t) = \boldsymbol{0}$ for an arbitrary matrix of polynomials $\boldsymbol{P}$ is carried out by Polderman and Willems [1998]. This can be thought of as a generalisation of the computation of the matrix exponential. Since we shall not benefit from a full-blown treatment of such systems, we do not give it, although some aspects of the theory certainly come up in our characterisation of the internal stability of an interconnected SISO linear system.

Before we proceed to give results that we can use to test for the stability of interconnected systems, let us see how the above definition of IBIBO stability covers our simple example.

6.32 Example **(Example 6.30 cont'd)** It is more convenient here to use the signal flow graph, and we show it in Figure 6.17 with the nodes relabelled to make it easier to apply the

$$x_1(s) \xrightarrow{\frac{s-1}{s+1}} x_2(s) \xrightarrow{\frac{1}{s-1}} x_3(s)$$

Figure 6.17 The signal flow diagram corresponding to the block diagram of Figure 6.15

definition of IBIBO stability. The appended systems are shown in Figure 6.18. We compute the graph transmittances to be

$$T_{21}(s) = \frac{s-1}{s+1}, \quad T_{31}(s) = \frac{1}{s+1}, \quad T_{32}(s) = \frac{1}{s-1}.$$

Note that the transfer function T_{32} is BIBO unstable by Proposition 5.13. Therefore, the interconnected system in Figure 6.17 is not IBIBO stable. •

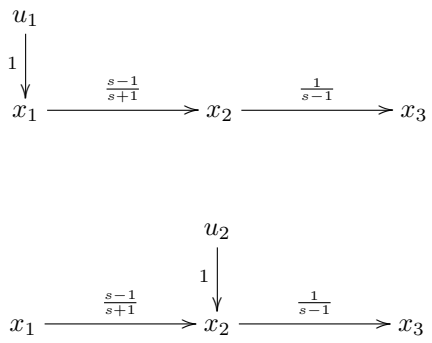


Figure 6.18 The ith-appended systems for Figure 6.17 with $i = 1$ (top) and $i = 2$ (bottom)

Conditions for internal stability

Let us now produce some results concerning the various types of stability. To get things rolling, we define the ***characteristic polynomial*** to be $P_{\mathcal{S},\mathcal{G}} = \det \boldsymbol{A}_{\mathcal{S},\mathcal{G}}$. The ***algebraic multiplicity*** of a root λ of $P_{\mathcal{S},\mathcal{G}}$ is the multiplicity of the root of a polynomial in the usual sense (see Section C.1). The ***geometric multiplicity*** of a root λ of $P_{\mathcal{S},\mathcal{G}}$ is the dimension of the kernel of the matrix $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}(\lambda)$. Note that $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}(\lambda) \in \mathbb{R}^{n\times n}$, so its kernel is defined as usual. Following what we do for $\mathbb{R}$-matrices, if $\lambda \in \mathbb{C}$ is a root of $P_{\mathcal{S},\mathcal{G}}$, let us denote the algebraic multiplicity by $m_a(\lambda)$ and the geometric multiplicity by $m_g(\lambda)$.

6.33 Theorem *Consider a proper interconnected SISO linear system $(\mathcal{S},\mathcal{G})$. The following statements hold.*

(i) $(\mathcal{S},\mathcal{G})$ *is internally unstable if* $\operatorname{spec}(P_{\mathcal{S},\mathcal{G}}) \cap \mathbb{C}_+ \neq \emptyset$.

(ii) $(\mathcal{S},\mathcal{G})$ *is internally asymptotically stable if* $\operatorname{spec}(P_{\mathcal{S},\mathcal{G}}) \subset \mathbb{C}_-$.

(iii) $(\mathcal{S},\mathcal{G})$ *is internally stable if* $\operatorname{spec}(P_{\mathcal{S},\mathcal{G}}) \cap \mathbb{C}_+ = \emptyset$ *and if* $m_g(\lambda) = m_a(\lambda)$ *for* $\lambda \in \operatorname{spec}(P_{\mathcal{S},\mathcal{G}}) \cap (i\mathbb{R})$.

(iv) $(\mathcal{S},\mathcal{G})$ *is internally unstable if* $m_g(\lambda) < m_a(\lambda)$ *for* $\lambda \in \operatorname{spec}(P_{\mathcal{S},\mathcal{G}}) \cap (i\mathbb{R})$.

Proof Just as the proof of Theorem 5.2 follows easily once one understands the nature of the matrix exponential, the present result follows easily once one understands the character of solutions of $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big) = \boldsymbol{0}$. The following lemma records those aspects of this that are useful for us. We simplify matters by supposing that we are working with complex signals. If the signals are real, then the results follow by taking real and imaginary parts of what we do here.

1 Lemma *Let $\boldsymbol{P} \in \mathbb{C}[s]^{n\times n}$ and suppose that $\det \boldsymbol{P}$ is not the zero polynomial. Let $\{\lambda_1,\dots,\lambda_\ell\}$ be the roots of $\det \boldsymbol{P}$ with respective multiplicities $m_1,\dots,m_\ell$. Every solution of $\boldsymbol{P}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$ has the form*

$$\boldsymbol{x}(t) = \sum_{i=1}^{\ell} \sum_{j=0}^{m_i-1} \boldsymbol{\beta}_{ij} t^j e^{\lambda_i t}$$

for some appropriate $\boldsymbol{\beta}_{ij} \in \mathbb{C}^n$, $i = 1,\dots,\ell$, $j = 0,\dots,m_i-1$. In particular, the set of solutions of $\boldsymbol{P}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$ forms a $\mathbb{R}$-vector space of dimension $\deg(\det \boldsymbol{P})$.

Proof We prove the lemma by induction on n. It is obvious for $n = 1$ by well-known properties of scalar differential equations as described in Section B.1. Now suppose the lemma true for $n \in \{1, \dots, k-1\}$ and let $\boldsymbol{P} \in \mathbb{R}[s]^{k \times k}$. Note that neither the solutions of the equations $\boldsymbol{P}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$ nor the determinant $\det \boldsymbol{P}$ are changed by the performing of elementary row operations on $\boldsymbol{P}$. Thus we may suppose that $\boldsymbol{P}$ has been row reduced to the form

$$\boldsymbol{P} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ 0 & & & \\ \vdots & & \tilde{\boldsymbol{P}} & \\ 0 & & & \end{bmatrix}$$

for some $\tilde{\boldsymbol{P}} \in \mathbb{R}[s]^{(k-1)\times(k-1)}$. By the induction hypothesis every solution to $\boldsymbol{Q}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$ has the form

$$\tilde{\boldsymbol{x}}(t) = \sum_{i=1}^{\tilde{\ell}} \sum_{j=0}^{m_i - 1} \tilde{\boldsymbol{\beta}}_{ij} t^j e^{\lambda_i t},$$

for some appropriate $\tilde{\boldsymbol{\beta}}_{ij} \in \mathbb{C}^n$, $i = 1, \dots, \tilde{\ell}$, $j = 0, \dots, m_i - 1$. Thus $\{\lambda_1, \dots, \lambda_{\tilde{\ell}}\}$ are the roots of $\det \tilde{\boldsymbol{P}}$ with respective multiplicities $m_1, \dots, m_{\tilde{\ell}}$. Substituting such a $\tilde{x}(t)$ into the first of the equations $\boldsymbol{P}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$ gives a differential equation of the form

$$P_{11}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_1(t) = \sum_{i=1}^{\tilde{\ell}} \sum_{j=0}^{m_i - 1} \alpha_{ij} t^j e^{\lambda_i t}, \tag{6.10}$$

for some appropriate $\alpha_{ij} \in \mathbb{C}$, $i = 1, \dots, \tilde{\ell}$, $j = 0, \dots, m_i - 1$. Thus the constants α_{ij} will depend on the components of the $\tilde{\boldsymbol{\beta}}_{ij}$'s, the coefficients of $P_{12}, \dots, P_{1k}$, and roots $\lambda_1, \dots, \lambda_{\tilde{\ell}}$. To solve this equation (or at least come up with the form of a solution) we recall that it will be the sum of a homogeneous solution $x_{ih}(t)$ satisfying $P_{11}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)x_{1h}(t) = 0$ and a particular solution. Let us investigate the form of both of these components of the solution. The homogeneous solution has the form

$$x_{1h}(t) = \sum_{i=1}^{\ell'} \sum_{j=0}^{m'_i} \beta'_{ij} t^j e^{\lambda'_i t},$$

where $\{\lambda'_1, \dots, \lambda'_{\ell'}\}$ are the roots of P_{11} with respective multiplicities $m'_1, \dots, m'_{\ell'}$. If there are no common roots between P_{11} and $\det \tilde{\boldsymbol{P}}$ then the particular solution will have the form

$$x_{1p}(t) = \sum_{i=1}^{\tilde{\ell}} \sum_{j=0}^{m_i - 1} \beta''_{ij} t^j e^{\lambda_i t},$$

i.e., the same form as the right-hand side of (6.10). However, if P_{11} and $\det \tilde{\boldsymbol{P}}$ *do* share roots, we have to be a little more careful. Suppose that P_{11} and $\det \tilde{\boldsymbol{P}}$ have a common root λ of multiplicity m' for P_{11} and multiplicity m for $\det \tilde{\boldsymbol{P}}$. A moments reflection shows that the method of undetermined coefficients, as outlined in Procedure B.2, then produces a particular solution corresponding to this common root of the form

$$\sum_{j=0}^{m+m'-1} \beta''_j t^j e^{\lambda t}.$$

Collecting this all together yields the lemma by induction once we realise that $\det \boldsymbol{P} = P_{11} \det \tilde{\boldsymbol{P}}$. ▼

(i) If $\lambda \in \operatorname{spec}(P_{\Sigma,\mathcal{G}}) \cap \mathbb{C}_+$ then there exists a vector $\boldsymbol{u} \in \mathbb{C}^n$ so that $\boldsymbol{A}_{\Sigma,\mathcal{G}}(\lambda)\boldsymbol{u} = \boldsymbol{0}$. If $\boldsymbol{x}(t) = e^{\lambda t}\boldsymbol{u}$ then we have

$$\boldsymbol{A}_{\Sigma,\mathcal{G}}\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = e^{\lambda t}\boldsymbol{A}_{\Sigma,\mathcal{G}}(\lambda)\boldsymbol{u} = \boldsymbol{0}.$$

Thus there is a solution of $\boldsymbol{A}_{\Sigma,\mathcal{G}}\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = \boldsymbol{0}$ that is unbounded.

(ii) From Lemma 1 we know that every solution of $\boldsymbol{A}_{\Sigma,\mathcal{G}}\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = \boldsymbol{0}$ is a vector function with components being linear combinations of functions of the form

$$t^j e^{\lambda t} \tag{6.11}$$

where $\lambda \in \operatorname{spec}(P_{\Sigma,\mathcal{G}})$. By hypothesis, it follows that every such function approaches $\boldsymbol{0}$ as $t \to \infty$, and so every solution of $\boldsymbol{A}_{\Sigma,\mathcal{G}}\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = \boldsymbol{0}$ approaches $\boldsymbol{0}$ as $t \to \infty$. We refer to the proof of Theorem 5.2 to see how this is done properly.

(iii) By Lemma 1, every solution of $\boldsymbol{A}_{\Sigma,\mathcal{G}}\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = \boldsymbol{0}$ is a vector functions whose components are linear combinations of terms of the form (6.11) for $\operatorname{Re}(\lambda) < 0$ and terms of the form

$$t^j e^{i\omega t}. \tag{6.12}$$

We must show that the condition that $m_a(i\omega) = m_g(i\omega)$ implies that the only solutions of the form (6.12) that are allowed occur with $j = 0$. Indeed, since $m_a(i\omega) = m_g(i\omega)$, it follows that there are $\ell \triangleq m_a(\lambda)$ linearly independent vectors, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell$, in $\ker(\boldsymbol{A}_{\Sigma,\mathcal{G}}(i\omega))$. Therefore, this implies that the functions

$$\boldsymbol{u}_i e^{i\omega t}, \ldots, \boldsymbol{u}_\ell e^{i\omega t}$$

are linearly independent solutions corresponding to the root $i\omega \in \operatorname{spec}(P_{\Sigma,\mathcal{G}})$. By Lemma 1, there are exactly ℓ such functions, so this implies that as long as $m_a(i\omega) = m_g(i\omega)$, all corresponding solutions of $\boldsymbol{A}_{\Sigma,\mathcal{G}}\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = \boldsymbol{0}$ have the form (6.12) with $j = 0$. Now we proceed as in the proof of Theorem 5.2 and easily show that this implies internal stability.

(iv) We must show that the hypothesis that $m_a(i\omega) > m_g(i\omega)$ implies that there is at least one solution of the form (6.12) with $j > 0$. However, we argued in the proof of part (iii) that the number of solutions of the form (6.12) with $j = 0$ is given exactly by $m_g(i\omega)$. Therefore, if $m_a(i\omega) > m_g(i\omega)$ there must be at least one solution of the form (6.12) with $j > 0$. From this, internal instability follows. ■

6.34 Remarks

1. Thus, just as with SISO linear systems, internal stability of interconnected systems is a matter checked by computing roots of a polynomial, and possibly checking the dimension of matrix kernels.
2. Note that Theorem 6.33 holds for arbitrary systems of the form $\boldsymbol{P}\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = \boldsymbol{0}$ where $\boldsymbol{P} \in \mathbb{R}[s]^{n\times n}$. •

Lemma 1 of the proof of Theorem 6.33 is obviously important in the study of the system $\boldsymbol{A}_{\Sigma,\mathcal{G}}\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x}(t) = \boldsymbol{0}$, as it infers the character of the set of solutions, even if it does not give an completely explicit formula for the solution as is accomplished by the matrix exponential. In

particular, if $\deg P_{\mathcal{S},\mathcal{G}} = N$ then there are N linearly independent solutions to $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$. Let us denote these solutions by

$$\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_N(t).$$

Linear independence implies that for each t the matrix

$$\boldsymbol{X}(t) = \left[\ \boldsymbol{x}_1(t) \ \middle|\ \cdots \ \middle|\ \boldsymbol{x}_N(t)\ \right]$$

has full rank in the sense that if there exists a function $\boldsymbol{c}\colon \mathbb{R} \to \mathbb{R}^N$ so that $\boldsymbol{X}(t)\boldsymbol{c}(t) = \boldsymbol{0}$ for all t, then it follows that $\boldsymbol{c}(t) = \boldsymbol{0}$ for all t.

Since the characteristic polynomial is clearly an interesting object, let us see how the computation of the characteristic polynomial goes for the systems we are working with.

6.35 Examples As always, we bypass the grotesque calculations, and simply produce the characteristic polynomial.

1. For the series interconnection of Figure 6.4 we ascertain that the characteristic polynomial is simply $P_{\mathcal{S},\mathcal{G}} = D_{21}D_{32}$.
2. For the parallel signal flow graph we have
$$P_{\mathcal{S},\mathcal{G}} = D_{21}D_{42}D_{31}D_{43}.$$
3. For the negative feedback loop of Figure 6.6 we compute
$$P_{\mathcal{S},\mathcal{G}} = D_{21}D_{54}(D_{32}D_{43}D_{54} - N_{32}N_{43}N_{24}).$$
In the case when $G_{21} = G_{54} = -G_{24} = 1$ we get
$$P_{\mathcal{S},\mathcal{G}} = D_{32}D_{43} + N_{32}N_{43}. \tag{6.13}$$
4. For the four loop interconnection of Figure 6.12 we compute
$$P_{\mathcal{S},\mathcal{G}} = D_{21}D_{54}D_{58}D_{61}\left(D_{23}D_{32}\left(D_{34}D_{43} - N_{34}N_{43}\right) - \left(D_{34}D_{43}N_{23}N_{32}\right)\right) * \left(D_{67}D_{76}\left(D_{78}D_{87} - N_{78}N_{87}\right) - \left(D_{78}D_{87}N_{67}N_{76}\right)\right)$$ •

It is important to note that the characteristic polynomial for an interconnected SISO linear system, is *not* the denominator of the transfer function, because in simplifying the transfer function, there may be cancellations that may occur between the numerator and the denominator. Let us illustrate this with a very concrete example.

6.36 Example We take the negative feedback loop of Figure 6.6 and we take as our gains the following rational functions:

$$G_{21}(s) = 1, \quad G_{32}(s) = \frac{s-1}{s+1}, \quad G_{43}(s) = \frac{1}{s-1}, \quad G_{54}(s) = 1, \quad G_{24}(s) = -1.$$

The characteristic polynomial, by (6.13), is

$$P_{\mathcal{S},\mathcal{G}}(s) = (s+1)(s-1) + (s-1)1 = s^2 + s - 2.$$

If we write the transfer function as given by (6.8) we get

$$T_{\mathcal{S},\mathcal{G}}(s) = \frac{s-1}{s^2+2-2} = \frac{1}{s+2}.$$

Note that $P_{\mathcal{S},\mathcal{G}}(s) \neq s+2$! •

Conditions for BIBO and IBIBO stability

Next we wish to parlay our understanding of the character of solutions to the equations of motion gained in the previous section into conditions on IBIBO stability. The following preliminary result relates $P_{\mathcal{S},\mathcal{G}}$ and $\Delta_{\mathcal{S},\mathcal{G}}$.

6.37 Proposition *Let $(\mathcal{S},\mathcal{G})$ be an interconnected SISO linear system with interconnections $\mathcal{I}$, and for $(i,j)\in\mathcal{I}$, let (N_{ij},D_{ij}) be the c.f.r. for the gain $G_{ij}\in\mathcal{G}$. Then*

$$P_{\mathcal{S},\mathcal{G}}=\Delta_{\mathcal{S},\mathcal{G}}\prod_{(i,j)\in\mathcal{I}}D_{ij}.$$

Proof This follows from the manner in which we arrived at $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}$ from $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$. Let us systematise this in a way that makes the proof easy. Let us order $\mathcal{I}$ by ordering $\boldsymbol{n}\times\boldsymbol{n}$ as follows:

$$(1,1),(1,2),\dots,(1,n),(2,1),(2,2),\dots,(2,n),\dots,(n,1),(n,2),\dots,(n,n).$$

Thus we order $\boldsymbol{n}\times\boldsymbol{n}$ first ordering the rows, then ordering by column if the rows are equal. This then gives a corresponding ordering of $\mathcal{I}\subset\boldsymbol{n}\times\boldsymbol{n}$, and let us denote the elements of $\mathcal{I}$ by $(i_1,j_1),\dots,(i_\ell,j_\ell)$ in order. Now, define $\boldsymbol{A}_0=\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ and inductively define $\boldsymbol{A}_k$, $k\in\{1,\dots,\ell\}$, by multiplying the i_kth row of $\boldsymbol{A}_{k-1}$ by D_{i_k,j_k}. Thus, in particular $\boldsymbol{A}_{\mathcal{S},\mathcal{G}}=\boldsymbol{A}_\ell$. Now note that by the properties of the determinant,

$$\det\boldsymbol{A}_k=D_{i_k,j_k}\det\boldsymbol{A}_{k-1}.$$

Thus we have

$$\begin{aligned}\det\boldsymbol{A}_1&=D_{i_1,j_1}\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}\\ \det\boldsymbol{A}_2&=D_{i_2,j_2}\det\boldsymbol{A}_1=D_{i_2,j_2}D_{i_1,j_1}\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}\\ &\vdots\\ \det\boldsymbol{A}_\ell&=D_{i_\ell,j_\ell}\det\boldsymbol{A}_{\ell-1}=\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}\prod_{k=1}^{\ell}D_{i_k,j_k}.\end{aligned}$$

The result now follows since $P_{\mathcal{S},\mathcal{G}}=\det\boldsymbol{A}_{\mathcal{S},\mathcal{G}}$ and $\Delta_{\mathcal{S},\mathcal{G}}=\det\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$. ■

Thus we see a strong relationship between the characteristic polynomial and the determinant. Since Theorem 6.33 tells us that the characteristic polynomial has much to do with stability, we expect that the determinant will have *something* to do with stability. This observation forms the basis of the Nyquist criterion of Chapter 7. In the following result, we clearly state how the determinant relates to matters of stability. The following theorem was stated for a simple feedback structure, but in the MIMO context, by Desoer and Chan [1975]. In the SISO context we are employing, the theorem is stated, but strangely not proved, by Wang, Lee, and He [1999].

6.38 Theorem *Let $(\mathcal{S},\mathcal{G})$ be a proper, well-posed interconnected SISO linear system with interconnections $\mathcal{I}$. The following statements are equivalent:*

(i) *$(\mathcal{S},\mathcal{G})$ is internally asymptotically stable;*

(ii) *$(\mathcal{S},\mathcal{G})$ is IBIBO stable;*

(iii) *the characteristic polynomial $P_{\mathcal{S},\mathcal{G}}$ is Hurwitz;*

(iv) *the following three statements hold:*

(a) *$\Delta_{\mathcal{S},\mathcal{G}}$ has all of its zeros in $\mathbb{C}_-$;*

(b) *there are no cancellations of poles and zeros in $\overline{\mathbb{C}}_+$ in the formation of the individual loop gains;*

(c) *for any path P connecting the input of $(\mathcal{S},\mathcal{G})$ to the output there are no cancellations of poles and zeros in $\overline{\mathbb{C}}_+$ in the formation of the gain G_P.*

Furthermore, each of the above four statements implies the following statement:

(v) *$(\mathcal{S},\mathcal{G})$ is BIBO stable.*

Proof Theorem 6.33 establishes the equivalence of (i) and (iii). Let us establish the equivalence of (ii) with parts (i) and (iii).

(iii) $\implies$ (ii) We must show that all graph transmittances are BIBO stable transfer functions. Let $\boldsymbol{T}_{\mathcal{S},\mathcal{G}}$ be the matrix of graph transmittances as in Corollary 6.12. From Corollary 6.12 and Remark 6.23 we know that

$$\boldsymbol{T}_{\mathcal{S},\mathcal{G}} = \boldsymbol{G}_{\mathcal{S},\mathcal{G}}^{-1} = \boldsymbol{A}_{\mathcal{S},\mathcal{G}}^{-1}\boldsymbol{B}_{\mathcal{S},\mathcal{G}}.$$

In particular, it follows that for each $(i,j) \in \mathcal{I}$ we have

$$T_{ji} = \frac{Q_{ji}}{P_{\mathcal{S},\mathcal{G}}}$$

for some $Q_{ji} \in \mathbb{R}[s]$. From this it follows that if $P_{\mathcal{S},\mathcal{G}}$ is Hurwitz then $(\mathcal{S},\mathcal{G})$ is IBIBO stable.

(ii) $\implies$ (i) From Corollary 6.12 and Remark 6.23 it follows that each of the graph transmittances can be written as

$$T_{ji} = \frac{Q_{ji}}{P_{\mathcal{S},\mathcal{G}}}$$

for some $Q_{ij} \in \mathbb{R}[s]$. We claim that there is at least one $(i,j) \in \mathcal{I}$ so that the polynomials Q_{ij} and $P_{\mathcal{S},\mathcal{G}}$ are coprime.

finish Now we use Proposition 6.37 to show that parts (iii) and (iv) are equivalent.

Finally, (v) follows from (ii), by definition of IBIBO stability. ■

6.39 Remark Note that BIBO stability of $(\mathcal{S},\mathcal{G})$ obviously does not necessarily imply IBIBO stability (cf. Example 6.30). This is analogous to SISO linear systems where internal stability is not implied by BIBO stability. This is a property of possible pole/zero cancellations when forming the transfer function $T_{\mathcal{S},\mathcal{G}}$. For SISO linear systems, we saw that this was related to controllability and observability. This then raises the question of whether one can talk intelligibly about controllability and observability for interconnected SISO linear systems. One can, but we will not do so, referring instead to [Polderman and Willems 1998]. •

If one determines the characteristic polynomial by computing the input/output transfer function, then simplifying, one must not cancel factors in the numerator and denominator. Let us recall why this is so.

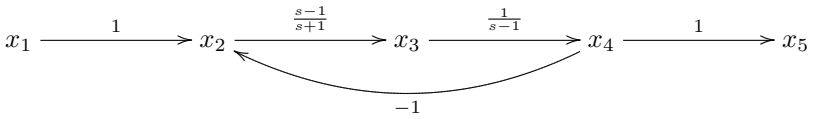


Figure 6.19 Be careful computing the characteristic polynomial

6.40 Example **(Example 6.36 cont'd)** We were looking at the negative feedback system depicted in Figure 6.19. Here the transfer function was determined to be

$$T_{\mathcal{S},\mathcal{G}}(s) = \frac{1}{s+2}$$

while the characteristic polynomial is $P_{\mathcal{S},\mathcal{G}}(s) = s^2 + s - 2 = (s+2)(s-1)$. Thus while the denominator of the transfer function has all roots in $\mathbb{C}_-$, the characteristic polynomial does not. This is also illustrated in this case by the conditions (iv a), (iv b) and (iv c) of Theorem 6.38. Thus we compute

$$\Delta_{\mathcal{S},\mathcal{G}} = 1 + \frac{1}{s+1} = \frac{s+2}{s+1}$$

and so condition (iv a) is satisfied. However, condition (iv b) is clearly not satisfied, and all three conditions must be met for IBIBO stability. •

6.3 Feedback for input/output systems with a single feedback loop

Although in the previous two sections we introduced a systematic way to deal with very general system interconnections, SISO control typically deals with the simple case where we have an interconnection with one loop. In this section we concentrate on this setting, and provide some details about such interconnections. We first look at the typical single loop control problem with a plant transfer function that is to be modified by a controller transfer function. In particular, we say what we mean by open-loop and closed-loop control. Then, in Section 6.3.2 we simplify things and look at a generic single loop configuration, identifying in it the features on which we will be concentrating for a large part of the remainder of these notes.

6.3.1 Open-loop versus closed-loop control

We shall mainly be interested in considering feedback as a means of designing a controller to accomplish a desired task. Thus we start with a rational function $R_P \in \mathbb{R}(s)$ that describes the ***plant*** (i.e., the system about whose output we care) and we look to design a controller $R_C \in \mathbb{R}(s)$ that stabilises the system. The plant rational function should be thought of as unchangeable. One could simply use an ***open-loop*** controller and design R_C so that the ***open-loop transfer function*** $R_P R_C$ has the desired properties. This corresponds to the situation of Figure 6.20.[1] However, as we saw in Section 1.2, there are serious drawbacks to this methodology. To get around these we design the controller to act not on the ***reference signal*** $\hat{r}$, but on the ***error signal*** $\hat{e} = \hat{r} - \hat{y}$. One may place the controller in other places in the block diagram, and one may have other rational functions in the block diagram.

[1] We place a • at a node in the signal flow graph that we do not care to name.

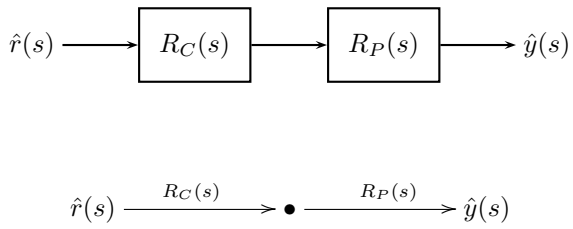


Figure 6.20 Open-loop control configuration as a block diagram (top) and a signal flow graph (bottom)

However, for such systems, the essential methodology is the same, and so let us concentrate on one type of system interconnection for the moment for simplicity. The block diagram configuration for the so-called ***closed-loop*** system we consider is depicted in Figure 6.21, and in Figure 6.22 we show some possible alternate feedback loops that we do not look at

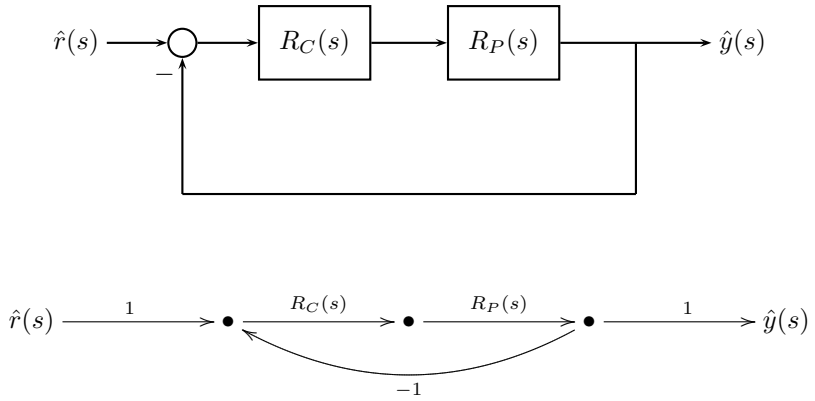


Figure 6.21 Closed-loop control configuration as a block diagram (top) and a signal flow graph (bottom)

in detail.

The ***closed-loop transfer function*** from $\hat{r}$ to $\hat{y}$ is readily computed as

$$\frac{\hat{y}}{\hat{r}} = \frac{R_P R_C}{1 + R_P R_C}.$$

The objective in the input/output scenario is described by the following problem statement.

6.41 Input/output control design problem Given a rational function R_P describing the plant, find a rational function R_C describing the controller, that make the closed-loop transfer function behave in a suitable manner. In particular, one will typically wish for the poles of the closed-loop transfer function to be in $\mathbb{C}_-$. •

When doing this, the concerns that we will raise in Section 6.2.3 need to be taken into account. What's more, there are other concerns one needs to be aware of, and some of these are addressed in Chapter 8. That is, one cannot look at the transfer function $\frac{\hat{y}}{\hat{r}}$ as being the only indicator of system performance.

Let us look at an example of designing a closed-loop control law for a given plant.

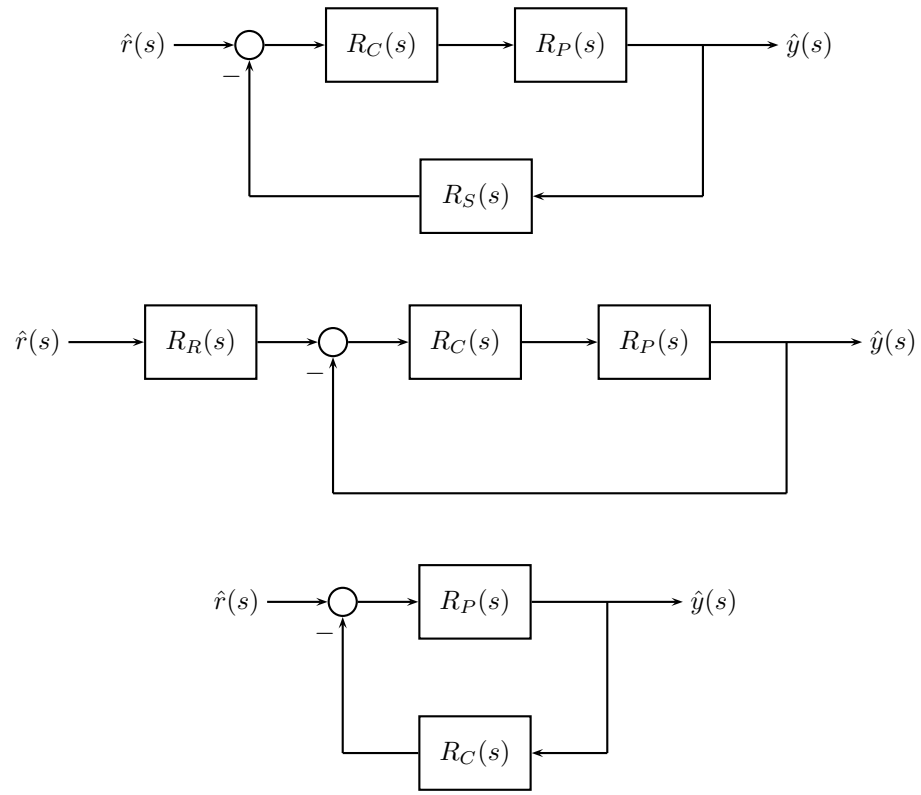


Figure 6.22 Some alternate feedback loops: transfer function in feedback path (top); modiifed reference signal (middle); controller in feedback path (bottom)

6.42 Example We consider the plant transfer function $R_P(s) = \frac{1}{s}$. Suppose we give the system a step input: $u(t) = 1(t)$. Then we have $\hat{u}(s) = \frac{1}{s}$, and so the output in the Laplace transform domain will be $\hat{y}(s) = \frac{1}{s^2}$. From this we determine that $y(t) = t$. Thus the output blows up as $t \to \infty$.

Let's try to repair this with an open-loop controller. We seek a plant rational function R_C so that $R_C R_P$ has all poles in $\mathbb{C}_-$. If (N_C, D_C) is the c.f.r. of R_C, we have

$$R_C(s)R_P(s) = \frac{N_C(s)}{sD_C(s)}.$$

Thus the partial fraction expansion of $R_C R_P$ will always contain a term like $\frac{a}{s}$ for some $a \in \mathbb{R}$ unless we cancel the denominator of the plant transfer function with the numerator of the controller transfer function. However, this is a bad idea. It essentially corresponds to introducing unobservable dynamics into the system. So this leaves us with the term $\frac{a}{s}$ in the partial fraction expansion, and with a step response, the output will still blow up as $t \to \infty$.

This motivates our trying a closed-loop scheme like that in Figure 6.21. We take $R_C(s) = \frac{1}{s+1}$ so that our closed-loop system is as depicted in Figure 6.23. The closed-loop transfer function in this case is readily computed to be

$$\frac{R_C(s)R_P(s)}{1 + R_C(s)R_P(s)} = \frac{\frac{1}{s+1}\frac{1}{s}}{1 + \frac{1}{s+1}\frac{1}{s}} = \frac{1}{s^2 + s + 1}.$$

The poles of the closed-loop transfer function are now $-\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$, which are both in $\mathbb{C}_-$. The

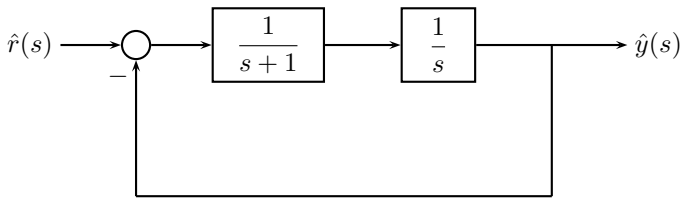


Figure 6.23 A closed-loop scheme for a simple plant

step response in the Laplace transform domain is

$$\hat{y}(s) = \frac{R_C(s)R_P(s)}{1 + R_C(s)R_P(s)}\hat{u}(s) = \frac{1}{s^2+s+1}\frac{1}{s}.$$

The inverse Laplace transform can be computed using partial fraction expansion. We have

$$\frac{1}{s^2+s+1}\frac{1}{s} = \frac{1}{s} - \frac{s}{s^2+s+1} - \frac{1}{s^2+s+1},$$

from which we can use our formulas of Section E.3 to ascertain that

$$y(t) = 1 - e^{t/2}\big(\cos\tfrac{\sqrt{3}}{2}t + \tfrac{1}{\sqrt{3}}\sin\tfrac{\sqrt{3}}{2}t\big).$$

We plot this response in Figure 6.24, Note that the closed-loop step response is now

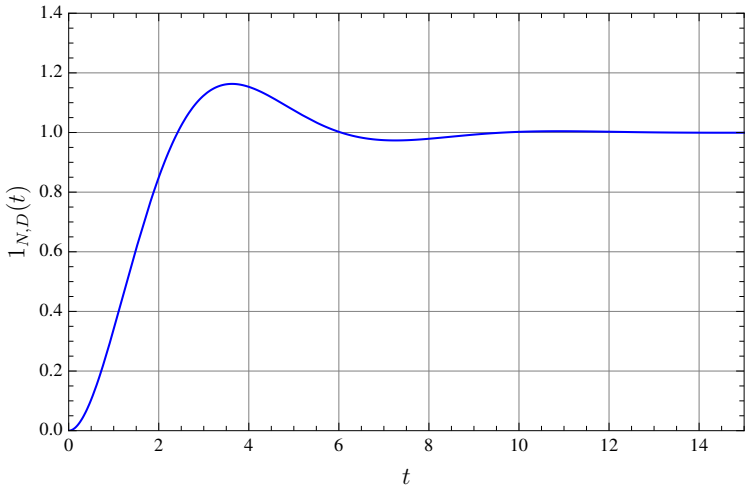


Figure 6.24 Closed-loop step response for simple plant and controller

bounded—something we were not able to legitimately accomplish with the open-loop controller. We shall develop ways of designing controllers for such systems later in the course. •

6.3.2 Unity gain feedback loops

In this section we focus on the feedback loop depicted in Figure 6.25. While in the previous section we had looked at the case where $R_L = R_C R_P$, in this section our interests

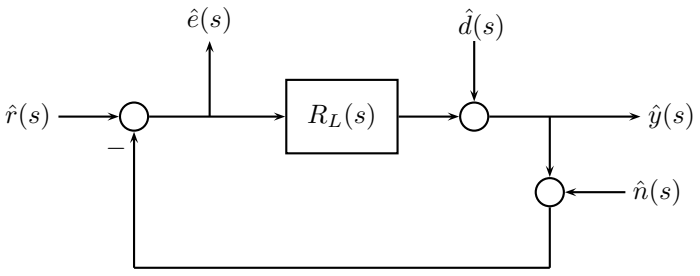


Figure 6.25 A unity gain feedback loop

are more in the structure of the block diagram than in the desire to design a controller rational function R_C. In this case R_L is often called the ***open-loop transfer function*** for the interconnection, meaning that it is the transfer function if the feedback connection is snipped. It is the relationship between the open-loop transfer function and the closed-loop transfer function that lies at the heart of classical control design. Although it is true that by restricting to such an interconnection we loose some generality, it is not difficult to adapt what we say here to more general single-loop configurations, perhaps with transfer functions in the feedback loop, or a transfer function between the reference and the input to the loop.

The signals in the block diagram Figure 6.25 are the ***reference*** $r(t)$, the ***output*** $y(t)$, the ***error*** $e(t)$, the ***disturbance*** $d(t)$, and the ***noise*** $n(t)$. Throughout this chapter we will encounter the various transfer functions associated with the block diagram Figure 6.25, so let us record them here so that we may freely refer to them in the sequel:

$$\frac{\hat{y}}{\hat{r}} = \frac{R_L}{1+R_L}, \quad \frac{\hat{y}}{\hat{d}} = \frac{1}{1+R_L}, \quad \frac{\hat{y}}{\hat{n}} = \frac{R_L}{1+R_L},$$
$$\frac{\hat{e}}{\hat{r}} = \frac{1}{1+R_L}, \quad \frac{\hat{e}}{\hat{d}} = \frac{1}{1+R_L}, \quad \frac{\hat{e}}{\hat{n}} = \frac{1}{1+R_L}.$$

We see that there are essentially two transfer functions involved here:

$$T_L = \frac{R_L}{1+R_L} \quad \text{and} \quad S_L = \frac{1}{1+R_L}.$$

These transfer functions are given the name of ***complementary sensitivity function***, and ***sensitivity function***, respectively. Note that

$$T_L + S_L = 1.$$

Of course, the complementary sensitivity function is simply the closed-loop transfer function from the input to the output. The sensitivity function is, in the parlance of Section 6.1.6, the sensitivity of T_L to R_L. In Chapters 8 and 9 we will be seeing the importance of each of these two transfer functions, and we will get a look at how they can interact in the design process to make things somewhat subtle.

6.3.3 Well-posedness and stability of single-loop interconnections

When making a single-loop interconnection of the type in Figure 6.26, we have the notions of stability and well-posedness given in Sections 6.2.3 and 6.2.2. Let us examine these ideas in our simple single-loop context.

We first show how to easily characterise well-posedness for single-loop interconnections.

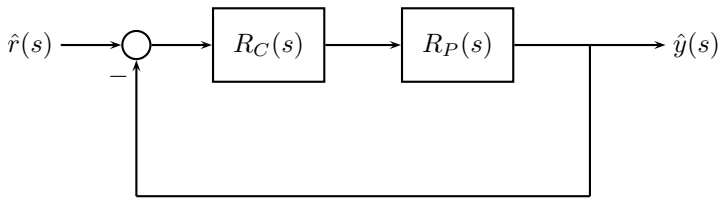


Figure 6.26 Block diagram single-loop feedback

6.43 Proposition *Let $R_L = R_C R_P \in \mathbb{R}(s)$ be proper and consider the interconnection of Figure 6.26. Let $\Sigma_L = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be the canonical minimal realisation for R_L. The following statements are equivalent:*

(i) the interconnection is well-posed;

(ii) $\lim_{s\to\infty} R_L(s) \neq -1$;

(iii) $\boldsymbol{D} \neq [-1]$.

Proof (i) $\implies$ (ii) Suppose that (ii) does not hold so that $\lim_{s\to\infty} R_L(s) = -1$. Then we have

$$\lim_{s\to\infty} T_L(s) = \lim_{s\to\infty} \frac{R_L(s)}{1 + R_L(s)} = \infty,$$

and similarly $\lim_{s\to\infty} S_L(s) = \infty$. This implies that both T_L and S_L must be improper, so the interconnection cannot be well-posed.

(ii) $\implies$ (iii) Recall from the proof of Theorem 3.20 that since Σ_L is the canonical minimal realisation of R_L we have

$$R_L(s) = T_{\Sigma_L}(s) = \frac{\tilde{c}_n s^n + \tilde{c}_{n-1} s^{n-1} + \cdots + \tilde{c}_1 s + \tilde{c}_0}{s^n + p_{n-1} s^{n-1} + \cdots + p_1 s + p_0}$$

where $\boldsymbol{D} = [\tilde{c}_n]$ and $\tilde{c}_i = c_i + \tilde{c}_n p_i$, and where $p_0, \ldots, p_{n-1}$ are the coefficients in the characteristic polynomial for $\boldsymbol{A}$. One now computes $\lim_{s\to\infty} T_{\Sigma_L}(s) = \tilde{c}_n$. Thus $\lim_{s\to\infty} R_L(s) \neq -1$ implies that $\boldsymbol{D} \neq [1]$, as desired.

(iii) $\implies$ (i) From the previous step in the proof, if $\boldsymbol{D} = [d]$ then $\lim_{s\to\infty} R_L(s) = d$. Therefore,

$$\lim_{s\to\infty} T_L(s) = \lim_{s\to\infty} \frac{R_L(s)}{1 + R_L(s)} = \frac{d}{1 + d}$$

is finite if $d \neq -1$. Also, if $d \neq -1$ then

$$\lim_{s\to\infty} S_L(s) = \frac{1}{1 + d}.$$

Thus indicates that if $d \neq -1$ then T_L and S_L are both proper. ■

The following obvious corollary indicates that well-posedness is not a consideration for strictly proper loop gains.

6.44 Corollary *If the loop gain R_L in Figure 6.25 is strictly proper, then the interconnection is well-posed.*

Proof This follows from Proposition 6.43 since if R_L is strictly proper than $\lim_{s\to\infty} R_L(s) = 0$. ■

Next let us turn to stability of single-loop interconnections. It will be convenient to introduce some notation. Given a plant R_P, let us denote by $\mathscr{S}(R_P)$ the collection of IBIBO stabilising controllers for which the closed-loop system is well-posed. Thus we consider the interconnection shown Figure 6.26 and we take

$$\mathscr{S}(R_P) = \{R_C \in \mathbb{R}(s) \text{ the interconnection of Figure 6.26 is IBIBO stable}\}.$$

Of course, one can characterise the set of stabilising controllers fairly concretely using the general machinery of Section 6.2.3. One readily sees that the following result follows directly from Theorem 6.38 (the reader can provide a direct proof of this in Exercise E6.7).

6.45 Proposition *Let $R_P \in \mathbb{R}(s)$ be proper. For $R_C \in \mathbb{R}(s)$ the following statements are equivalent:*

(i) $R_C \in \mathscr{S}(R_P)$;

(ii) *the following two statements hold:*

  (a) *the characteristic polynomial $D_C D_P + N_C N_P$ is Hurwitz;*

  (b) $\lim_{s\to\infty} R_C(s)R_P(s) \neq -1$;

(iii) *the following three statements hold:*

  (a) *R_C and R_P have no pole/zero cancellations in $\overline{\mathbb{C}}_+$;*

  (b) *$1 + R_C R_P$ has no zeros in $\overline{\mathbb{C}}_+$;*

  (c) $\lim_{s\to\infty} R_C(s)R_P(s) \neq -1$.

Here (N_C, D_C) and (N_P, D_P) are the c.f.r.'s of R_C and R_P, respectively.

6.46 Remark While Proposition 6.45 characterises the stabilising controllers, it does not answer the question concerning their existence, and if they exist, how many there are. We shall deal with this in subsequent parts of the text, but since the questions are so fundamental, let us now point to where the answers can be found.

1. For any strictly proper plant R_P of order n, there exists a strictly proper controller $R_C \in \mathscr{S}(R_P)$ of order n (Theorems 10.27 and 13.2) and a proper controller $R_C \in \mathscr{S}(R_P)$ of order $n-1$ (Theorem 13.2).
2. For any proper plant R_P of order n, there exists a strictly proper controller $R_C \in \mathscr{S}(R_P)$ of order n (Theorems 10.27 and 13.2).
3. For a proper plant R_P, there is (essentially) a bijection from the set $\mathscr{S}(R_P)$ to RH_∞^+ (Theorem 10.37). •

6.4 Feedback for SISO linear systems

We look at our state-space model

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t), \end{aligned} \tag{6.14}$$

and ask, "What should feedback mean for such a system?" One can, of course, write the transfer function T_Σ as a quotient of coprime polynomials, and then proceed like we did above in the input/output case. However, it is not immediately clear what this means in the framework of the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$. Indeed, it is not even clear what types of transfer functions ought to be achievable via a SISO linear system Σ because the inputs for such a system appear in a specific way.

What needs to be undertaken is a description of feedback for SISO linear systems, independent of those in input/output form. The idea is that for the system (6.14) we should take as feedback a linear function of the state $\boldsymbol{x}$ and the output y. We first consider the case of pure state feedback, then allow the output to be fed back.

6.4.1 Static state feedback for SISO linear systems

State feedback should be of the form $u(t) = r(t) - \boldsymbol{f}^t\boldsymbol{x}(t)$ for some $\boldsymbol{f} \in \mathbb{R}^n$, where r is the reference signal. In block diagram form, the situation is illustrated in Figure 6.27. One

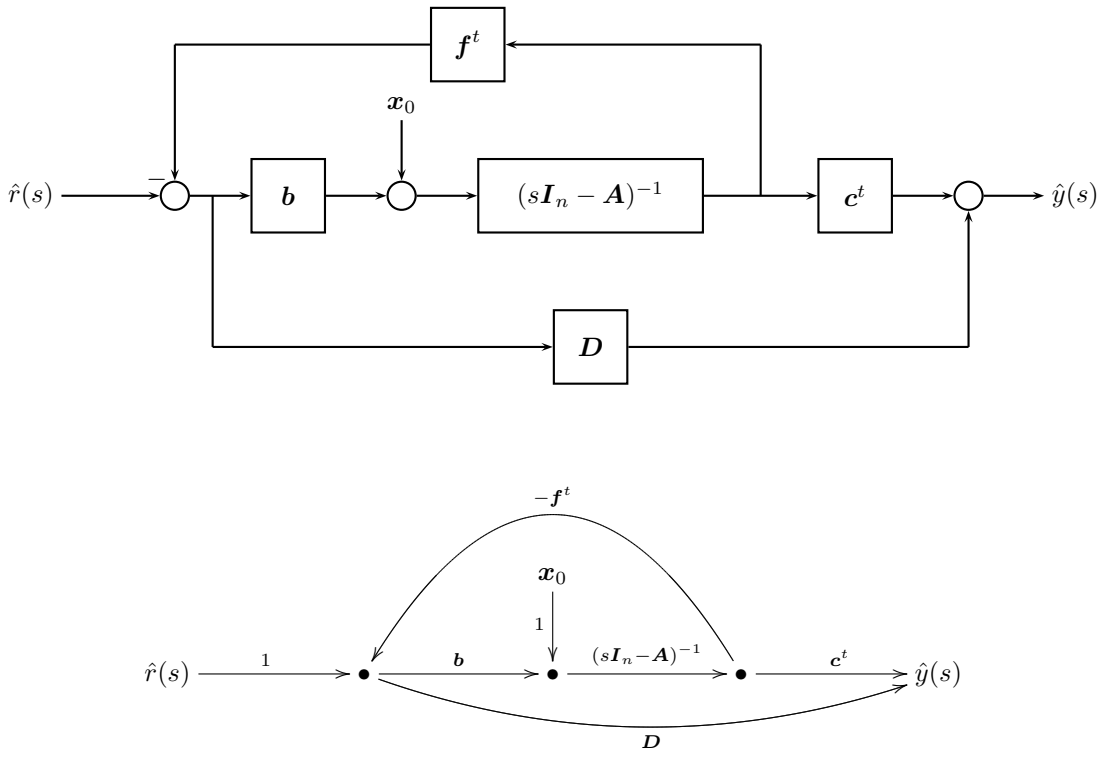


Figure 6.27 The static state feedback configuration for the SISO linear system (6.14) as a block diagram (top) and a signal flow graph (bottom)

readily ascertains that the closed-loop equations are

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)\boldsymbol{x}(t) + \boldsymbol{b}r(t)\\ y(t) &= (\boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t)\boldsymbol{x}(t) + \boldsymbol{D}r(t).\end{aligned}$$

Motivated by this, we have the following definition.

6.47 Definition Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system. A ***state feedback vector*** is a vector $\boldsymbol{f} \in \mathbb{R}^n$, and to a state feedback vector $\boldsymbol{f}$ we assign the ***closed-loop*** SISO linear system $\Sigma_{\boldsymbol{f}} = (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t, \boldsymbol{b}, \boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t, \boldsymbol{D})$. The transfer function for $\Sigma_{\boldsymbol{f}}$ is called the ***closed-loop transfer function***. A rational function $R \in \mathbb{R}(s)$ is ***state compatible*** with Σ is there exists a state feedback vector $\boldsymbol{f}$ with the property that $T_{\Sigma_{\boldsymbol{f}}} = R$. •

The control problem here is a bit different than that of the input/output problem.

6.48 Static state feedback design problem Given the system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, find a state feedback vector $\boldsymbol{f}$ so that

(i) the closed-loop transfer function $\Sigma_{\boldsymbol{f}}$ has desirable properties and

(ii) the state variables are behaving in a nice fashion.

In particular, one will typically want the matrix $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ to be Hurwitz. •

By recognising that we have states, we are forced to confront the issue of their behaviour, along with the behaviour of the input/output system. Following our notation for stabilising controller transfer functions for input/output systems, given $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, let us denote by $\mathscr{S}_{\mathrm{s}}(\Sigma)$ the set of stabilising state feedback vectors. Thus

$$\mathscr{S}_{\mathrm{s}}(\Sigma) = \{\boldsymbol{f} \in \mathbb{R}^n \mid \boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t \text{ is Hurwitz}\}.$$

The subscript "s" means state, as we shall shortly look at output feedback as well.

The following result says, at least, what we can do with the input/output relation. Note that the result also says that for a controllable system, the eigenvalues of the closed-loop system can be arbitrarily placed.

6.49 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system with $(\boldsymbol{A}, \boldsymbol{b})$ controllable. A rational function $R \in \mathbb{R}(s)$ is state compatible with Σ if and only if there exists a monic polynomial $P \in \mathbb{R}[s]$ of degree n so that*

$$R(s) = \frac{\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} + dP_{\boldsymbol{A}}(s)}{P(s)},$$

where $\boldsymbol{D} = [d]$, and where $P_{\boldsymbol{A}}$ is the characteristic polynomial for $\boldsymbol{A}$. In particular, if $(\boldsymbol{A}, \boldsymbol{b})$ is controllable, then $\mathscr{S}_{\mathrm{s}}(\Sigma) \neq \emptyset$.

Proof The closed-loop transfer function for is

$$T_{\Sigma_{\boldsymbol{f}}}(s) = \frac{(\boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t)\mathrm{adj}(s\boldsymbol{I}_n - (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t))\boldsymbol{b}}{\det(s\boldsymbol{I}_n - (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t))} + \boldsymbol{D}.$$

Since $(\boldsymbol{A}, \boldsymbol{b})$ is controllable, we may as well suppose that

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

and let us write $\boldsymbol{c} = (c_0, c_1, \ldots, c_{n-1})$. If $\boldsymbol{f} = (f_0, f_1, \ldots, f_{n-1})$ a simple calculation gives

$$\boldsymbol{b}\boldsymbol{f}^t = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_0 & f_1 & f_2 & \cdots & f_{n-1} \end{bmatrix}$$

and so

$$\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t=\begin{bmatrix}0&1&0&0&\cdots&0\\0&0&1&0&\cdots&0\\0&0&0&1&\cdots&0\\\vdots&\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&0&\cdots&1\\-p_0-f_0&-p_1-f_1&-p_2-f_2&-p_3-f_3&\cdots&-p_{n-1}-f_{n-1}\end{bmatrix}.$$

This shows that by choosing $\boldsymbol{f}$ appropriately, we may make the characteristic polynomial of $\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t$ anything we like.

Now we note that

$$(\boldsymbol{c}^t-\boldsymbol{D}\boldsymbol{f}^t)\mathrm{adj}(s\boldsymbol{I}_n-\boldsymbol{A})\boldsymbol{b}=(\boldsymbol{c}^t-\boldsymbol{D}\boldsymbol{f}^t)\mathrm{adj}(s\boldsymbol{I}_n-(\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t))\boldsymbol{b}$$

for *any* vector $\boldsymbol{f}^t$. This is because both $\boldsymbol{A}$ and $\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t$ are in controller canonical form, which means that the polynomials $(\boldsymbol{c}^t-\boldsymbol{D}\boldsymbol{f}^t)\mathrm{adj}(s\boldsymbol{I}_n-\boldsymbol{A})\boldsymbol{b}$ and $(\boldsymbol{c}^t-\boldsymbol{D}\boldsymbol{f}^t)\mathrm{adj}(s\boldsymbol{I}_n-(\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t))\boldsymbol{b}$ are both given by

$$(c_{n-1}-df_{n-1})s^{n-1}+\cdots+(c_1-df_1)s+(c_0-df_0).$$

if $\boldsymbol{c}=(c_0,c_1,\dots,c_{n-1})$ and $\boldsymbol{D}=[d]$. Now we observe that if $P(s)=\det(s\boldsymbol{I}_n-(\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t))$ then we have

$$f_{n-1}s^{n-1}+\cdots+f_1s+f_0=P(s)-P_{\boldsymbol{A}}(s).$$

Therefore

$$(\boldsymbol{c}^t-\boldsymbol{D}\boldsymbol{f}^t)\mathrm{adj}(s\boldsymbol{I}_n-(\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t))\boldsymbol{b}=\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n-\boldsymbol{A})\boldsymbol{b}-d\big(P(s)-P_{\boldsymbol{A}}(s)\big).$$

The theorem now follows by straightforward simplification. ■

This result is important because it demonstrates that by choosing the appropriate static state feedback for a controllable system Σ, we may do as we please with the poles of the closed-loop transfer function. And, as we have seen in Proposition 3.24 and Corollary 5.7, the poles of the transfer function have a great deal of effect on the behaviour of the system.

Let us do this in an *ad hoc* way in an example.

6.50 Example We take

$$\boldsymbol{A}=\begin{bmatrix}0&1\\-1&0\end{bmatrix},\quad \boldsymbol{b}=\begin{bmatrix}0\\1\end{bmatrix}.$$

Note that $\boldsymbol{A}$ is not Hurwitz as it has characteristic polynomial s^2+1. Without any justification (for this, refer ahead to Example 10.15) we take as state feedback vector $\boldsymbol{f}=(3,4)$. We then have

$$\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t=\begin{bmatrix}0&1\\-1&0\end{bmatrix}-\begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}3&4\end{bmatrix}=\begin{bmatrix}0&1\\-4&-4\end{bmatrix}.$$

Since $(\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t,\boldsymbol{b})$ is in controller canonical form, the characteristic polynomial can be read from the bottom row: s^2+4s+4.

Let's look at the behaviour of the open-loop system. We compute

$$e^{\boldsymbol{A}t}=\begin{bmatrix}\cos t&\sin t\\-\sin t&\cos t\end{bmatrix}.$$

If we provide the periodic input $u(t) = 1(t)\cos t$ and zero initial condition, the time-response of the state of the system is

$$\boldsymbol{x}(t) = \begin{bmatrix} \frac{1}{2}t\sin t \\ \frac{1}{2}(t\cos t + \sin t) \end{bmatrix}$$

which we plot in Figure 6.28. Taking $\boldsymbol{c} = (1, 0)$, the corresponding output is

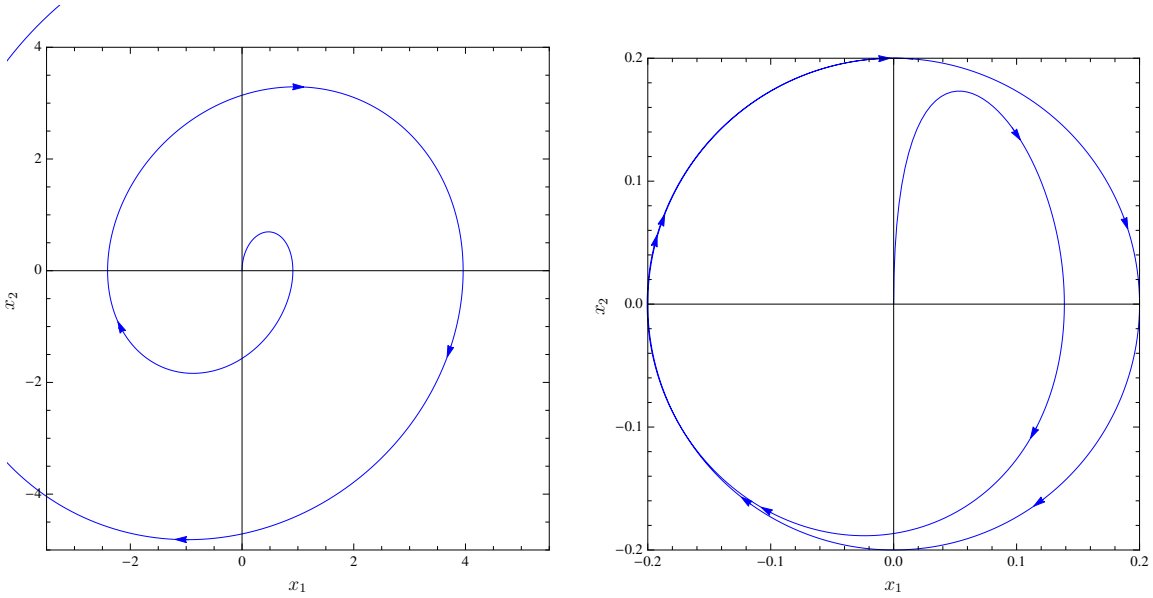


Figure 6.28 State response for open-loop system (left) and closed-loop system (right) under static state feedback

$$y(t) = \tfrac{1}{2}t\sin t$$

which we plot in Figure 6.29.

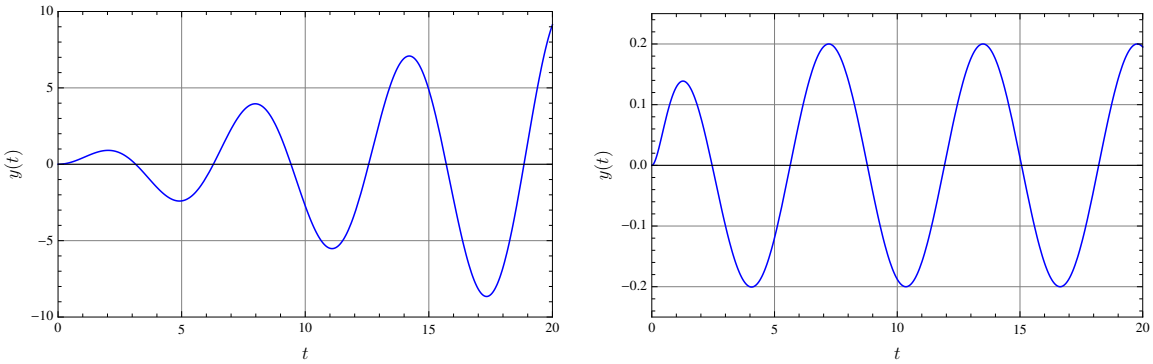


Figure 6.29 Output response for open-loop system (left) and closed-loop system (right) under static state feedback

For the closed-loop system we compute

$$e^{(\boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t)t} = \begin{bmatrix} e^{-2t} + 2te^{-2t} & te^{-2t} \\ -4te^{-2t} & e^{-2t} - 2te^{-2t} \end{bmatrix}$$

from which we ascertain the state behaviour to be

$$\boldsymbol{x}(t) = \begin{bmatrix} -\frac{3}{25}e-2t - \frac{2}{5}te^{-2t} + \frac{3}{25}\cos t + \frac{4}{25}\cos t \\ -\frac{4}{25}e^{-2t} - \frac{4}{5}te^{-2t} + \frac{4}{20}\cos t - \frac{3}{25}\sin t \end{bmatrix}$$

and the output to be

$$y(t) = -\frac{3}{25}e-2t - \frac{2}{5}te^{-2t} + \frac{3}{25}\cos t + \frac{4}{25}\cos t.$$

These are shown beside the open-loop response in Figures 6.28 and 6.29, respectively.

As expected, the addition of static state feedback has caused the system to behave in a more suitable manner. In fact, for this example, it has taken an BIBO unstable system and made it BIBO stable. •

The matter of static state feedback is also attended to in Section 10.1.1—where a better understanding of when static state feedback can make a closed-loop system stable—and in Section 10.2.1—where methods of constructing such feedback laws are discussed. We also mention that those attracted to the signal flow graph technology for feedback might be interested in looking at [Reinschke 1988] where a presentation of static state and static output feedback (see the next section for the latter) appears in terms of graphs.

6.4.2 Static output feedback for SISO linear systems

Now we consider feeding back not the state, but the output itself. Thus we consider as feedback for the system the quantity $u(t) = r(t) - Fy(t)$ for $F \in \mathbb{R}$, and where r is the reference signal. This is illustrated diagrammatically in Figure 6.30.

Using the equations (6.14) we may determine the closed-loop equations. It turns out that we require that $F\boldsymbol{D} \neq -1$, a condition that is true, for instance, when $\boldsymbol{D} = \mathbf{0}_1$. This condition is related to the well-posedness condition for input/output systems discussed in Section 6.3.3. In any event, when the conputations are carried out we get

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \Big(\boldsymbol{A} - \frac{F}{1+F\boldsymbol{D}}\boldsymbol{b}\boldsymbol{c}^t\Big)\boldsymbol{x}(t) - \Big(1 - \frac{F\boldsymbol{D}}{1+F\boldsymbol{D}}\Big)\boldsymbol{b}r(t) \\ y(t) &= (1+F\boldsymbol{D})^{-1}\boldsymbol{c}^t\boldsymbol{x}(t) + (1+F\boldsymbol{D})^{-1}\boldsymbol{D}r(t), \end{aligned}$$

for a reference signal $r(t)$. These expressions simplify somewhat in the usual case when $\boldsymbol{D} = \mathbf{0}_1$. With this in mind, we make the following definition which is the analogue of Definition 6.47.

6.51 Definition Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system. An ***output feedback constant*** is a number $F \in \mathbb{R}$ with the property that $F\boldsymbol{D} \neq -1$. To an output feedback number F we assign the ***closed-loop*** SISO linear system $\Sigma_F = (\boldsymbol{A} - \frac{F}{1+F\boldsymbol{D}}\boldsymbol{b}\boldsymbol{c}^t, \big(1 - \frac{F\boldsymbol{D}}{1+F\boldsymbol{D}}\big)\boldsymbol{b}, (1+F\boldsymbol{D})^{-1}\boldsymbol{c}^t, (1+F\boldsymbol{D})^{-1}\boldsymbol{D})$. The transfer function for Σ_F is called the ***closed-loop transfer function***. A rational function $R \in \mathbb{R}(s)$ is ***output compatible*** with Σ is there exists an output feedback number F with the property that $T_{\Sigma_F} = R$. •

6.52 Remarks Note that static output feedback is somewhat uninteresting for SISO systems. This is because the feedback parameter is simply a scalar in this case. For MIMO systems, the feedback is not via a scalar, but by a matrix, so things are more interesting in this case. Nevertheless, as we shall see in Section 10.2.2, the static output feedback problem is difficult, even for SISO systems. •

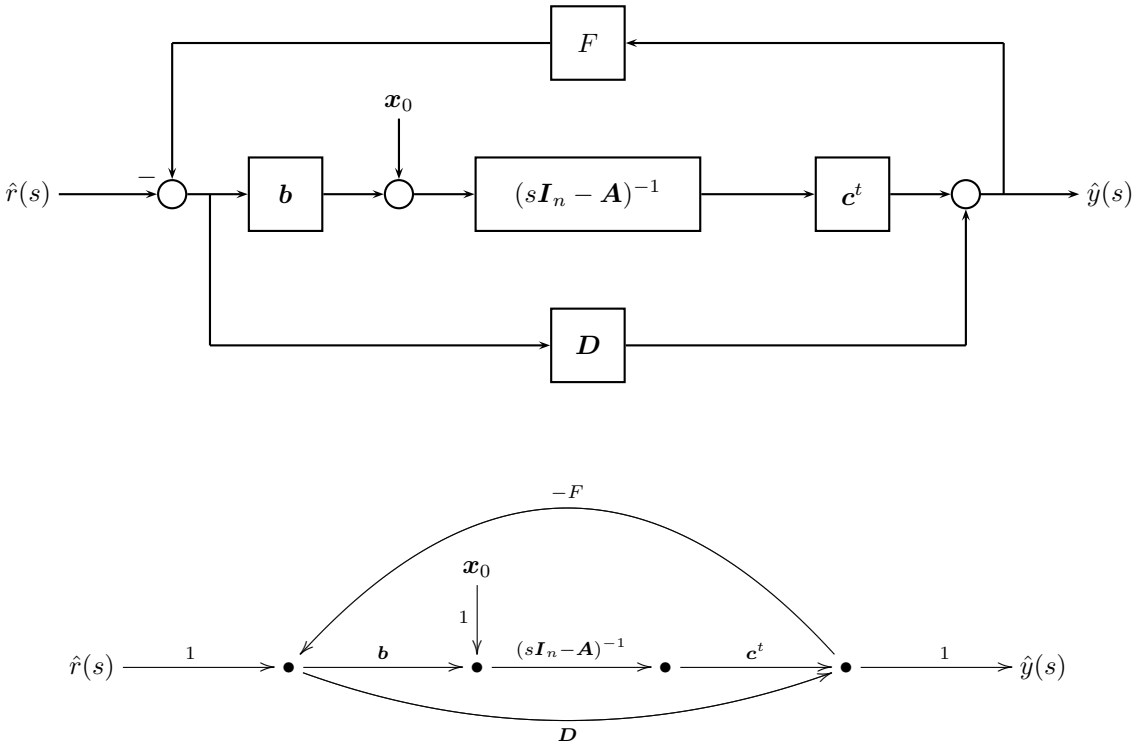


Figure 6.30 The static output feedback configuration for the SISO linear system (6.14) as a block diagram (top) and a signal flow graph (bottom)

Let us say what is the objective with this type of feedback.

6.53 Static output feedback design problem Given the system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, find an output feedback constant F so that

(i) the closed-loop transfer function Σ_F has desirable properties and

(ii) the state variables are behaving in a nice fashion.

In particular, one typically want the matrix $\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t$ to be Hurwitz. •

Following our earlier notation, given $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, we denote by $\mathscr{S}_{\text{o}}(\Sigma)$ the set of stabilising output feedback constants. That is,

$$\mathscr{S}_{\text{o}}(\Sigma) = \{F \in \mathbb{R} \mid \boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t \text{ is Hurwitz}\}.$$

Let us look at the form of rational functions compatible with a system under static output feedback. This is analogous to Theorem 6.49, although we cannot make a statement concerning the nature of the stabilising output feedback constants.

6.54 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system with $(\boldsymbol{A}, \boldsymbol{c})$ observable and $\boldsymbol{D} \neq \boldsymbol{0}_1$. A rational function $R \in \mathbb{R}(s)$ is output compatible with Σ if and only if*

$$R(s) = \frac{\big(1 - \frac{F\boldsymbol{D}}{1+F\boldsymbol{D}}\big)\boldsymbol{c}^t \text{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} + \boldsymbol{D}P(s)}{P(s)},$$

where $P(s) = P_{\boldsymbol{A}}(s) + \frac{F}{1+F\boldsymbol{D}}\boldsymbol{c}^t \text{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ and $F \in \mathbb{R}$.

Proof We without loss of generality assume that $(\boldsymbol{A}, \boldsymbol{c})$ are in observer canonical form:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -p_0 \\ 1 & 0 & 0 & \cdots & 0 & -p_1 \\ 0 & 1 & 0 & \cdots & 0 & -p_2 \\ 0 & 0 & 1 & \cdots & 0 & -p_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -p_{n-2} \\ 0 & 0 & 0 & \cdots & 1 & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{c}^t = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Let us write $\boldsymbol{b} = (b_0, b_1, \ldots, b_{n-1})$. We then have

$$\boldsymbol{b}\boldsymbol{c}^t = \begin{bmatrix} 0 & 0 & 0 & \cdots & b_0 \\ 0 & 0 & 0 & \cdots & b_1 \\ 0 & 0 & 0 & \cdots & b_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & b_{n-1} \end{bmatrix}$$

so that

$$\boldsymbol{A} - \frac{F}{1+F\boldsymbol{D}}\boldsymbol{b}\boldsymbol{c}^t = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -p_0 - \frac{F}{1+F\boldsymbol{D}}b_0 \\ 1 & 0 & 0 & \cdots & 0 & -p_1 - \frac{F}{1+F\boldsymbol{D}}b_1 \\ 0 & 1 & 0 & \cdots & 0 & -p_2 - \frac{F}{1+F\boldsymbol{D}}b_2 \\ 0 & 0 & 1 & \cdots & 0 & -p_3 - \frac{F}{1+F\boldsymbol{D}}b_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -p_{n-2} - \frac{F}{1+F\boldsymbol{D}}b_{n-2} \\ 0 & 0 & 0 & \cdots & 1 & -p_{n-1} - \frac{F}{1+F\boldsymbol{D}}b_{n-1} \end{bmatrix}.$$

Thus the characteristic polynomial of $\boldsymbol{A} - \frac{F}{1+F\boldsymbol{D}}\boldsymbol{b}\boldsymbol{c}^t$ is $P_{\boldsymbol{A}}(s) + \frac{F}{1+F\boldsymbol{D}}\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ since

$$\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} = b_{n-1}s^{n-1} + \cdots + b_1 s + b_0.$$

We also have

$$\boldsymbol{c}^t\Big(s\boldsymbol{I}_n - \Big(\boldsymbol{A} - \frac{F}{1+F\boldsymbol{D}}\boldsymbol{b}\boldsymbol{c}^t\Big)\Big)\Big(1 - \frac{F\boldsymbol{D}}{1+F\boldsymbol{D}}\Big)\boldsymbol{b} = \Big(1 - \frac{F\boldsymbol{D}}{1+F\boldsymbol{D}}\Big)\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b},$$

given that $\boldsymbol{A} - \frac{F}{1+F\boldsymbol{D}}\boldsymbol{b}\boldsymbol{c}^t$ is in observer canonical form. Therefore, the closed-loop transfer function is

$$T_{\Sigma_F}(s) = \frac{\big(1 - \frac{F\boldsymbol{D}}{1+F\boldsymbol{D}}\big)\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}}{P_{\boldsymbol{A}}(s) + \frac{F}{1+F\boldsymbol{D}}\boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}} + \boldsymbol{D},$$

and this is exactly as stated in the theorem. ■

Unsurprisingly, static output feedback for SISO systems does not give the same freedom for pole placement as does static state feedback. Nevertheless, it is possible to have a positive effect on a system's behaviour by employing static output feedback, as is indicated by the following cooked example.

6.55 Example **(Example 6.50 cont'd)** We consider the SISO system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1.$$

Thus $\boldsymbol{A}$ and $\boldsymbol{b}$ are just as in Example 6.50. Therefore, if we give the open-loop system the same input $u(t) = 1(t)\cos t$ with the same state initial condition $\boldsymbol{x}(0) = \boldsymbol{0}$, the open-loop state evolution will be the same as that in Figure 6.28. However, the output vector $\boldsymbol{c}$ we now use differs from that of Example 6.50. The open-loop output is

$$2\sin t + 2t\cos t + \tfrac{3}{2}t\sin t$$

and shown in Figure 6.31. Let us use static output feedback with $F = 1$. We then compute

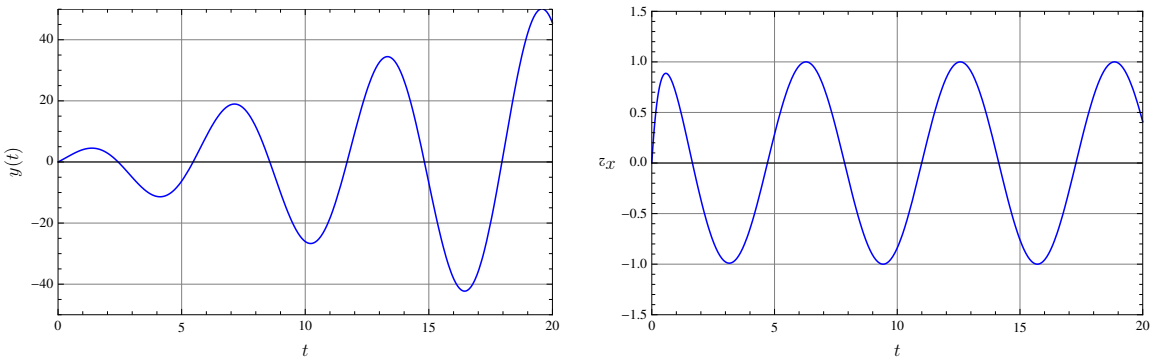


Figure 6.31 Output response for open-loop system (left) and closed-loop system (right) under static output feedback

$$\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t = \begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix}.$$

This is the same closed-loop matrix as obtained in Example 6.50, so the closed-loop state response will be the same as that shown in Figure 6.28. The closed-loop output response is now computed to be

$$y(t) = (2t - 1)e^{-2t} + \cos t,$$

and this is shown in Figure 6.31. The salient fact, of course, is that where the open-loop output was unbounded with the given input, the closed-loop output is now bounded. •

It is also not hard to come up with examples where static output feedback is *not* capable of providing stable closed-loop behaviour (see Examples E6.17 and E6.18). The matter of static output feedback is also attended to in Section 10.2.2, where methods of constructing such feedback laws are discussed.

6.4.3 Dynamic output feedback for SISO linear systems

Thus far, for SISO linear systems, we have investigated only *static* feedback. This feedback is static because we have not use any derivatives of the quantity being fed back. Now let us consider introducing dynamics into the mix. The objective is not just to feedback the output, but also maybe derivatives of the output, and to maybe have the fed back quantity

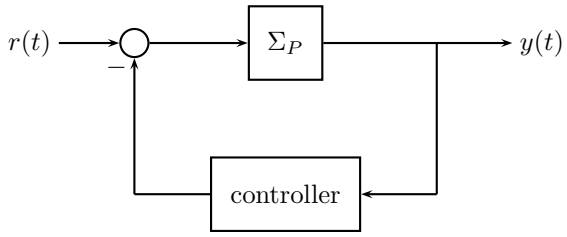


Figure 6.32 A proposed feedback loop for dynamic output feedback

also depend on the input. Thus we start schematically with a block diagram as depicted in Figure 6.32. In the diagram, Σ_P is an abbreviation for the block diagram for a SISO linear plant $\Sigma_P = (\boldsymbol{A}_P, \boldsymbol{b}_P, \boldsymbol{c}_P^t, \boldsymbol{D}_P)$ (i.e., for a block diagram like Figure 3.6). Let us address the question of what lies within the block labelled "controller." For dynamic output feedback, in this block sits Σ_C, a controller SISO system $\Sigma_C = (\boldsymbol{A}_C, \boldsymbol{b}_C, \boldsymbol{c}_C^t, \boldsymbol{D}_C)$. Now note that the diagram of Figure 6.32 looks schematically just like the bottom input/output feedback loop in Figure 6.22. Thus, in being consistent with our discussion of Section 6.3, let us agree to consider a block diagram schematic like Figure 6.33 to model dynamic output feedback.

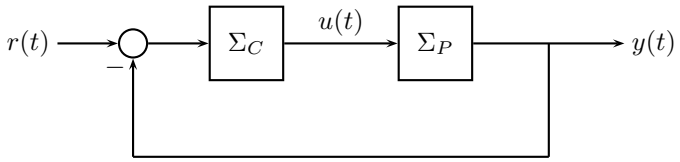


Figure 6.33 Schematic for dynamic output feedback used in text

Therefore, dynamic output feedback consists of connection two SISO linear systems. Note, however, that we have come full circle back to the situation in Section 6.3 where we talked about feedback for input/output systems. Indeed, one can view the designing of a controller rational function R_C as being equivalent to specifying its (say) canonical minimal realisation Σ_C.

Let us get a little more mathematical and write the interconnection of Figure 6.33 is differential equation form. Let us denote the states for the plant by $\boldsymbol{x}_P$ and the states for the controller by $\boldsymbol{x}_C$. We let $u(t)$ be the input to Σ_P which is also the output from Σ_C. We then have

$$\begin{aligned}
\dot{\boldsymbol{x}}_P(t) &= \boldsymbol{A}_P\boldsymbol{x}_P(t) + \boldsymbol{b}_P u(t)\\
\dot{\boldsymbol{x}}_C(t) &= \boldsymbol{A}_C\boldsymbol{x}_C(t) + \boldsymbol{b}_C(r(t) - y(t))\\
y(t) &= \boldsymbol{c}_P^t\boldsymbol{x}_P(t) + \boldsymbol{D}_P u(t)\\
u(t) &= \boldsymbol{c}_C^t\boldsymbol{x}_C(t) + \boldsymbol{D}_C(r(t) - y(t)).
\end{aligned}$$

To obtain the ***closed-loop system*** we use the last two equations to solve for y and u in terms of the other variables. The equations to be solved are

$$\begin{bmatrix} 1 & -\boldsymbol{D}_P \\ \boldsymbol{D}_C & 1 \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_P^t\boldsymbol{x}_P(t) \\ \boldsymbol{c}_C^t\boldsymbol{x}_C(t) + \boldsymbol{D}_C r(t) \end{bmatrix}.$$

We see that in order to solve this equation for $y(t)$ and $u(t)$ we must have $1 + \boldsymbol{D}_C\boldsymbol{D}_P \neq 0$. If this condition is satisfied, we say the interconnection is ***well-posed***. This, it turns out, is exactly the same as the definition of well-posed made in Section 6.3.3 (see Exercise E6.9). After eliminating u, we are left with the input $r(t)$, the output $y(t)$, and the state $(\boldsymbol{x}_P, \boldsymbol{x}_C)$. The following result says that the resulting equations are those for a SISO linear system, and gives the form of the system. The proof is a direct calculation following the outline above.

6.56 Proposition *Suppose that $\boldsymbol{x}_P \in \mathbb{R}^n$ and $\boldsymbol{x}_C \in \mathbb{R}^m$. The closed-loop system for Figure 6.33 is a SISO linear system $\Sigma_{\text{cl}} = (\boldsymbol{A}_{\text{cl}}, \boldsymbol{b}_{\text{cl}}, \boldsymbol{c}_{\text{cl}}^t, \boldsymbol{D}_{\text{cl}})$ where*

$$\begin{aligned}
\boldsymbol{A}_{\text{cl}} &= \begin{bmatrix} \boldsymbol{A}_P & \boldsymbol{0}_{n,m} \\ \boldsymbol{0}_{m,n} & \boldsymbol{A}_C \end{bmatrix} + (1 + \boldsymbol{D}_C\boldsymbol{D}_P)^{-1} \begin{bmatrix} \boldsymbol{b}_P & \boldsymbol{0}_m \\ \boldsymbol{0}_n & \boldsymbol{b}_C \end{bmatrix} \begin{bmatrix} 1 & \boldsymbol{D}_C \\ -\boldsymbol{D}_P & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{0}_n^t & \boldsymbol{c}_C^t \\ -\boldsymbol{c}_P^t & \boldsymbol{0}_m \end{bmatrix}, \\
\boldsymbol{b}_{\text{cl}} &= (1 + \boldsymbol{D}_C\boldsymbol{D}_P)^{-1} \begin{bmatrix} \boldsymbol{b}_P & \boldsymbol{0}_n \\ \boldsymbol{0}_m & \boldsymbol{b}_C \end{bmatrix} \begin{bmatrix} 1 & \boldsymbol{D}_C \\ -\boldsymbol{D}_P & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_C \\ 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{0}_n \\ \boldsymbol{b}_C \end{bmatrix}, \\
\boldsymbol{c}_{\text{cl}}^t &= (1 + \boldsymbol{D}_C\boldsymbol{D}_P)^{-1} \begin{bmatrix} \boldsymbol{c}_P^t & \boldsymbol{D}_P\boldsymbol{c}_C^t \end{bmatrix}, \\
\boldsymbol{D}_{\text{cl}} &= (1 + \boldsymbol{D}_C\boldsymbol{D}_P)^{-1}\boldsymbol{D}_C\boldsymbol{D}_P.
\end{aligned}$$

Note that this is the content of Exercise E2.3, except that in that exercise no feedforward terms were included.

Now that we have the closed-loop system on hand, we may state a problem one often wishes to resolve by the use of dynamic output feedback.

6.57 Dynamic output feedback design problem Given a plant SISO linear system Σ_P, find a controller SISO linear system Σ_C so that the closed-loop system is internally asymptotically stable and well-posed. •

As our final symbolic representation for stabilising controllers, given $\Sigma_P = (\boldsymbol{A}_P, \boldsymbol{b}_P, \boldsymbol{c}_P^t, \boldsymbol{D}_P)$, let us denote by $\mathscr{S}(\Sigma_P)$ the set of SISO linear systems $\Sigma_C = (\boldsymbol{A}_C, \boldsymbol{b}_C, \boldsymbol{c}_C^t, \boldsymbol{D}_C)$ for which the closed-loop system Σ_{cl} is internally asymptotically stable. The set $\mathscr{S}(\Sigma_P)$ is closely related to the set of stabilising controller rational functions, $\mathscr{S}(R_P)$, if $R_P = T_{\Sigma_P}$. In fact, the only essential difference is that all controllers $\Sigma_C \in \mathscr{S}(\Sigma_P)$ will give rise to *proper* controller transfer functions T_{Σ_C}.

6.58 Remarks

1. Note that if the closed-loop system is internally asymptotically stable then it is IBIBO stable. This may not appear obvious, but it actually is. Because all the possible inputs and outputs in the system will simply be linear combinations of the states $\boldsymbol{x}_P$ and $\boldsymbol{x}_C$, it follows that if the states behave stably, then so too will all the inputs and outputs.
2. As we mentioned above, the dynamic output feedback control problem, Problem 6.57, and the input/output control problem, Problem 6.41, are *very* closely related. In the input/output control problem, we were a little more vague, and asked for the closed-loop transfer function to behave in a "suitable manner." For the dynamic output feedback control problem, we were a little more specific because we could be. Nevertheless, even for the dynamic output feedback problem, one often simply wants internal asymptotic stability as a matter of course, and additional requirements will be imposed additionally.
3. Related to the question we asked at the end of Section 6.3.1 is the question of whether given a plant SISO linear system Σ_P, it is alway possible to come up with a controller SISO linear system Σ_C so that the closed-loop system is internally asymptotically stable. This

question is answered in the affirmative in Section 10.2.3, at least under mild assumptions. For example, if Σ_P is controllable and observable, this is possible. However, unlike the analogous situation for input/output systems, it is possible for $\mathscr{S}(\Sigma_P)$ to be empty. •

Let us give an example of how dynamic output feedback can be used to do good. As with all the controllers we have designed thus far, we simply give an *ad hoc* controller that does the job. The matter of coming up with these in a systematic manner is something that we get into in detail in subsequent chapters.

6.59 Example We look at the problem of a mass with no gravitational effects with the control being a force applied to the mass. The differential equation is thus

$$m\ddot{x} = u.$$

As output, let us use the position $x(t)$ of the mass. Putting this into the form $\Sigma_P = (\boldsymbol{A}_P, \boldsymbol{b}_P, \boldsymbol{c}_P^t, \boldsymbol{D}_P)$ gives

$$\boldsymbol{A}_P = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{b}_P = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}, \quad \boldsymbol{c}_P = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{D}_P = \boldsymbol{0}_1.$$

Let us (magically) choose a controller SISO linear system $\Sigma_C = (\boldsymbol{A}_C, \boldsymbol{b}_C, \boldsymbol{c}_C^t, \boldsymbol{D}_C)$ given by

$$\boldsymbol{A}_C = \begin{bmatrix} -2 & 1 \\ -4 & -2 \end{bmatrix}, \quad \boldsymbol{b}_C = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \boldsymbol{c}_C = \begin{bmatrix} 2m \\ 2m \end{bmatrix}, \quad \boldsymbol{D}_C = \boldsymbol{0}_1.$$

A tedious calculation using Proposition 6.56 then gives

$$\boldsymbol{A}_{\text{cl}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ -2 & 0 & -2 & 1 \\ -2 & 0 & -4 & -2 \end{bmatrix}.$$

One checks that the eigenvalues of $\boldsymbol{A}_{\text{cl}}$ are $\{-1+i, -1+i, -1-i, -1-i\}$. Thus $\boldsymbol{A}_{\text{cl}}$ is Hurwitz as desired. This calculation is explained in Example 10.30. The step response for the plant is shown in Figure 6.34, alongside the step response for the closed-loop system. •

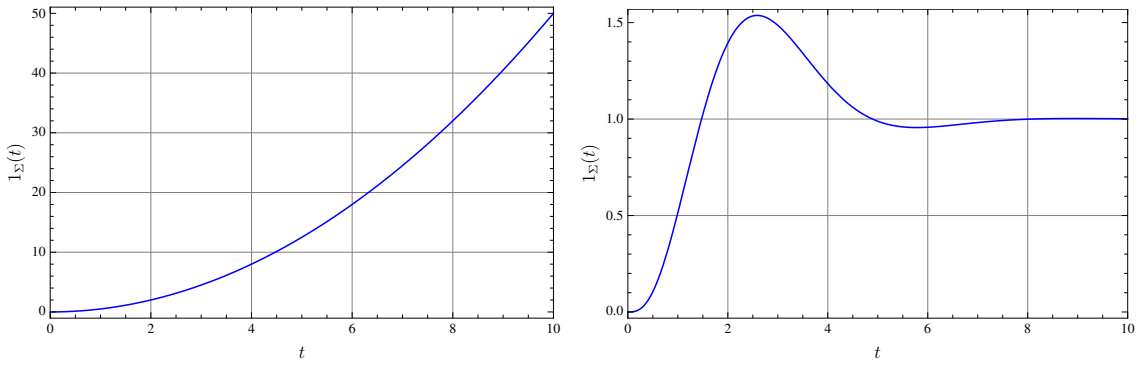


Figure 6.34 Step response for unit mass before (left) and after (right) using dynamic output feedback

6.5 The PID control law

Minorsky, 1922

The so-called PID, for *P*roportional-*I*ntegral-*D*erivative, feedback is very popular, mainly because it is simple and intuitive, and very often quite effective. The PID controller is intended to apply to systems in input/output form, and so if one is dealing with a SISO linear system Σ, one needs to be aware that the PID controller only knows about the input/output behaviour, and not the state behaviour. The idea is that one designs a controller that provides an input to the plant based upon a sum of three terms, one of which is proportional to the error, one of which is proportional to the time-derivative of the error, and the other of which is proportional to the integral of the error with respect to time. In this section we investigate each of these terms, and how they contribute to the controller's performance, and why one should exercise some caution is choosing gains for the PID controller. Knowing what differentiation and integration look like in the Laplace transform domain, we may represent the controller transfer function for a PID control law as in Figure 6.35. The transfer function

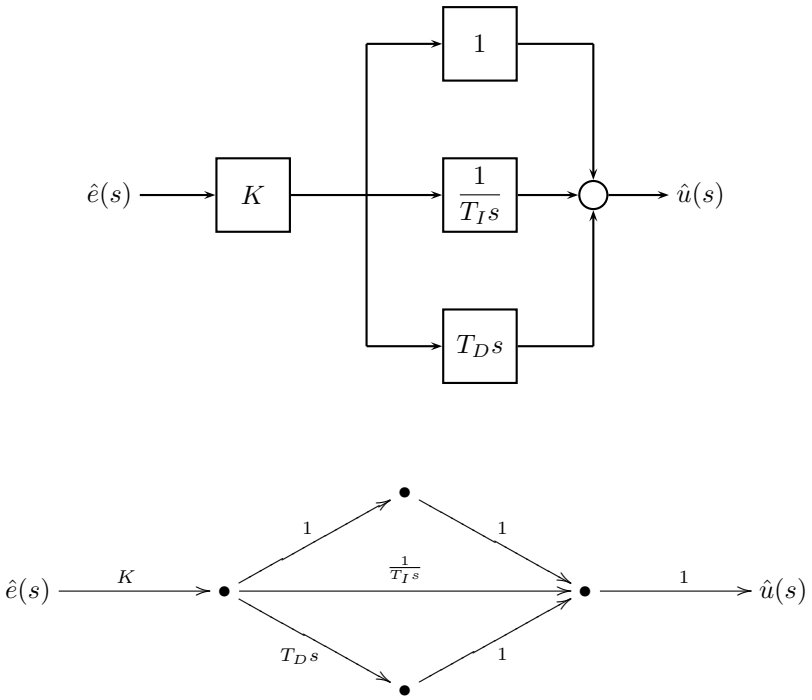


Figure 6.35 The block diagram (top) and signal flow graph for a PID controller

for this controller is

$$R_C(s) = K\Big(1 + T_D s + \frac{1}{T_I s}\Big).$$

The constant K is the ***gain***, and T_D and T_I are the ***derivative time*** and ***reset time***, respectively. Note that the transfer function for the term $T_D s$ in the controller is not proper. This can cause problems, and sometimes one considers the form

$$R_C(s) = K\Big(1 + \frac{T_D s}{\tau_D s + 1} + \frac{1}{T_I s}\Big)$$

for the PID controller, ensuring that all terms in the controller are proper. One can think of the additional factor $(\tau_D s + 1)^{-1}$ as being a low-pass filter added to the second term.

In classical PID design, one makes τ_D small compared to T_D, thus minimising its effects. However, more modern practise allows τ_D to be set as a design parameter. In this chapter, however, we shall always take $\tau_D = 0$, and deal with a straight, classical PID controller. In the design of such controllers, one wishes to determine these constants to accomplish certain objectives.

In investigating the PID controller types, we utilise a specific example. In general, the behaviour of a given controller transfer function will depend to a very large extent upon the nature of the plant being controlled. However, by looking at an example, we hope that we can capture some of the essential features of each of "P," "I," and "D." The example we use will have a plant transfer function

$$R_P(s) = \frac{1}{s^2 + 3s + 2}$$

The closed-loop transfer function is then

$$\frac{KR_C(s)\frac{1}{s^2+3s+2}}{1 + KR_C(s)\frac{1}{s^2+3s+2}}, \tag{6.15}$$

and we consider how choosing a certain type of controller rational function R_C and fiddling with the gain K effects the closed-loop response.

6.5.1 Proportional control

For proportional control we take $R_C(s) = 1$. We compute the closed-loop transfer function to be

$$\frac{K}{s^2 + 3s + 2 + K}$$

and so the characteristic polynomial is $s^2 + 3s + 2 + K$. The roots of the characteristic polynomial are the poles of the closed-loop transfer function, and so these are of great interest to us. We compute them to be $-\frac{3}{2} \pm \frac{1}{2}\sqrt{1 - 4K}$. In Figure 6.36 we plot the set of roots as K varies. Note that the roots are imaginary when $K > \frac{1}{4}$. In Figure 6.36, $K = \frac{1}{4}$ thus corresponds to where the branches meet at $-\frac{3}{2} + i0$.

Observe that by making K large we end up in a situation where the damping remains the same as it was for K small (i.e., the value of the real part when the poles are complex is always $-\frac{3}{2}$), but the frequency of the oscillatory component increases. This could be a rather destructive effect in many systems, and indicates that the system response can suffer when the proportional gain is too large. This is especially a problem for higher-order plants. Proportional control can also suffer from giving a nonzero steady-state error.

6.5.2 Derivative control

Here we take $R_C(s) = T_D s$ for a constant T_D called the ***derivative time***. This transfer function clearly corresponds to differentiating the error signal. The closed-loop transfer function is

$$\frac{Ks}{s^2 + (3 + KT_D)s + 2}$$

which yields the roots of the characteristic polynomial as

$$\frac{-3 - KT_D}{2} \pm \frac{\sqrt{1 + 6KT_D + (KT_D)^2}}{2}.$$

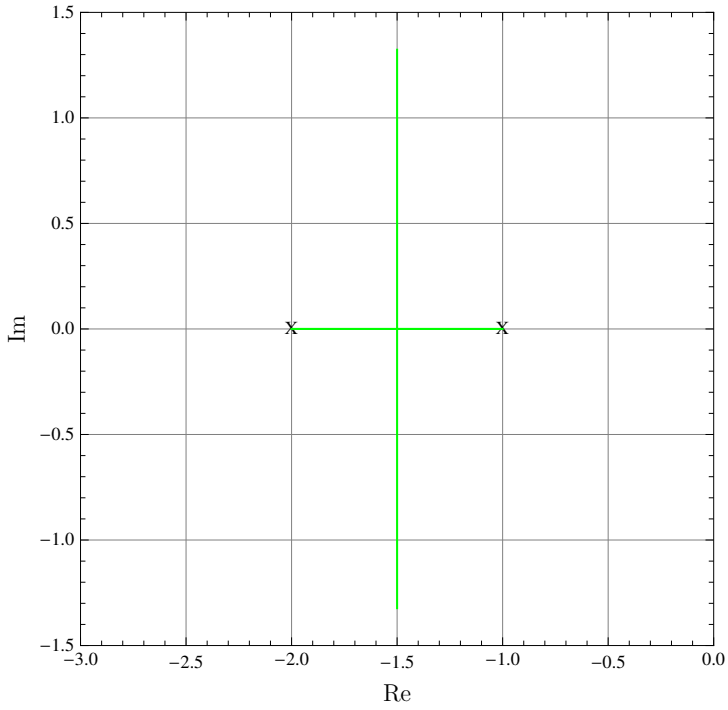


Figure 6.36 The locus of roots for (6.15) with proportional control

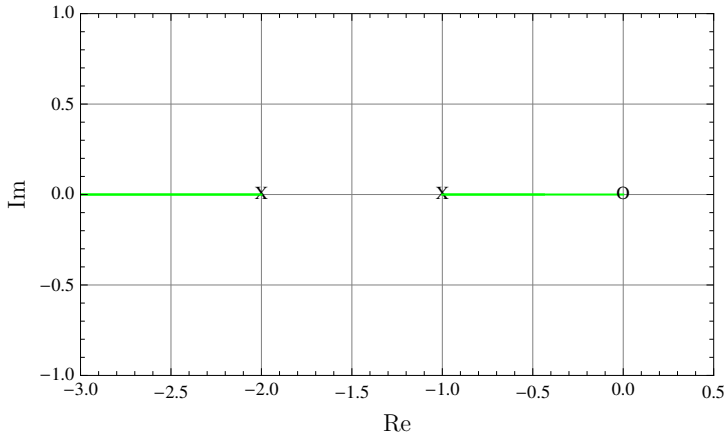


Figure 6.37 The locus of roots for (6.15) with derivative control

The roots are always real for $K > 0$, and the locus of roots is shown in Figure 6.37 as K varies and T_D is fixed to be 1. As we increase K, one root of the characteristic polynomial gets closer and closer to zero. This suggests that for large derivative time, the response will be slow. Derivative control also can suffer from being difficult to implement because accurately measuring velocity is sometimes a difficult task. Furthermore, if used alone, derivative control is incapable of determining steady-state error.

6.5.3 Integral control

We finally look at integral control where we take $R_C(s) = \frac{1}{T_I s}$ where the constant T_I is called the ***reset time***. The closed-loop transfer function is

$$\frac{K}{T_I s^3 + 3T_I s^2 + 2T_I s + K}.$$

The roots of this equation are too complicated to represent conveniently in closed form. However, it is still possible to numerically plot the locus of roots, and we do this in Figure 6.38. We have fixed in this plot $T_I = 1$ and are varying K. Note that as we increase the

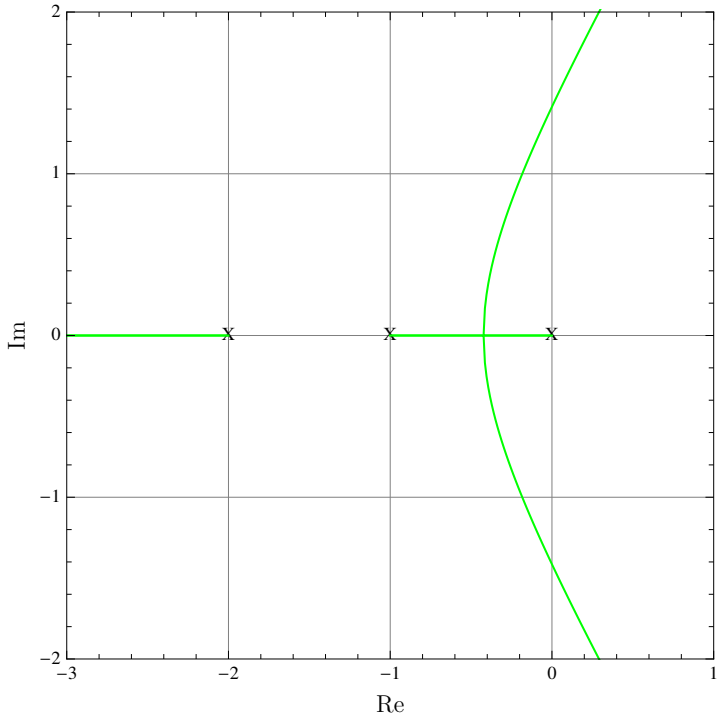


Figure 6.38 The locus of roots for (6.15) with integral control

gain K, the roots of the characteristic polynomial become oscillatory with decreasing real part. Thus integral control by itself will tend to lead to poor behaviour. But it does have the advantage of always giving zero steady-state error to certain common input types.

6.5.4 Characteristics of the PID control law

We will discuss more in Chapter 8 the types of behaviour we are looking for in a good controller, but we summarise some of the features of the various elements of a PID controller in Table 6.1. Often, while each component of the PID controller may by itself not have completely desirable properties, one can obtain satisfactory results by combining them appropriately. The table should be regarded as providing some information on what to look for to better tune a PID controller.

Let's provide a simple example of a PID control application.

6.60 Example We consider a mass m falling under the influence of gravity as in Figure 6.39. At time $t = 0$ the mass is at a height $y = y_0$ and moving with velocity $\dot{y}(0) = v_0$. The mass

Table 6.1 Features of PID control

Type	Advantages	Disadvantages
Proportional	1. Fast response	1. Potentially unstable for large gains and higher-order plants 2. Possible steady-state error for certain types of input
Derivative	1. Good stability properties	1. Difficult to implement 2. Does not correct steady-state errors 3. Magnifies high-frequency noise (see Exercise E6.20)
Integral	1. Corrects steady-state error	1. Large gain leads to lightly damped oscillatory response

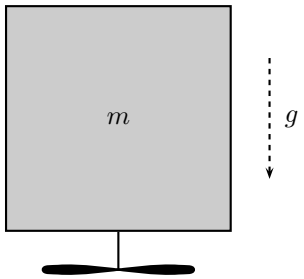


Figure 6.39 Falling mass

has attached to it a fan that is able to provide an upward force u. The differential equations governing the system are

$$m\ddot{y}(t) = -mg + u.$$

We take $y(t)$ as our output. Note that this system is not quite of the type we have been considering because of the presence of the gravitational force. For the moment, therefore, let's suppose that $g = 0$, or equivalently that the motion of the mass is taking place in a direction orthogonal to the direction of gravity. We will reintroduce the gravitational force in Section 8.4. At time $t = 0$ we assume the mass to have state $y(0) = 1$ and $\dot{y}(0) = 1$, and it is our goal to make it move to the height $y = 0$. Thus we take the reference signal $r(t) = 0$. We shall investigate the effects of using proportional, derivative, and integral control.

First we look at making the applied force u proportional to the error. Thus the force we apply has the form $u = -Ky$. The differential equation is then

$$m\ddot{y}(t) + Ky(t) = 0, \quad y(0) = 1, \ \ \dot{y}(0) = 1$$

which has the solution

$$y(t) = \cos\sqrt{\tfrac{K}{m}}t + \sqrt{\tfrac{m}{K}}\sin\sqrt{\tfrac{K}{m}}t.$$

In this case the mass simply oscillates forever, only returning to its desired position periodically. This output is plotted in Figure 6.40 for $m = 1$ and $K = 28$. Note that when we

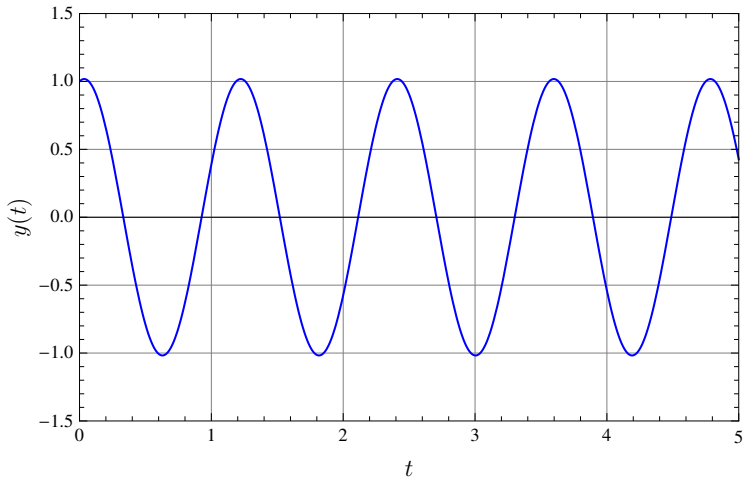


Figure 6.40 Falling mass under proportional control

increase K this has the effect of decreasing the magnitude of the oscillations while increasing their frequency.

Now we examine the situation when we have derivative control. Thus we take $u = -KT_D\dot{y}$. The differential equation is

$$m\ddot{y}(t) + KT_D\dot{y}(t) = 0, \quad y(0) = 1, \;\; \dot{y}(0) = 1$$

which has the solution

$$y(t) = 1 + \frac{m}{KT_D}\left(1 - e^{-(KT_D/m)t}\right).$$

This output is shown in Figure 6.41 for $K = 28$, $m = 1$, and $T_D = \frac{9}{28}$. Note here that the

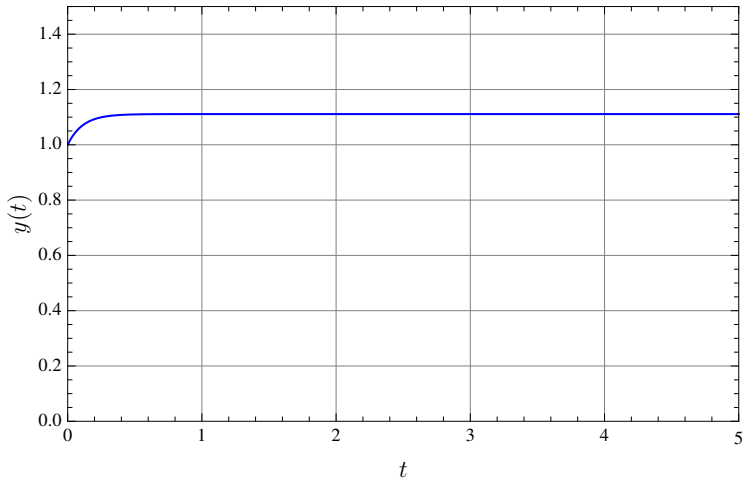


Figure 6.41 Falling mass under derivative control

mass ends up maintaining a steady height that is not the desired height. This is the problem with derivative control—once there is no velocity, the controller stops reacting.

We can also consider integral control. Here we take $u = -\frac{K}{T_I}\int_0^t y(\tau)\,\mathrm{d}\tau$. The differential equation is thus

$$m\ddot{y}(t) + \frac{K}{T_I}\int_0^t y(\tau)\,\mathrm{d}\tau = 0, \quad y(0) = 1,\ \dot{y}(0) = 1. \tag{6.16}$$

To integrate this equation it is most convenient to differentiate it once to get

$$m\dddot{y} + \frac{K}{T_I}y(t) = 0, \quad y(0) = 1,\ \dot{y}(0) = 1,\ \ddot{y}(0) = 0.$$

The initial condition $\ddot{y}(0) = 0$ comes to us from evaluating the equation (6.16) at $t = 0$. This third-order equation may be explicitly solved, but the resulting expression is not particularly worth recording. However, the equation may be numerically integrated and the resulting output is plotted in Figure 6.42 for $K = 28$, $m = 1$, $T_D = \frac{9}{28}$, and $T_I = \frac{7}{10}$. The behaviour

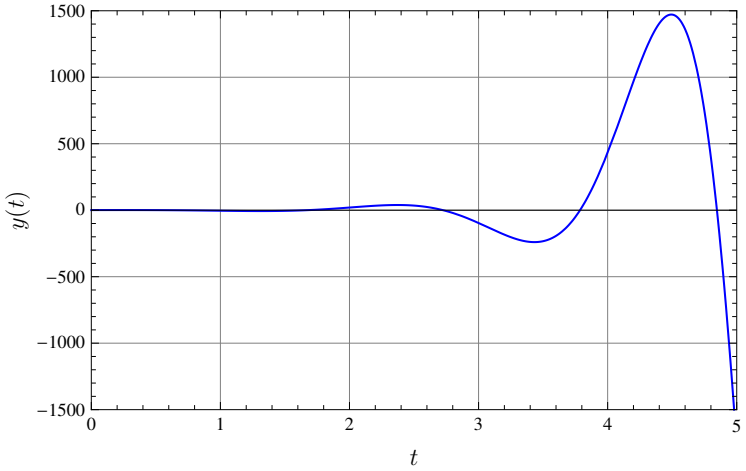


Figure 6.42 Falling mass under integral control

here is oscillatory and unstable.

While none of the controllers individually performed in a reasonable manner, when we combine them, we can get satisfactory performance. When we combine the three controllers, the differential equation is

$$m\dddot{y} + KT_D\ddot{y}(t) + K\dot{y}(t) + \frac{K}{T_I}\int_0^t y(\tau)\,\mathrm{d}\tau = 0, \quad y(0) = 1,\ \dot{y}(0) = 1.$$

Again, to get rid of the integral sign, we differentiate to get

$$m\ddddot{y} + KT_D\dddot{y}(t) + K\ddot{y}(t) + \frac{K}{T_I}y(t) = 0, \quad y(0) = 1,\ \dot{y}(0) = 1,\ \ddot{y}(0) = -K(T_D + 1).$$

This equation may in principle be solved explicitly, but we shall just numerically integrate and show the results in Figure 6.43 for the same parameter values as were used in the plots for the individual controllers. Note that the controller is behaving quite nicely, bringing the mass to the desired height in a reasonable time.

The parameters in the PID controller were not chosen completely by trial and error here—I don't expect the number $\frac{9}{28}$ often gets generated by trial and error. In Chapter 12 we will see how to do select controller parameters in a more systematic manner. That we should expect to be able to do what we wish can be seen by the following result.

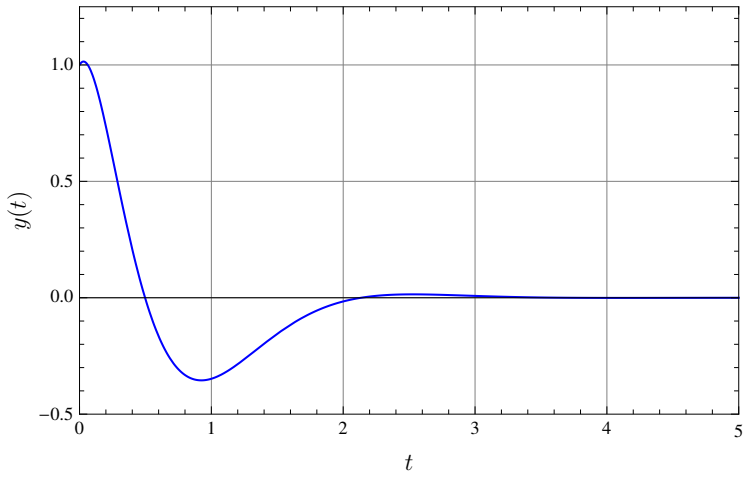


Figure 6.43 Falling mass under combined PID control

6.61 Proposition *For the falling mass under PID control, the poles of the closed-loop system can be chosen to duplicate the roots of any cubic polynomial except those of the form* s^3+as^2+bs+c *where* $b \neq 0$. *If the closed-loop polynomial has the coefficient* $b = 0$ *then it must also be the case that* $a = c = 0$.

Proof The block diagram for the system with the PID feedback is shown in Figure 6.44. The closed-loop transfer function is therefore

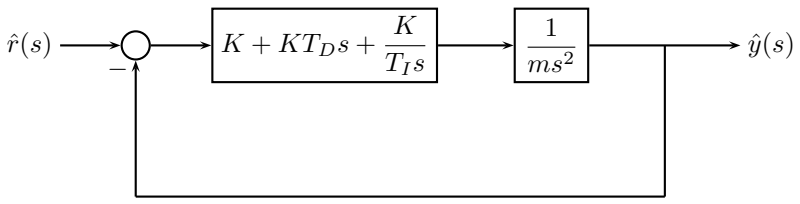


Figure 6.44 Block diagram for falling mass with PID control

$$\frac{\hat{y}(s)}{\hat{r}(s)} = \frac{\frac{1}{ms^2}\left(K + KT_Ds + \frac{K}{T_Is}\right)}{1 + \frac{1}{ms^2}\left(K + KT_Ds + \frac{K}{T_Is}\right)} = \frac{KT_D}{m}\,\frac{s^2 + \frac{1}{T_D}s + \frac{1}{T_DT_I}}{s^3 + \frac{K}{m}T_Ds^2 + \frac{K}{m}s + \frac{K}{mT_I}}.$$

We wish to show that appropriate choices for K, T_D, and T_I can be made so that

$$s^3 + \frac{K}{m}T_Ds^2 + \frac{K}{m}s + \frac{K}{mT_I} = s^3 + as^2 + bs + c$$

for arbitrary $a, b, c \in \mathbb{R}$. If $b \neq 0$ then

$$T_D = \frac{a}{b}, \quad T_I = \frac{b}{c}, \quad K = bm$$

accomplishes the task. If $b = 0$ case that $K = 0$ and so a and c must also be zero. ■

Thus we can place poles for the closed-loop transfer function pretty much anywhere we wish. I used the proposition to select poles at $\{-5, -2 \pm 2i\}$.

And with that we leave the wee falling mass to its own devices. •

6.6 Summary

Of course, the notion of feedback is an important one in control. In this section, we have discussed some of the issues surrounding feedback in a fairly general way. In later chapters, we will be deciding how to deploy feedback in an effective manner. Let us do as we have been doing, and list some of the more essential points raised in this chapter.

1. The first few sections of the chapter dealt with the setup surrounding "interconnected SISO linear systems." The generality here is somewhat extreme, but it serves to properly illustrate the problems that can arise when interconnecting systems.
2. You might find it helpful to be able to switch freely from block diagrams to signal flow graphs. Sometimes the latter are the more useful form, although the former are more commonly used in practice.
3. One should know immediately how to compute the determinant and characteristic polynomial for simple block diagram configurations. One should also be able to use the signal flow graph technology to determine the transfer function between various inputs and outputs in a block diagram configuration.
4. One should be able to test a block diagram configuration for IBIBO stability.
5. Although we have not said much about controller design, one should understand the issues surrounding the design problem for both SISO linear systems in input/output form and for SISO linear systems.
6. Much of what we say as we go on will apply to the simple unity gain feedback loop. This is simple, and one ought to be able to work with these fluently.
7. You should know what the PID control law is, and have some feel for how its proportional, derivative, and integral components affect the performance of a system.

Exercises

E6.1 Let $(\mathcal{S}, \mathcal{G})$ be a signal flow graph with $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$ its corresponding matrix. Show that the gains in the ith column correspond to branches that originate from the ith node, and that the gains in the jth row correspond to branches that terminate at the jth node.

E6.2 Consider the block diagram of Exercise E3.2.

(a) Draw the corresponding signal flow graph.

(b) Write the signal flow graph as $(\mathcal{S}, \mathcal{G})$ as we describe in the text—thus you should identify the nodes and how the nodes are connected.

(c) Write the structure matrix $\boldsymbol{G}_{\mathcal{S},\mathcal{G}}$.

(d) Determine the pair $(\boldsymbol{A}_{\mathcal{S},\mathcal{G}}, \boldsymbol{B}_{\mathcal{S},\mathcal{G}})$ as per Procedure 6.22.

(e) Write all simple paths through the signal flow graph which connect the input $\hat{r}$ with the output $\hat{y}$.

(f) Identify all loops in the graph by writing their gains—that is, determine $\mathrm{Loop}(\mathcal{S}, \mathcal{G})$.

(g) For $k \geq 1$ determine $\mathrm{Loop}_k(\mathcal{S}, \mathcal{G})$.

(h) Find $\Delta_{\mathcal{S},\mathcal{G}}$.

(i) Find $T_{\mathcal{S},\mathcal{G}}$.

(j) Write each of the rational functions $R_1, \dots, R_6$ in Figure E3.1 as a numerator polynomial over a denominator polynomial and then determine $P_{\mathcal{S},\mathcal{G}}$.

E6.3 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system.

(a) Show that there exists $\boldsymbol{P}_\Sigma \in \mathbb{R}[s]^{n \times n}$ so that $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$ if and only if $\boldsymbol{P}_\Sigma\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)\boldsymbol{x}(t) = \boldsymbol{0}$.

(b) Show that Theorem 5.2 then follows from Theorem 6.33. That is, show that the hypotheses of Theorem 5.2 imply the hypotheses of Theorem 6.33.

E6.4 For the block diagram of Figure E6.1, determine the values of a, b, and c for which

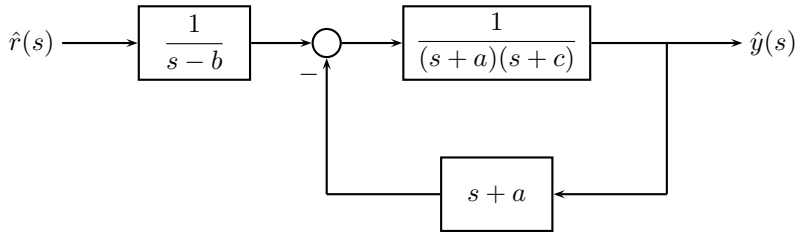


Figure E6.1 A block diagram for determining IBIBO stability

the interconnected system is IBIBO stable. For which values of a and b can stability be inferred from the zeros of the determinant, without having to resort to looking at the characteristic polynomial.

E6.5 For the block diagram of Exercise E3.2 (which you investigated as a signal flow graph in Exercise E6.2), consider the following assignment of specific rational functions to the block transfer functions:

$$R_1(s) = \frac{1}{s}, \quad R_2(s) = s - 1, \quad R_3(s) = \frac{1}{s+b},$$
$$R_4(s) = \frac{1}{s+2}, \quad R_5(2) = \frac{1}{s+a}, \quad R_6(s) = s + a.$$

Answer the following questions.

(a) Use Theorem 6.38 and the Routh/Hurwitz criteria to ascertain for which values of a and b the interconnected system is IBIBO stable.

(b) For which values of a and b can IBIBO stability of the interconnected system be inferred from looking only at the determinant of the graph without having to resort to using the characteristic polynomial?

E6.6 In this exercise, you will investigate the matter of relating BIBO stability of individual branch gains in a signal flow graph $(\mathcal{S}, \mathcal{G})$ to the IBIBO stability of the interconnection.

(a) Is it possible for $(\mathcal{S}, \mathcal{G})$ to be IBIBO stable, and yet have branch gains that are not BIBO stable? If it is not possible, explain why not. If it is possible, give an example.

(b) Is it true that BIBO stability of all branch gains for $(\mathcal{S}, \mathcal{G})$ implies IBIBO stability of the interconnection? If it is true, prove it. If it is not true, give a counterexample.

E6.7 Prove Proposition 6.45 directly, without reference to Theorem 6.38.

E6.8 Well-posedness and existence and uniqueness of solutions. finish

E6.9 For the feedback interconnection of Figure E6.2, let $\Sigma_C = (\boldsymbol{A}_1, \boldsymbol{b}_1, \boldsymbol{c}_1^t, \boldsymbol{D}_1)$ be the

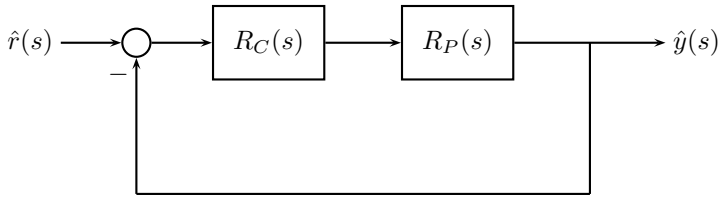


Figure E6.2 Plant/controller feedback loop

canonical minimal realisation for R_C and let $\Sigma_P = (\boldsymbol{A}_2, \boldsymbol{b}_2, \boldsymbol{c}_2^t, \boldsymbol{D}_2)$ be the canonical minimal realisation for R_P. Show that the interconnection is well-posed if and only if $\boldsymbol{D}_1\boldsymbol{D}_2 \neq [-1]$.

E6.10 For the block diagram configuration of Figure E6.3, show that as $K \to 0$ the poles

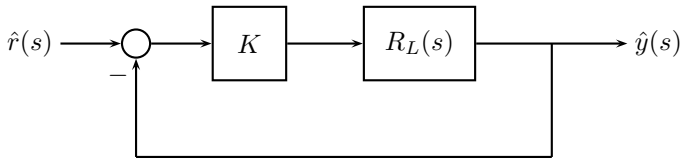


Figure E6.3 Unity gain feedback loop with variable gain

and zeros of the closed-loop transfer function approach those of R_L.

E6.11 Let $R_P \in \mathbb{R}(s)$ be a proper plant. For $R_C \in \mathbb{R}(s)$ define $R_L = R_C R_P$, as usual, and let T_L be the corresponding closed-loop transfer function. Suppose that R_P has zeros $z_1, \ldots, z_\ell \in \overline{\mathbb{C}}_+ \cup \{\infty\}$ and poles $p_1, \ldots, p_k \in \overline{\mathbb{C}}_+$ (there may be other poles and zeros of the plant, but we do not care about these). Prove the following result.

Proposition *$R_C \in \mathscr{S}(R_P)$ if and only if the following four statements hold:*

(i) $T_L \in \mathrm{RH}^+_\infty$;

(ii) *the zeros of* $1 - T_L$ *contain* $\{p_1, \ldots, p_k\}$, *including multiplicities;*

(iii) *the zeros of* T_L *contain* $\{z_1, \ldots, z_\ell\}$, *including multiplicities;*

(iv) $\lim_{s\to\infty} R_L(s) \neq -1$.

E6.12 Let $R_P \in \mathbb{R}(s)$ be a proper plant. For $R_C \in \mathbb{R}(s)$ define $R_L = R_C R_P$, as usual, and let S_L be the corresponding sensitivity function. Suppose that R_P has zeros $z_1, \ldots, z_\ell \in \overline{\mathbb{C}}_+ \cup \{\infty\}$ and poles $p_1, \ldots, p_k \in \overline{\mathbb{C}}_+$ (there may be other poles and zeros of the plant, but we do not care about these). Prove the following result.

Proposition *$R_C \in \mathscr{S}(R_P)$ if and only if the following four statements hold:*

(i) $S_L \in \mathrm{RH}^+_\infty$;

(ii) *the zeros of* S_L *contain* $\{p_1, \ldots, p_k\}$, *including multiplicities;*

(iii) *the zeros of* $1 - S_L$ *contain* $\{z_1, \ldots, z_\ell\}$, *including multiplicities;*

(iv) $\lim_{s\to\infty} R_L(s) \neq -1$.

E6.13 Let $R_P \in \mathbb{R}(s)$ be a BIBO stable plant. Show that there exists a controller $R_C \in \mathbb{R}(s)$ for which the interconnection of Figure E6.2 is IBIBO stable with closed-loop transfer function T_L if and only if $T_L, \frac{T_L}{R_P} \in \mathrm{RH}^+_\infty$.

E6.14 Let $R_P \in \mathbb{R}(s)$ be a proper plant transfer function, and let $(\mathcal{S}, \mathcal{G})$ be an interconnected SISO linear system with the property that every forward path from the input to the output passes through the plant. Show that IBIBO stability of $(\mathcal{S}, \mathcal{G})$ implies that $T_{\mathcal{S},\mathcal{G}}, \frac{T_{\mathcal{S},\mathcal{G}}}{R_P} \in \mathrm{RH}^+_\infty$.

E6.15 In this exercise you will show that by feedback it is possible to move into $\mathbb{C}_-$ the poles of a closed-loop transfer function, even when the poles of the plant are in $\mathbb{C}_+$. Consider the closed-loop system as depicted in Figure 6.21 with

$$R_C(s) = 1, \quad R_P(s) = \frac{1}{(s+1)(s-a)},$$

with $a > 0$. Determine for which values of the gain K the closed-loop system has all poles in $\mathbb{C}_-$. Is the system IBIBO stable when all poles of the closed-loop system are in $\mathbb{C}_-$?

In this exercise we explore the relationship between performing static state feedback for SISO linear systems, and performing design for controller rational functions for input/output systems.

E6.16 Consider a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ with $(\boldsymbol{A}, \boldsymbol{b})$ controllable and $(\boldsymbol{A}, \boldsymbol{c})$ observable. Let $R_P = T_\Sigma$. For $\boldsymbol{f} = (f_0, f_1, \ldots, f_{n-1}) \in \mathbb{R}^n$ define the polynomial

$$F(s) = f_{n-1}s^{n-1} + \cdots + f_1 s + f_0 \in \mathbb{R}[s].$$

Suppose that $(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t, \boldsymbol{c})$ is observable.

(a) Show that there exists a controller rational function R_C with the property that the poles of the two transfer functions

$$T_{\Sigma_{\boldsymbol{f}}}, \quad T = \frac{R_C R_P}{1 + R_C R_P}$$

agree if and only if the polynomial N_P divides the polynomial F over $\mathbb{R}[s]$. In particular, show that if N_P is a constant polynomial, then it is always possible to find a controller rational function R_C with the property that the poles of T_{Σ_f} and T agree.

Hint: *Without loss of generality suppose that* $(\mathbf{A}, \mathbf{b})$ *is in controller canonical form, and look at the proof of Proposition 10.13.*

(b) Suppose that N_P divides F over $\mathbb{R}[s]$ and by part (a) choose a controller rational function R_C with the property that the poles of T_{Σ_f} and T agree. What is the difference of the numerators polynomials for T_{Σ_f} and T.

Thus the problem of placement of poles in feedback design for SISO linear systems can sometimes be realised as feedback design for input/output systems. The following parts of the problem show that there are some important cases where controller rational function design cannot be realised as design of a state feedback vector.

Let (N, D) be a strictly proper SISO linear system in input/output form with $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$ the canonical minimal realisation. Suppose that N has no root at $s = 0$. Let $R_P = T_{N,D}$.

(c) Is it possible, if R_C is the controller rational function for a PID controller, to find $\boldsymbol{f} \in \mathbb{R}^n$ so that the poles of the transfer functions of part (a) agree?

E6.17 Consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$ with

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

For this system, answer the following.

(a) Show that there is no continuous function $u(x_1)$ with the property that for every solution $\boldsymbol{x}(t)$ of the differential equation

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(x_1(t)) \tag{E6.1}$$

satisfies $\lim_{t\to\infty} \|\boldsymbol{x}\|(t) = 0$.

Hint: *First prove that the function*

$$\mathrm{V}(\mathbf{x}) = \frac{1}{2}\mathrm{x}_2^2 - \int_0^{\mathrm{x}_1} \mathrm{u}(\xi)\,\mathrm{d}\xi$$

is constant along solutions of the differential equation (E6.1).

(b) If $\boldsymbol{f} \in \mathbb{R}^2$ is a state feedback vector for which $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ is Hurwitz, what can be said about the form of $\boldsymbol{f}$ from part (a).

E6.18 In this exercise we generalise Exercise E6.17 for linear feedback. We let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$ be a SISO linear system with $\mathrm{tr}(\boldsymbol{A}) = 0$ and $\boldsymbol{c}^t\boldsymbol{b} = 0$.

(a) Show that there is no output feedback number F with the property that the closed-loop system is internally asymptotically stable.

Hint: *Show that if*

$$\mathrm{P}_{\mathbf{A}}(\mathrm{s}) = \mathrm{s}^{\mathrm{n}} + \mathrm{p}_{\mathrm{n}-1}\mathrm{s}^{\mathrm{n}-1} + \cdots + \mathrm{p}_1\mathrm{s} + \mathrm{p}_0$$

is the characteristic polynomial of $\mathbf{A}$*, then* $-\mathrm{p}_{\mathrm{n}-1} = \mathrm{tr}(\mathbf{A})$ *(think of putting the matrix in complex Jordan canonical form and recall that trace is invariant under similarity transformations).*

(b) Something with Liapunov finish

E6.19 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a complete SISO linear system and suppose that $z \in \mathbb{C}$ is a zero of T_Σ. Show that by static output feedback it is not possible to obtain a closed-loop system with a pole at z.

One of the potential problems with derivative control is that differentiation can magnify high frequency noise, as the following exercise points out.

E6.20 For a time signal

$$y(t) = A_\mathrm{s} \sin(\omega_\mathrm{s} t) + A_\mathrm{n} \sin(\omega_\mathrm{n} t + \phi_\mathrm{n}),$$

consisting of a sinusoidal signal (the first term) along with sinusoidal noise (the second term), the ***signal-to-noise ratio*** is defined by $S/N = \frac{|A_\mathrm{s}|}{|A_\mathrm{n}|}$.

(a) Show that for any such signal, the signal-to-noise ratio for $\dot{y}$ tends to zero as ω_n tends to infinity.

(b) Indicate in terms of Bode plots why differentiation is bad in terms of amplifying high frequency noise.

E6.21 In this exercise we will investigate in detail the DC servo motor example that was used in Section 1.2 to provide an illustration of some control concepts.

We begin by making sure we know how to put the model in a form we can deal with. We model the system as a SISO linear system whose single state is the angular velocity of the motor. The output is the angular velocity of the motor (i.e., the value of the system's only state), and the input is the voltage to the motor.

(a) Determine $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$. Your model should incorporate the time-constant τ and motor gain k_E as in Section 1.2, but do not include any effects from external disturbances.

(b) Determine T_Σ.

We let the plant transfer function R_P be T_Σ whose c.f.r. we write as (N_P, D_P). For the reasons we discussed in Section 1.2, an open-loop control scheme, while fine in an idealised environment, lacks robustness. Thus we employ a closed-loop control scheme like that depicted in Figure 6.21. For proportional control we use $R_C(s) = 1$.

(c) Determine the closed-loop transfer function with gain K.

(d) Assuming that τ and k_E are both positive, determine the range of gains K for which the closed-loop system has all poles in $\mathbb{C}_-$. That is, determine K_min and K_max so that the closed-loop system has poles in $\mathbb{C}_-$ if and only if $K \in (K_\mathrm{min}, K_\mathrm{max})$.

Now we will see how the closed-loop system's frequency response represents its ability to track sinusoidal inputs.

(e) Determine the frequency response for the closed-loop transfer function.

(f) Determine the output response to system when the reference signal is $r(t) = \cos \omega t$ for some $\omega > 0$. Assuming that $K \in (K_\mathrm{min}, K_\mathrm{max})$, what is the steady-state response, $y_\mathrm{ss}(t)$.

(g) Show that $\lim_{K \to \infty} y_\mathrm{ss}(t) = \cos \omega t$, and so as we boost the gain higher, we can in principle exactly track a sinusoidal reference signal. Can you see this behaviour reflected in the frequency response of the system?

E6.22 Consider the coupled masses of Exercise E1.4 (assume no friction). As input take the situation in Exercise E2.19 with $\alpha = 0$. Thus the input is a force applied only to the leftmost mass.

We wish to investigate the effect of choosing an output on our ability to manipulate the poles of the closed-loop transfer function. We first consider the case when the output is the displacement of the rightmost mass.

(a) Determine the transfer function T_Σ for the system with this output.

(b) What are the poles for the transfer function? What does this imply about the uncontrolled behaviour of the coupled mass system?

First we look at an open-loop controller as represented by Figure 6.20. We seek a controller rational function R_C whose c.f.r. we denote (N_C, D_C). We also let $R_P = T_\Sigma$, and denote by (N_P, D_P) its c.f.r.

(c) Can you find a controller transfer function R_C so that

1. D_P and N_C are coprime and
2. the open-loop transfer function has all poles in $\mathbb{C}_-$?

Why do we impose the condition 1?

Now we look for a closed-loop controller as represented by Figure 6.21. For simplicity, we begin using a proportional control.

(d) For proportional control, suppose that $R_C(s) = 1$, and derive the closed-loop transfer function with gain K.

(e) Show that it is impossible to design a proportional control law for the system with the properties

1. D_P and N_C are coprime and
2. the closed-loop transfer function has all poles in $\mathbb{C}_-$?

Hint: *Show that for a polynomial* $\mathrm{s}^4 + \mathrm{a}\mathrm{s}^2 + \mathrm{b} \in \mathbb{R}[\mathrm{s}]$*, if* $\mathrm{s}_0 = \sigma_0 + \mathrm{i}\omega_0$ *is a root, then so are* $\sigma_0 - \mathrm{i}\omega_0$*,* $-\sigma_0 + \mathrm{i}\omega_0$*, and* $-\sigma_0 - \mathrm{i}\omega_0$*.*

It turns out, in fact, that introducing proportional and/or derivative control into the problem described to this point does not help. The difficulty is with our plant transfer function. To change it around, we change what we measure.

Thus, for the remainder of the problem, suppose that the output is the *velocity* of the *leftmost* mass (make sure you use the correct output).

(f) Determine the transfer function T_Σ for the system with this output.

(g) What are the poles for the transfer function?

The open-loop control problem here is "the same" as for the previous case where the output was displacement of the rightmost mass. So now we look for a closed-loop proportional controller for this transfer function.

(h) For proportional control, suppose that $R_C(s) = 1$, and derive the closed-loop transfer function with gain K.

(i) Choose $m = 1$ and $k = 1$, and show numerically that there exists $K > 0$ so that the poles of the closed-loop transfer function all lie in $\mathbb{C}_-$.

E6.23 Refer to Exercise E6.21. Set the time constant $\tau = 1$ and the motor constant $k_E = 1$. Produce the Bode plots for plant transfer function, and for the closed-loop system with the proportional controller with gains $K = \{1, 10, 100\}$. Describe the essential differences in the Bode plots. How is your discovery of Exercise E6.21(g) reflected in your Bode plots?

E6.24 Refer to Exercise E6.22, taking $m = 1$ and $k = 1$, and use the second output (i.e., the velocity of the leftmost mass). Produce the Bode plots for plant transfer function, and for the closed-loop system with the proportional controller with gains

$K = \{1, 10, 100\}$. Describe the essential differences in the Bode plots. In what way does the open-loop transfer function differ from the rest?

E6.25 Consider the controller transfer function $R_C(s) = K\big(1 + T_D s + \frac{1}{T_I s}\big)$.

(a) Can you find a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ so that $T_\Sigma = R_C$? (Assume that K, T_D, and T_I are finite and nonzero.)

(b) What does this tell you about the nature of the relationship between the Problems 6.41 and 6.57?

PID control is widely used in many industrial settings, due to its easily predictable behaviour, at least when used with "simple" plants. In the next exercise you will see what one might mean by simple.

E6.26 Consider the interconnection in Figure E6.4 with R_P a proper plant. Suppose that if

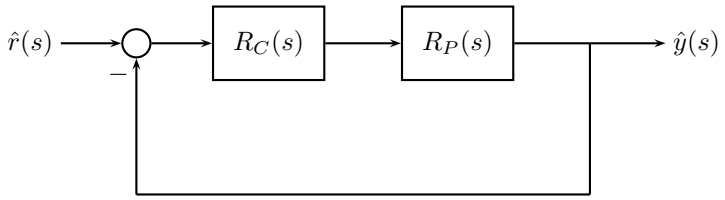


Figure E6.4 Feedback loop for studying properties of PID control

$R_C = 1$ then the interconnection is IBIBO stable. Show that there exists $K_0 > 1$ and $T_{D,0}, T_{I,0} > 0$ so that the controller

$$R_C(s) = K\Big(1 + T_D s + \frac{1}{T_I s}\Big)$$

IBIBO stabilises the interconnection for all $K \in [1, K_0]$, $T_D \in [0, T_{D,0}]$, and $T_I \in [T_{I,0}, \infty)$.

Hint: *Use the Nyquist criterion of Section 7.1 to show that the number of encirclements of* $-1+\mathrm{i}0$ *does not change for a PID controller with the parameters satisfying* $\mathrm{K} \in [1, \mathrm{K}_0]$, $\mathrm{T_D} \in [0, \mathrm{T_{D,0}}]$, *and* $\mathrm{T_I} \in [\mathrm{T_{I,0}}, \infty)$.

In the next exercise we will consider a "difficult" plant; one that is unstable and nonminimum phase. For this plant you will see that any "conventional" strategies for designing a PID controller, based on the intuitive ideas about PID control as discussed in Section 6.5, are unlikely to meet with success.

E6.27 Consider, still using the interconnection of Figure E6.4, the plant

$$R_P(s) = \frac{1 - s}{s(s - 2)}.$$

Answer the following questions.

(a) Show that it is not possible to IBIBO stabilise the system using a PID controller with positive parameters K, T_D, and T_I.

Hint: *One can use one of the several polynomial stability tests of Section 5.5. However, it turns out that the Routh test provides the simplest way of getting at what we want here.*

Thus we must take at least one of the PID parameters to be negative. Let us consider the simplest situation where we take $K < 0$, so that perhaps *some* of our intuition about PID controllers persists.

(b) Show that if $K < 0$ then it is necessary for IBIBO stability that $T_I > 1$.

(c) Show that if $K < 0$ and $T_I > 1$ then it is necessary for IBIBO stability that $T_D < 0$.

Chapter 7

Frequency domain methods for stability

In Chapter 5 we looked at various ways to test various notions of stability of SISO control systems. Our stability discussion in that section ended with a discussion in Section 6.2.3 of how interconnecting systems in block diagrams affects stability of the resulting system. The criterion developed by Nyquist [1932] deals further with testing stability in such cases, and we look at this in detail in this chapter. The methods in this chapter rely heavily on some basic ideas in complex variable theory, and these are reviewed in Appendix D.

Contents

7.1 The Nyquist criterion

The Nyquist criterion is a method for testing the closed-loop stability of a system based on the frequency response of the open-loop transfer function.

7.1.1 The Principle of the Argument

In this section we review one of the essential tools in dealing with closed-loop stability as we shall in this chapter: the so-called Principle of the Argument. This is a result from the theory of complex analytic functions. That such technology should be useful to us has been made clear in the developments of Section 4.4.2 concerning Bode's Gain/Phase Theorem. The Principle of the Argument has to do, as we shall use it, with the image of closed contours under analytic functions. However, let us first provide its form in complex analysis. Let $U \subset \mathbb{C}$ be an open set, and let $f : U \to \mathbb{C}$ be an analytic function. A ***pole*** for f is a point $s_0 \in \overline{U}$ with the property that the limit $\lim_{s \to s_0} f(s)$ does not exist, but there exists a $k \in \mathbb{N}$ so that the limit $\lim_{s \to s_0} (s - s_0)^k f(s)$ does exist. Recall that a ***meromorphic*** function on an open subset $U \subset \mathbb{C}$ is a function $f : U \to \mathbb{C}$ having the property that it is defined and analytic except at isolated points (i.e., except at poles). Now we can state the

Principle of the Argument, which relies on the Residue Theorem stated in Appendix D.

7.1 Theorem **(Principle of the Argument)** *Let U be a simply connected open subset of $\mathbb{C}$ and let C be a contour in U. Suppose that f is a function which*

(i) is meromorphic in U,

(ii) has no poles or zeros on the contour C, and

(iii) has n_p poles and n_z zeros in the interior of C, counting multiplicities of zeros and poles.

Then

$$\int_C \frac{f'(s)}{f(s)}\,\mathrm{d}s = 2\pi i(n_z - n_p),$$

provided integration is performed in a counterclockwise direction.

Proof Since f is meromorphic, $\frac{f'}{f}$ is also meromorphic, and is analytic except at the poles and zeroes of f. Let s_0 be such a pole or zero, and suppose that it has multiplicity k. Then there exists a meromorphic function $\tilde{f}$, analytic an s_0, with the property that $f(s) = (s - s_0)^k \tilde{f}(s)$. One then readily determines that

$$\frac{f'(s)}{f(s)} = \frac{k}{s - s_0} + \frac{\tilde{f}'(s)}{\tilde{f}(s)}$$

for s in a neighbourhood of s_0. Now by the Residue Theorem the result follows since k is positive if s_0 is a zero and negative if s_0 is a pole. ■

The use we will make of this theorem is in ascertaining the nature of the image of a closed contour under an analytic function. Thus we let $U \subset \mathbb{C}$ be an open set and $c\colon [0, T] \to U$ a closed curve. The image of c we denote by C, and we let f be a function satisfying the hypotheses of Theorem 7.1. Let us denote by $\tilde{c}$ the curve defined by $\tilde{c}(t) = f \circ c(t)$. Since f has no zeros on C, the curve $\tilde{c}$ does not pass through the origin and so the function $F\colon [0, T] \to \mathbb{C}$ defined by $F(t) = \ln(\tilde{c}(t))$ is continuous. By the chain rule we have

$$F'(t) = \frac{f'(c(t))}{f(c(t))} c'(t), \quad t \in [0, T].$$

Therefore

$$\int_C \frac{f'(s)}{f(s)}\,\mathrm{d}s = \int_C \frac{f'(c(t))}{f(c(t))} c'(t)\,\mathrm{d}t = \ln(f(c(t)))\Big|_0^T.$$

Using the definition of the logarithm we have

$$\ln(f(c(t)))\Big|_0^T = \ln|f(c(t))|\Big|_0^T + i\measuredangle f(c(t))\Big|_0^T.$$

Since c is closed, the first term on the right-hand side is zero. Using Theorem 7.1 we then have

$$2\pi(n_z - n_p) = \measuredangle f(c(t))\Big|_0^T. \tag{7.1}$$

In other words, we have the following.

7.2 Proposition *If C and f are as in Theorem 7.1, then the image of C under f encircles the origin $n_z - n_p$ times, with the convention that counterclockwise is positive.*

Let us illustrate the principle with an example.

7.3 Example We take C to be the circle of radius 2 in $\mathbb{C}$. This can be parameterised, for example, by $c\colon t \mapsto 2e^{it}$, $t \in [0, 2\pi]$. For f we take

$$f(s) = \frac{1}{(s+1)^2 + a^2}, \quad a \in \mathbb{R}.$$

Let's see what happens as we allow a to vary between 0 and 1.5. The curve $\tilde{c} = f \circ c$ is defined by

$$t \mapsto \frac{1}{(e^{2it}+1)^2 + a^2}, \quad t \in [0, 2\pi].$$

Proposition 7.2 says that for $a < 1$ the image of C should encircle the origin in $\mathbb{C}$ two times in the counterclockwise direction, and for $a > 1$ there should be no encirclements of the origin. Of course, it is problematic to determine the image of a closed contour under a given analytic function. Here we let the computer do the work for us, and the results are shown in Figure 7.1. We see that the encirclements are as we expect. •

7.1.2 The Nyquist criterion for single-loop interconnections

Now we apply the Principle of the Argument to determine the stability of a closed-loop transfer function. The block diagram configuration we consider here is shown in Figure 7.2. The key observation is that if the system is to be IBIBO stable then the poles of the closed-loop transfer function

$$T(s) = \frac{R_C(s)R_P(s)}{1 + R_C(s)R_P(s)}$$

must all lie in $\mathbb{C}_-$. The idea is that we examine the determinant $1 + R_C R_P$ to ascertain when the poles of T are stable.

We denote the loop gain by $R_L = R_C R_P$. Suppose that R_L has poles on the imaginary axis at $\pm i\omega_1, \dots, \pm i\omega_k$ where $\omega_k > \cdots > \omega_1 \geq 0$. Let $r > 0$ have the property that

$$r < \frac{1}{2} \min_{\substack{i,j \in \{1,\dots,k\} \\ i \neq j}} \{|\omega_i - \omega_j|\}. \tag{7.2}$$

That is, r is smaller than half the distance separating the two closest poles on the imaginary axis. Now we choose R so that

$$R > \omega_k + \frac{r}{2}. \tag{7.3}$$

With r and R so chosen we may define a contour $\Gamma_{R,r}$ which will be comprised of a collection of components. For $i = 1, \dots, k$ we define

$$\Gamma_{r,k,+} = \left\{ i\omega_i + re^{i\theta} \;\middle|\; -\tfrac{\pi}{2} \leq \theta \leq \tfrac{\pi}{2} \right\}, \quad \Gamma_{r,k,-} = \left\{ -i\omega_i + re^{i\theta} \;\middle|\; -\tfrac{\pi}{2} \leq \theta \leq \tfrac{\pi}{2} \right\}.$$

Now for $i = 1, \dots, k-1$ define

$$\bar{\Gamma}_{r,i,+} = \{i\omega \mid \omega_i + r < \omega < \omega_{i+1} - r\}, \quad \bar{\Gamma}_{r,i,-} = \{i\omega \mid -\omega_i - r > \omega > -\omega_{i+1} + r\},$$

and also define

$$\bar{\Gamma}_{r,i,+} = \{i\omega \mid \omega_k + r < \omega < R\}, \quad \bar{\Gamma}_{r,i,-} = \{i\omega \mid -\omega_k - r > \omega > -R\}.$$

Finally we define

$$\Gamma_R = \left\{ Re^{i\theta} \;\middle|\; -\tfrac{\pi}{2} \leq \theta \leq \tfrac{\pi}{2} \right\}.$$

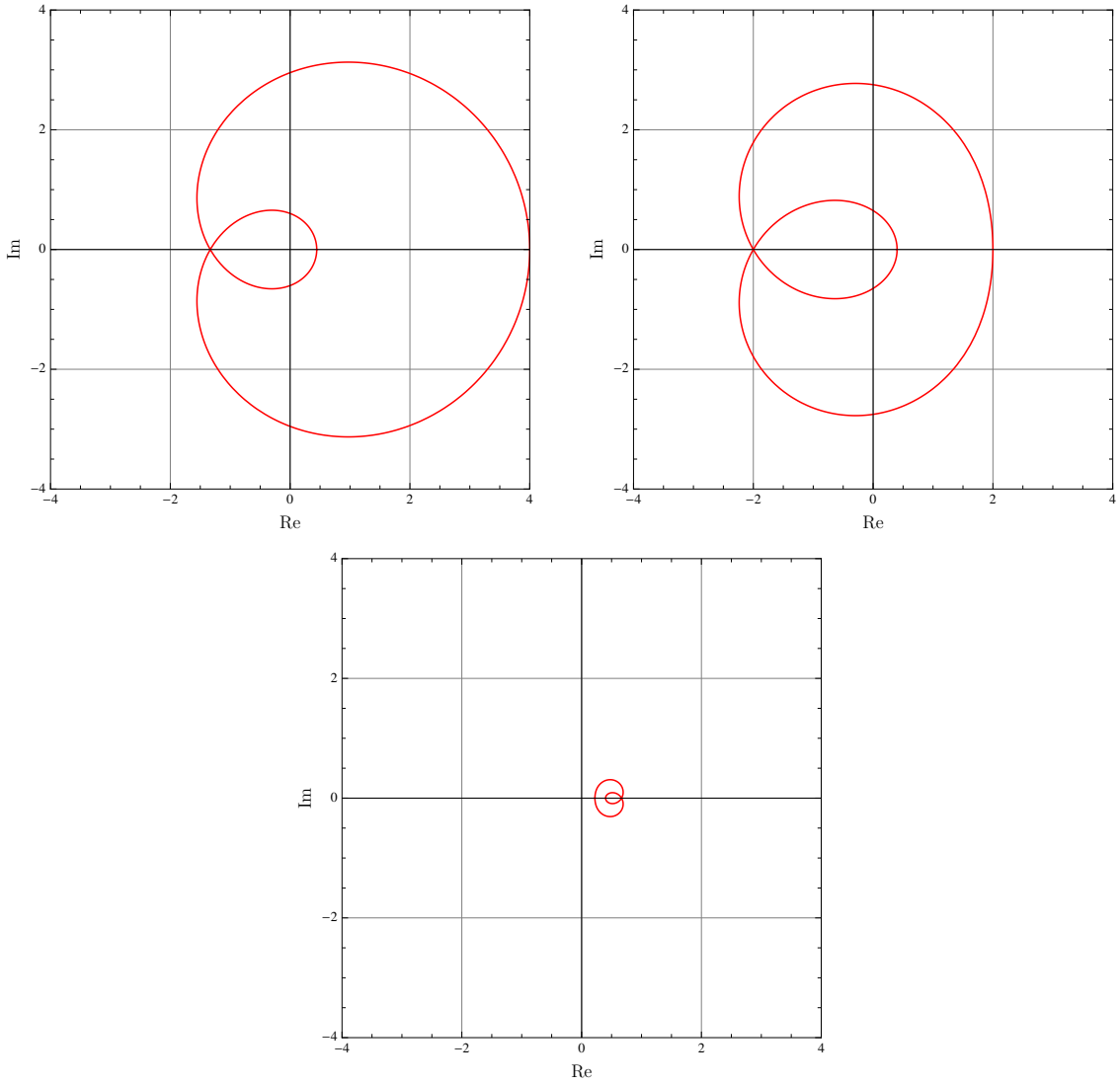


Figure 7.1 Images of closed contours for $a = 0$ (top left), $a = 0.5$ (top right), and $a = 1.5$ (bottom)

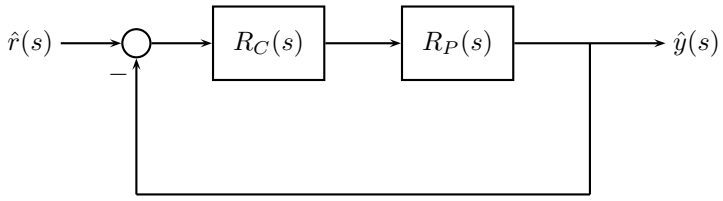


Figure 7.2 A unity feedback loop

The union of all these various contours we denote by $\Gamma_{R,r}$:

$$\Gamma_{R,r} = \bigcup_{i=1}^{k} (\Gamma_{r,i,+} \cup \Gamma_{r,i,-}) \bigcup_{i=1}^{k} (\bar{\Gamma}_{r,i,+} \cup \bar{\Gamma}_{r,i,-}) \bigcup \Gamma_R.$$

When R_L has no poles on the imaginary axis, for $R > 0$ we write

$$\Gamma_{R,0} = \{i\omega \mid \ -R < \omega < R\} \bigcup \Gamma_R.$$

Note that the orientation of the contour $\Gamma_{R,r}$ is taken by convention to be positive in the clockwise direction. This is counter to the complex variable convention, and we choose this convention because, for reasons will soon see, we wish to move along the positive imaginary axis from bottom to top. In any case, the idea is that we have a semicircular contour extending into $\mathbb{C}_+$, and we need to make provisions for any poles of R_L which lie on the imaginary axis. The situation is sketched in Figure 7.3. With this notion of a contour behind

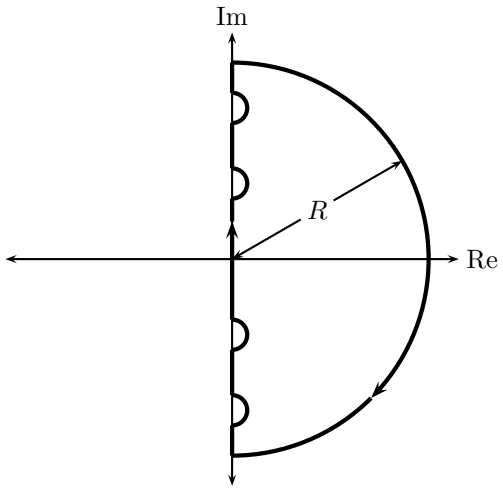


Figure 7.3 The contour $\Gamma_{R,r}$

us, we can define what we will call the Nyquist contour.

7.4 Definition For the unity feedback loop of Figure 7.2, let R_L be the rational function $R_C R_P$, and let R and r satisfy the conditions (7.3) and (7.2). The ***(R, r)-Nyquist contour*** is the contour $R_L(\Gamma_{R,r}) \subset \mathbb{C}$. We denote the (R, r)-Nyquist contour by $\mathcal{N}_{R,r}$. •

When we are willing to live with the associated imprecision, we shall often simply say "Nyquist contour" in place of "(R, r)-Nyquist" contour.

Let us first state some general properties of the (R, r)-Nyquist contour. At the same time we introduce some useful notation. Since we are interested in using the Nyquist criterion for determining IBIBO stability, we shall suppose R_L to be proper, as in most cases we encounter.

7.5 Proposition *Let R_L be a proper rational function, and for $\delta > 0$ let $\bar{D}(0, \delta) = \{s \in \mathbb{C} \mid \ |s| \leq \delta\}$ be the disk of radius δ centred at the origin in $\mathbb{C}$. The following statements hold.*

(i) If R_L has no poles on the imaginary axis then there exists $M > 0$ so that for any $R > 0$ the $(R, 0)$-Nyquist contour is contained in the disk $\bar{D}(0, M)$. Furthermore $\lim_{R\to\infty} \mathcal{N}_{R,0}$ is well-defined and we denote the limit by $\boldsymbol{\mathcal{N}_{\infty,0}}$.

(ii) If R_L is strictly proper, then for any $r > 0$ satisfying (7.2) and for any $\epsilon > 0$ there exists $R_0 > 0$ so that $\mathcal{N}_{R,r} \setminus \mathcal{N}_{R_0,r} \subset \bar{D}(0, \epsilon)$ for any $R > R_0$. Thus $\lim_{R\to\infty} \mathcal{N}_{R,r}$ is well-defined and we denote the limit by $\boldsymbol{\mathcal{N}_{\infty,r}}$.

(iii) If R_L is both strictly proper and has no poles on the imaginary axis, then the consequences of (ii) hold with $r = 0$, and we denote by $\mathbf{N_{\infty,0}}$ the limit $\lim_{R\to\infty} \mathcal{N}_{R,0}$.

Proof (i) Define $R(s) = R_L(\frac{1}{s})$ for $s \neq 0$. Since R_L is proper, the limit $\lim_{s\to 0} R(s)$ exists. But, since R_L is continuous, this is nothing more than the assertion we are trying to prove.

(ii) If R_L is strictly proper then $\lim_{s\to\infty} R_L(s) = 0$. Therefore, by continuity of R_L we can choose R_0 sufficiently large that, for $R > R_0$, those points which lie in the (R, r)-Nyquist contour but do not lie in the (R_0, r)-Nyquist contour reside in the disk $\bar{D}(0, \epsilon)$. This is precisely what we have stated.

(iii) This is a simple consequence of (i) and (ii). ■

The punchline here is that the Nyquist contour is always bounded for proper loop gains, provided that there are no poles on the imaginary axis. When there *are* poles on the imaginary axis, then the Nyquist contour will be unbounded, but for any fixed $r > 0$ sufficiently small, we may still consider letting $R \to \infty$.

Let us see how the character of the Nyquist contour relates to stability of the closed-loop system depicted in Figure 7.2.

7.6 Theorem **(Nyquist Criterion)** *Let R_C and R_P be rational functions with $R_L = R_C R_P$ proper. Let n_p be the number of poles of R_L in $\mathbb{C}_+$. First suppose that $1+R_L$ has no zeros on $\mathrm{i}\mathbb{R}$. Then the interconnected SISO linear system represented by the block diagram Figure 7.2 is IBIBO stable if and only if*

(i) there are no cancellations of poles and zeros in $\overline{\mathbb{C}}_+$ between R_C and R_P;

(ii) $\lim_{s\to\infty} R_L(s) \neq -1$;

(iii) there exists $R_0, r_0 > 0$ satisfying (7.3) and (7.2) with the property that for every $R > R_0$ and $r < r_0$, the (R, r)-Nyquist contour encircles the point $-1 + i0$ in the complex plane n_p times in the counterclockwise direction as the contour $\Gamma_{R,r}$ is traversed once in the clockwise direction.

Furthermore, if for any R and r satisfying (7.3) and (7.2) the (R, r)-Nyquist contour passes through the point $-1 + \mathrm{i}0$, and in particular if $1 + R_L$ has zeros on $\mathrm{i}\mathbb{R}$, then the closed-loop system is IBIBO unstable.

Proof We first note that the condition (ii) is simply the condition that the closed-loop transfer function be proper. If the closed-loop transfer function is not proper, then the resulting interconnection cannot be IBIBO stable.

By Theorem 6.38, the closed-loop system is IBIBO stable if and only if (1) all the closed-loop transfer function is proper, (2) the zeros of the determinant $1 + R_L$ are in $\mathbb{C}_-$, and (3) there are no cancellations of poles and zeros in $\overline{\mathbb{C}}_+$ between R_C and R_P. Thus the first statement in the theorem will follow if we can show that, when $1 + R_L$ has no zeros on $\mathrm{i}\mathbb{R}$, the condition (iii) is equivalent to the condition

(iv) all zeros of the determinant $1 + R_L$ are in $\mathbb{C}_-$.

Since there are no poles or zeros of $1 + R_L$ on $\Gamma_{R,r}$, provided that $R > R_0$ and $r < r_0$, we can apply Proposition 7.2 to the contour $\Gamma_{R,r}$ and the function $1 + R_L$. The conclusion is that the image of $\Gamma_{R,r}$ under $1 + R_L$ encircles the origin $n_z - n_p$ times, with n_z being the number of zeros of $1 + R_L$ in $\mathbb{C}_+$ and n_p being the number of poles of $1 + R_L$ in $\mathbb{C}_+$. Note that the poles of $1 + R_L$ are the same as the poles of R_L, so n_p is the same as in the statement of the theorem. The conclusion in this case is that $n_z = 0$ if and only if the image of $\Gamma_{R,r}$ under $1 + R_L$ encircles the origin n_p times, with the opposite orientation of $\Gamma_{R,r}$. This, however, is

equivalent to the image of $\Gamma_{R,r}$ encircling $-1+\mathrm{i}0$ n_p times, with the opposite orientation of $\Gamma_{R,r}$.

Finally, if the (R,r)-Nyquist contour passes through the point $-1+i0$, this means that the contour $1+R_L(\Gamma_{R,r})$ passes through the origin. Thus this means that there is a point $s_0 \in \Gamma_{R,r}$ which is a zero of $1+R_L$. However, since all points on $\Gamma_{R,r}$ are in $\overline{\mathbb{C}}_+$, the result follows by Theorem 6.38. ■

Let us make a few observations before working out a few simple examples.

7.7 Remarks

1. Strictly proper rational functions *always* satisfy the condition (ii).
2. Of course, the matter of producing the Nyquist contour may not be entirely a straightforward one. What one can certainly do is produce it with a computer. As we will see, the Nyquist contour provides a graphical representation of some important properties of the closed-loop system.
3. By parts (i) and (ii) of Proposition 4.13 it suffices to plot the Nyquist contour only as we traverse that half of $\Gamma_{R,r}$ which sits in the positive imaginary plane, i.e., only for those values of s along $\Gamma_{R,r}$ which have positive imaginary part. This will be borne out in the examples below.
4. When R_L is proper and when there are no poles for R_L on the imaginary axis (so we can take $r=0$), the $(R,0)$-Nyquist contour is bounded as we take the limit $R \to \infty$. If we further ask that R_L be *strictly* proper, that portion of the Nyquist contour which is the image of Γ_R under R_L will be mapped to the origin as $R \to \infty$. Thus in this case it suffices to determine the image of the imaginary axis under R_L, along with the origin in $\mathbb{C}$. Given our remark 3, this essentially means that in this case we only determine the polar plot for the loop gain R_L. Thus we see the important relationship between the Nyquist criterion and the Bode plot of the loop gain.
5. Here's one way to determine the number of times the (R,r)-Nyquist contour encircles the point $-1+i0$. From the point $-1+i0$ draw a ray in *any* direction. Choose this ray so that it is nowhere tangent to the (R,r)-Nyquist contour, and so that it does not pass through points where the (R,r)-Nyquist contour intersects itself. The number of times the (R,r)-Nyquist contour intersects this ray while moving in the counterclockwise direction is the number of counterclockwise encirclements of $-1+i0$. A crossing in the clockwise direction is a negative counterclockwise crossing. •

The Nyquist criterion can be readily demonstrated with a couple of examples. In each of these examples we use Remark 7.7–(5).

In the Nyquist plots below, the solid contour is the image of points in the positive imaginary plane under R_L, and the dashed contour is the image of the points in the negative imaginary plane.

7.8 Examples

1. We first take $R_C(s)=1$ and $R_P(s)=\frac{1}{s+a}$ for $a \in \mathbb{R}$. Note that conditions (i) and (ii) of Theorem 7.6 are satisfied for all a, so stability can be check by verifying the condition (iii). We note that for $a<0$ there is one pole of R_L in $\mathbb{C}_+$, and otherwise there are no poles in $\mathbb{C}_+$.

   The loop gain $R_L(s)=\frac{1}{s+a}$ is strictly proper with no poles on the imaginary axis unless $a=0$. So let us first consider the situation when $a \neq 0$. We need only consider

the image under R_L of points on the imaginary axis. The corresponding points on the Nyquist contour are given by

$$\frac{1}{i\omega + a} = \frac{a}{\omega^2 + a^2} - i\frac{\omega}{\omega^2 + a^2}, \quad \omega \in (-\infty, \infty).$$

This is a parametric representation of a circle of radius $\frac{1}{2|a|}$ centred at $\frac{1}{2a}$. This can be checked by verifying that

$$\Big(\frac{a}{\omega^2 + a^2} - \frac{1}{2a}\Big)^2 + \Big(\frac{\omega}{\omega^2 + a^2}\Big)^2 = \frac{1}{4a^2}.$$

The Nyquist contour is shown in Figure 7.4 for various nonzero a. From Figure 7.4 we make the following observations:

(a) for $a < -1$ there are no encirclements of the point $-1 + i0$;
(b) for $a = -1$ the Nyquist contour passes through the point $-1 + i0$;
(c) for $-1 < a < 0$ the Nyquist contour encircles the point $-1 + i0$ one time in the counterclockwise direction (to see this, one must observe the sign of the imaginary part as ω runs from $-\infty$ to $+\infty$);
(d) for $a > 0$ there are no encirclements of the point $-1 + i0$.

Now let us look at the case where $a = 0$. In this case we have a pole for R_L at $s = 0$, so this must be taken into account. Choose $r > 0$. The image of $\{i\omega \mid \omega > r\}$ is

$$\big\{ -i\omega \;\big|\; 0 < \omega < \tfrac{1}{r} \big\}.$$

Now we need to look at the image of the contour Γ_r given by $s = re^{i\theta}$, $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. One readily sees that the image of Γ_r is

$$\frac{e^{-i\theta}}{r}, \quad \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$$

which is a large semi-circle centred at the origin going from $+i\infty$ to $-\infty$ in the clockwise direction. This is shown in Figure 7.5.

This allows us to conclude the following:

(a) for $a < -1$ the system is IBIBO unstable since $n_p = 1$ and the number of counterclockwise encirclements is -1;
(b) for $a = -1$ the system is IBIBO unstable since the Nyquist contour passes through the point $-1 + i0$;
(c) for $-1 < a < 0$ the system is IBIBO stable since $n_p = 1$ and there is one counterclockwise encirclement of the point $-1 + i0$;
(d) for $a \geq 0$ the system is IBIBO stable since $n_p = 0$ and there are no encirclements of the point $-1 + i0$.

We can also check this directly by using Theorem 6.38. Since there are no unstable pole/zero cancellations in the interconnected system, IBIBO stability is determined by the zeros of the determinant, and the determinant is

$$1 + R_L(s) = \frac{s + a + 1}{s + a}.$$

The zero of the determinant is $-a - 1$ which is in $\mathbb{C}_-$ exactly when $a > -1$, and this is exactly the condition we derive above using the Nyquist criterion.

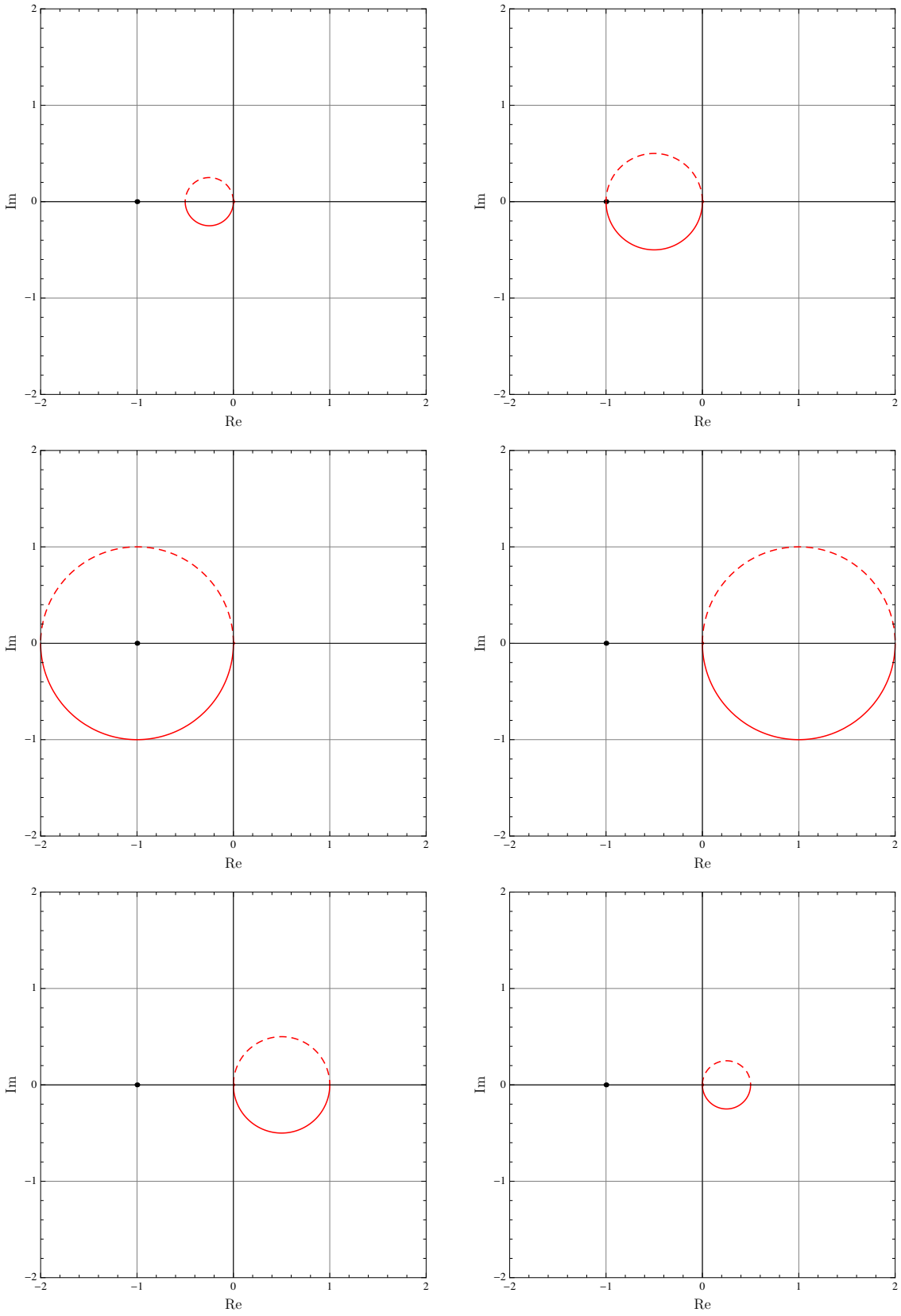


Figure 7.4 The $(\infty, 0)$-Nyquist contour for $R_L(s) = \frac{1}{s+a}$, $a = -2$ (top left), $a = -1$ (top right), $a = -\frac{1}{2}$ (middle left), $a = \frac{1}{2}$ (middle right), $a = 1$ (bottom left), and $a = 2$ (bottom right)

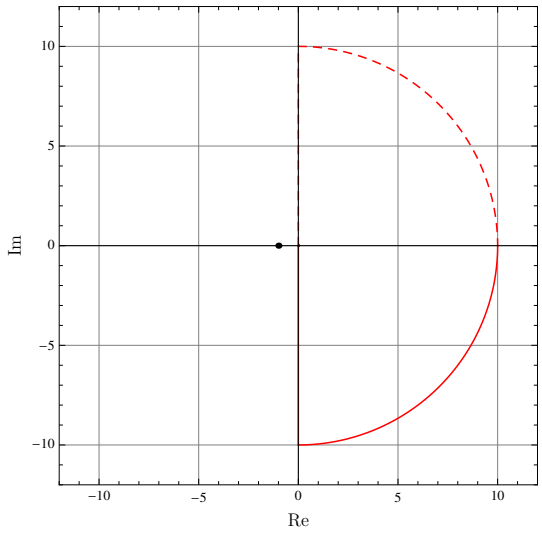


Figure 7.5 The $(\infty, 0.1)$-Nyquist contour for $R_L(s) = \frac{1}{s}$

2. The previous example can be regarded as an implementation of a proportional controller. Let's give an integral controller a try. Thus we take $R_C(s) = \frac{1}{s}$ and $R_P(s) = \frac{1}{s+a}$. Once again, the conditions (i) and (ii) of Theorem 7.6 are satisfied, so we need only check condition (iii).
In this example the loop gain R_L is strictly proper, and there is a pole of R_L at $s = 0$. Thus we need to form a modified contour to take this into account. Let us expand the loop gain into its real and imaginary parts when evaluated on the imaginary axis away from the origin. We have

$$R_L(i\omega) = -\frac{1}{\omega^2 + a^2} - i\frac{a}{\omega(\omega^2 + a^2)}.$$

Let us examine the image of $\{i\omega \mid \omega > r\}$ as r becomes increasingly small. For ω near zero, the real part of the Nyquist contour is near $-\frac{1}{\omega^2+a^2}$, and as ω increases, it shrinks to zero. Near $\omega = 0$ the imaginary part is at $-\text{sgn}(a)\infty$, and as ω increases, it goes to zero. Also note that the imaginary part does not change sign. The image of $\{i\omega \mid \omega < -r\}$ reflects this about the real axis. It only remains to examine the image of the contour Γ_r around $s = 0$ given by $s = re^{i\theta}$ where $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. The image of this contour under R_L is

$$\frac{1}{re^{i\theta}(re^{i\theta} + a)}, \quad \theta \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}].$$

For sufficiently small r we have

$$\begin{aligned}\frac{1}{re^{i\theta}(re^{i\theta} + a)} &= \frac{1}{re^{i\theta}} \frac{\frac{1}{a}}{1 + \frac{r}{a}e^{i\theta}} \\ &= \frac{1}{re^{i\theta}} \frac{1}{a}\Big(1 - \frac{r}{a}e^{i\theta} + \cdots\Big) \\ &= \frac{e^{-i\theta}}{ar} - \frac{1}{a^2} + \cdots.\end{aligned}$$

Thus, as $r \to 0$, the contour Γ_r gets mapped into a semi-circle of infinite radius, centred at $\frac{1}{a^2} + i0$, which goes clockwise from $\frac{1}{a^2} + i\text{sgn}(a)\infty$ to $\frac{1}{a^2} - i\text{sgn}(a)\infty$. In Figure 7.6 we

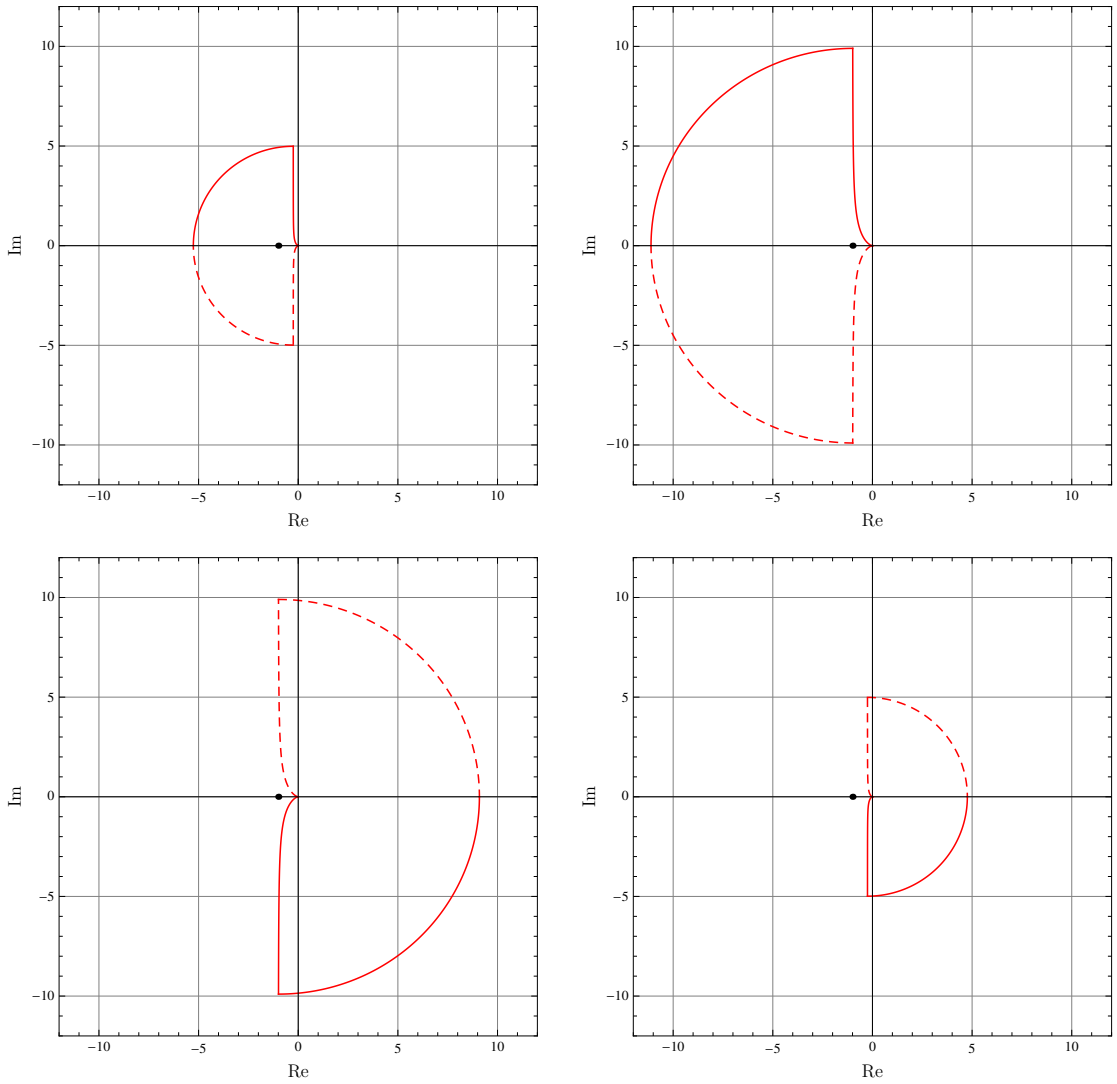


Figure 7.6 The $(\infty, 0.1)$-Nyquist contours for $R_L(s) = \frac{1}{s(s+a)}$, $a = -2$ (top left), $a = -1$ (top right), $a = 1$ (bottom left), and $a = 2$ (bottom right)

plot the Nyquist contours.

Now we look at the particular case when $a = 0$. In this case $R_L(s) = \frac{1}{s^2}$, and so the Nyquist contour is the image of the imaginary axis, except the origin, under R_L, along with the image of the contour Γ_r as $r \to 0$. On the imaginary axis we have $R_L(i\omega) = -\frac{1}{\omega^2}$. As ω goes from $-\infty$ to 0_- the Nyquist contour goes from $0_- + i0$ to $-\infty + i0$, and as ω goes from 0_+ to $+\infty$ the Nyquist contour goes from $-\infty + i0$ to $0_- + i0$. On the contour Γ_r we have

$$R_L(re^{i\theta}) = \frac{e^{-2i\theta}}{r^2}, \quad \theta \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}].$$

As $r \to 0$ this describes an infinite radius circle centred at the origin which starts at $-\infty + i0$ and goes around once in the clockwise direction. In particular, when $a = 0$ the Nyquist contour passes through the point $-1 + i0$. The contour is shown in Figure 7.7.

We can now make the following conclusions regarding stability:

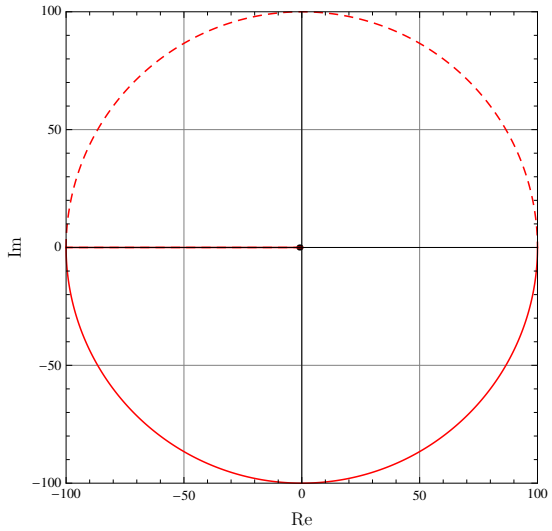


Figure 7.7 The $(\infty, 0.1)$-Nyquist contour for $R_L(s) = \frac{1}{s^2}$

(a) when $a < 0$ the system is IBIBO unstable since $n_p = 1$ and there is one clockwise encirclement of the point $-1 + i0$;

(b) when $a = 0$ the system is IBIBO unstable since the Nyquist contour passes through the point $-1 + i0$;

(c) when $a > 0$ the system is IBIBO stable since $n_p = 0$ and there are no encirclements of the point $-1 + i0$.

We can also check IBIBO stability of the system via Theorem 6.38. Since there are no unstable pole/zero cancellations, we can look at the zeros of the determinant which is

$$1 + R_L(s) = \frac{s^2 + as + 1}{s^2 + as}.$$

By the Routh/Hurwitz criterion, the system is IBIBO stable exactly when $a > 0$, and this is what we ascertained using the Nyquist criterion.

3. Our final example combines the above two examples, and implements a PI controller where we take $R_C(s) = 1 + \frac{1}{s}$ and $R_P(s) = \frac{1}{s+a}$. Here the condition (ii) of Theorem 7.6 holds, but we should examine (i) just a bit carefully before moving on. We have $R_L(s) = \frac{s+1}{s}\frac{1}{s+a}$, and so there is a pole/zero cancellation here when $a = 1$. However, it is a stable pole/zero cancellation, so condition (i) still holds. We also ascertain that R_L has one pole in $\mathbb{C}_+$ when $a < 0$.

In this example, R_L is again strictly proper, and has a pole on the imaginary axis at the origin. Thus to compute the Nyquist contour we determine the image of the imaginary axis minus the origin, and tack on the image of the contour $\Gamma_r = re^{i\theta}$, $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. We first compute

$$R_L(i\omega) = \frac{a-1}{\omega^2 + a^2} - i\frac{\omega^2 + a}{\omega(\omega^2 + a^2)}.$$

Let's first consider the situation when $a \neq 0$. In this case, as ω goes from $-i\infty$ to 0_-, the real part goes from 0 to $\frac{a-1}{a^2}$ and the imaginary part goes from 0 to $-\text{sgn}(a)\infty$. While the real part never changes sign, the imaginary part *can* change sign, so we must take care about how it is behaving. As ω goes from 0_+ to $+\infty$, the resulting part of the Nyquist

contour is the mirror image about the real axis of what we have already computed. Now let us look at the image of the contour Γ_r. When $a \neq 0$ we have

$$\begin{aligned} R_L(re^{i\theta}) &= \frac{re^{i\theta}+1}{re^{i\theta}} \frac{\frac{1}{a}}{1+\frac{r}{a}e^{i\theta}} \\ &= \frac{re^{i\theta}+1}{are^{i\theta}} \Big(1 - \frac{r}{a}e^{i\theta} + \cdots\Big) \\ &= \frac{1}{a} + \frac{e^{-i\theta}}{ar} - \frac{re^{i\theta}}{a^2} - \frac{1}{a^2} + \cdots \\ &= \frac{e^{-i\theta}}{ar} + \frac{a-1}{a^2} + \cdots \end{aligned}$$

From this we conclude that for $a \neq 0$ the image as $r \to 0$ of Γ_r under R_L is an infinite radius semi-circle centered at $\frac{a-1}{a^2}$ and going clockwise from $\frac{a-1}{a^2} + i\mathrm{sgn}(a)\infty$ to $\frac{a-1}{a^2} - i\mathrm{sgn}(a)\infty$. In Figure 7.8 we show the Nyquist contours for various values of $a \neq 0$.

Now let us consider the image of the imaginary axis when $a = 0$. In this case we have

$$R_L(i\omega) = -\frac{1}{\omega^2} - i\frac{1}{\omega}.$$

Thus the image of those points on the imaginary axis away from the origin describe a parabola, sitting in $\overline{\mathbb{C}}_-$, passing through the origin, and symmetric about the real axis. As concerns Γ_r when $a = 0$ we have

$$R_L(re^{i\theta}) = \Big(1 + \frac{e^{-i\theta}}{r}\Big)\frac{e^{-i\theta}}{r}$$

which, as $r \to 0$, describes an infinite radius circle centred at the origin and going clockwise from $-\infty + i0$ to $-\infty + i0$. This Nyquist contour is shown in Figure 7.9.

With the above computations and the corresponding Nyquist plots, we can make the following conclusions concerning IBIBO stability of the closed-loop system.

(a) For $a < -1$ the system is IBIBO unstable since $n_p = 1$ and there is one clockwise encirclement of $-1 + i0$.

(b) For $a = -1$ the system is IBIBO unstable since the Nyquist contour passes through the point $-1 + i0$.

(c) For $-1 < a < 0$ the system is IBIBO stable since $n_p = 1$ and there is one counter-clockwise encirclement of the point $-1 + i0$.

(d) For $a \geq 0$ the system is IBIBO stable since $n_p = 0$ and there are no encirclements of $-1 + i0$.

As always, we can evaluate IBIBO stability with Theorem 6.38. To do this we can still simply look at the zeros of the determinant, because although there *is* a pole zero cancellation when $a = 1$, it is a cancellation of stable factors so it does not hurt us. We compute the determinant to be

$$1 + R_L(s) = \frac{s^2 + (a+1)s + 1}{s^2 + as}.$$

An application of the Routh/Hurwitz criterion suggests that we have IBIBO stability for $a > -1$, just as we have demonstrated with the Nyquist criterion. •

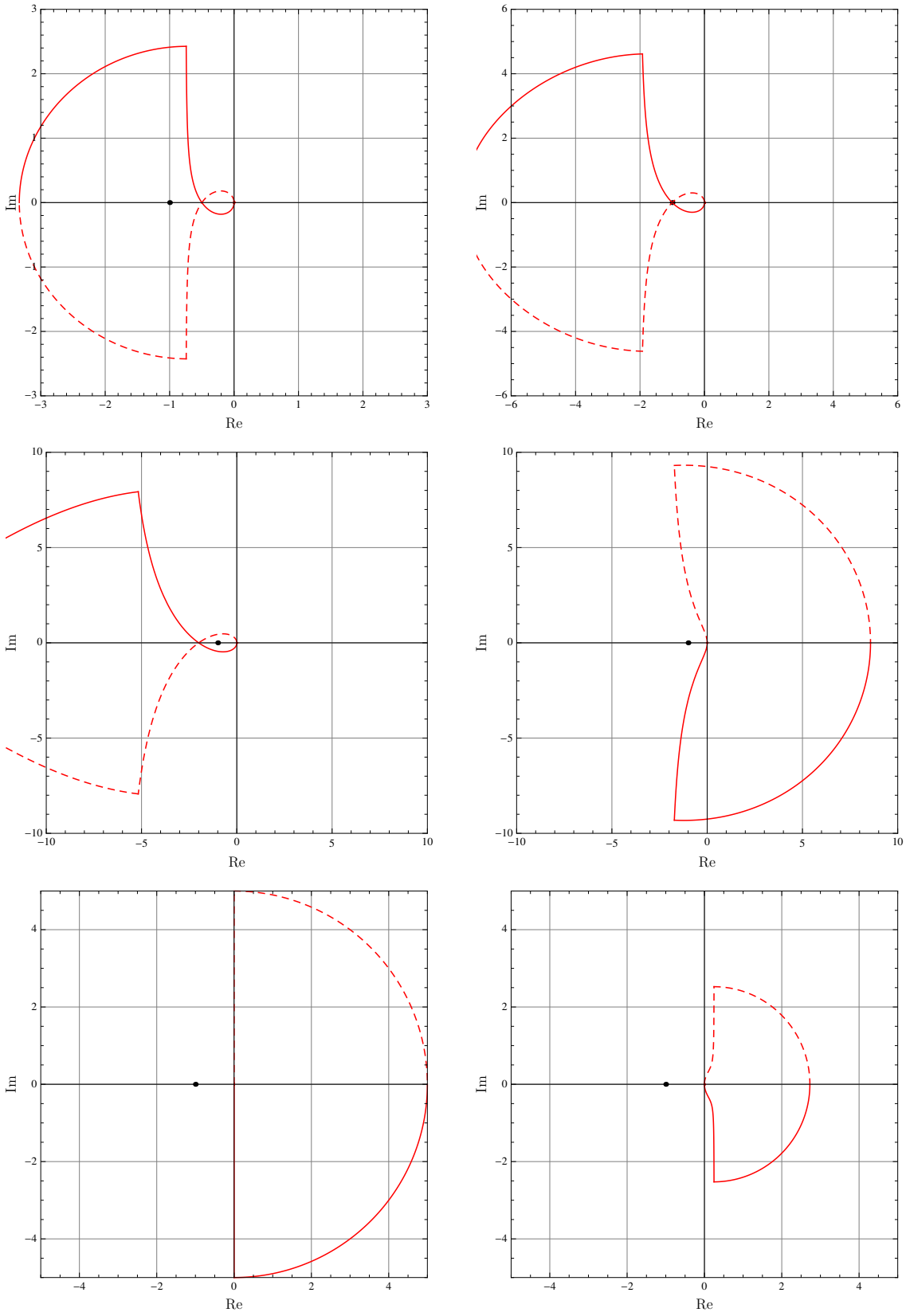


Figure 7.8 The $(\infty, 0.2)$-Nyquist contours for $R_L(s) = \frac{s+1}{s(s+a)}$, $a = -2$ (top left), $a = -1$ (top right), $a = -\frac{1}{2}$ (middle left), $a = \frac{1}{2}$ (middle right), $a = 1$ (bottom left), and $a = 2$ (bottom right)

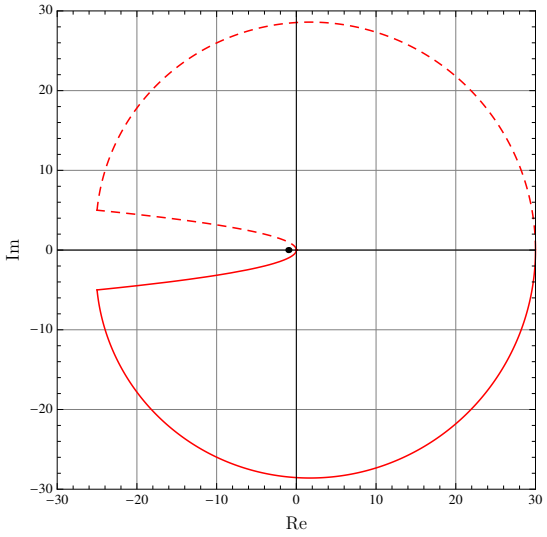


Figure 7.9 The $(\infty, 0.2)$-Nyquist contour for $R_L(s) = \frac{s+1}{s^2}$

Note that the Nyquist criterion as we have shown is applicable only to interconnections with a simple structure, namely a single loop. Bode [1945] discusses a version of the Nyquist criterion for systems with multiple loops, and this is explored further by Zadeh and Desoer [1979]. However, the development is too significant, and the outcome too modest (what are obtained are sufficient conditions for IBIBO stability under restrictive hypotheses) to make a presentation of these results worthwhile.

7.2 The relationship between the Nyquist contour and the Bode plot

The above examples, although simple, demonstrate that obtaining the Nyquist contour can be problematic, at least by hand. This is especially well illustrated by the third of the three examples where the capacity to change sign of the imaginary part of the restriction of R_L to the imaginary axis causes some difficulties which must be accounted for. A useful observation here is that the Nyquist contour is in essence the polar plot for the loop gain, taking care of the possibility of poles on the imaginary axis. The matter of constructing a Bode plot is often an easier one than that of building the corresponding polar plot, so a plausible approach for making a Nyquist contour is to first make a Bode plot, and convert this to a polar plot as discussed in Section 4.3.3.

7.2.1 Capturing the essential features of the Nyquist contour from the Bode plot

Let us illustrate this with the third of our examples from the previous section.

7.9 Example **(Example 7.8–3 cont'd)** The loop gain, recall, is $R_L(s) = \frac{s+1}{s(s+a)}$. Let us write this transfer function in the recommended form for making Bode plots. For $a \neq 0$ we have

$$R_L(s) = \frac{1}{a}\frac{1}{s}\frac{1}{\frac{s}{a}+1}(s+1).$$

Thus the frequency response for R_L is a product of four terms:

$$H_1(\omega) = \frac{1}{a}, \quad H_2(\omega) = -\frac{i}{\omega}, \quad H_3(\omega) = \frac{1}{1+i\frac{\omega}{a}}, \quad H_4(\omega) = 1 + i\omega.$$

Each of these is a simple enough function for the purpose of plotting frequency response, and the effect of a is essentially captured in H_3.

Let us see if from Bode plot considerations we can infer when the imaginary part of the transfer function changes sign as ω goes from 0_+ to $+\infty$. In Example 7.8–3 we determined that as we vary ω in this way, the real part of the frequency response goes from $\frac{a-1}{a^2}$ to 0 and the imaginary part goes from $\operatorname{sgn}(a)\infty$ to 0. The question is, "For which values of a does the imaginary part of the frequency response change sign as ω goes from 0_+ to $+\infty$?" Provided $\frac{a-1}{a^2} < 0$ this will happen if and only if the phase is $\pm\pi$ at some finite frequency $\bar{\omega}$. Let's take a look at the phase of the frequency response function.

1. First we take $a < 0$. Since $\angle H_1(\omega) = \pi$ and $\angle H_2(\omega) = -\frac{\pi}{2}$, in order to have the total phase equal $\pm\pi$, it must be the case that

$$\angle H_3(\bar{\omega}) + \angle H_4(\bar{\omega}) \in \{\tfrac{\pi}{2}, -\tfrac{3\pi}{2}\}.$$

   The phase of H_4 varies from 0 to $\frac{\pi}{2}$, and the phase of H_3 varies from 0 to $\frac{\pi}{2}$ (because $a < 0$!) Therefore we should aim for conditions on when $\angle H_3(\bar{\omega}) + \angle H_4(\bar{\omega}) = \frac{\pi}{2}$ for some finite frequency $\bar{\omega}$. But one easily sees that for *any* $a < 0$ there will always be a finite frequency $\bar{\omega}$ so that this condition is satisfied since

$$\lim_{\omega\to\infty} \big(\angle H_3(\omega) + \angle H_4(\omega)\big) = \pi.$$

   Thus as long as $a < 0$ there will always be a sign change in the imaginary part of the frequency response as we vary ω from 0_+ to $+\infty$. The Bode plots for R_L are shown in Figure 7.10 for various $a < 0$.

2. Now we consider when $a > 0$. In this case we have $\angle H_1(\omega) = 0$ and so we must seek $\bar{\omega}$ so that

$$\angle H_3(\bar{\omega}) + \angle H_4(\bar{\omega}) \in \{-\tfrac{\pi}{2}, \tfrac{3\pi}{2}\}.$$

   However, since $a > 0$ the phase of H_3 will go from 0 to $-\frac{\pi}{2}$. Therefore it will not be possible for $\angle H_3(\omega) + \angle H_4(\omega)$ to equal either $-\frac{\pi}{2}$ or $\frac{3\pi}{2}$, and so we conclude that for $a > 0$ the imaginary part of $R_L(i\omega)$ will not change sign as we vary ω from 0_+ to $+\infty$. The Bode plots for R_L are shown in Figure 7.11 for various $a > 0$. •

7.2.2 Stability margins

The above example illustrates how the Bode plot can be useful in determining certain aspects of the behaviour of the Nyquist contour. Indeed, if one gives this a short moments' thought, one reaches the realisation that one can benefit a great deal by looking at the Bode plot for the loop gain. Let us provide the proper nomenclature for organising such observations.

7.10 Definition Let $R_L \in \mathbb{R}(s)$ be a proper rational function forming the loop gain for the interconnection of Figure 7.12. Assume that there is no frequency $\omega > 0$ for which $R_L(i\omega) = -1 + i0$.

(i) A ***phase crossover frequency***, $\omega_{\mathrm{pc}} \in [0, \infty)$, for R_L is a frequency for which $\angle R_L(i\omega_{\mathrm{pc}}) = 180°$.

Let $\omega_{\mathrm{pc},1}, \ldots, \omega_{\mathrm{pc},\ell}$ be the phase crossover frequencies for R_L, and assume these are ordered so that

$$R_L(i\omega_{\mathrm{pc},1}) < \cdots < R_L(i\omega_{\mathrm{pc},\ell}).$$

Also suppose that for some $k \in \{1, \ldots, \ell\}$ we have

$$R_L(i\omega_{\mathrm{pc},k}) < -1 < R_L(i\omega_{\mathrm{pc},k+1}).$$

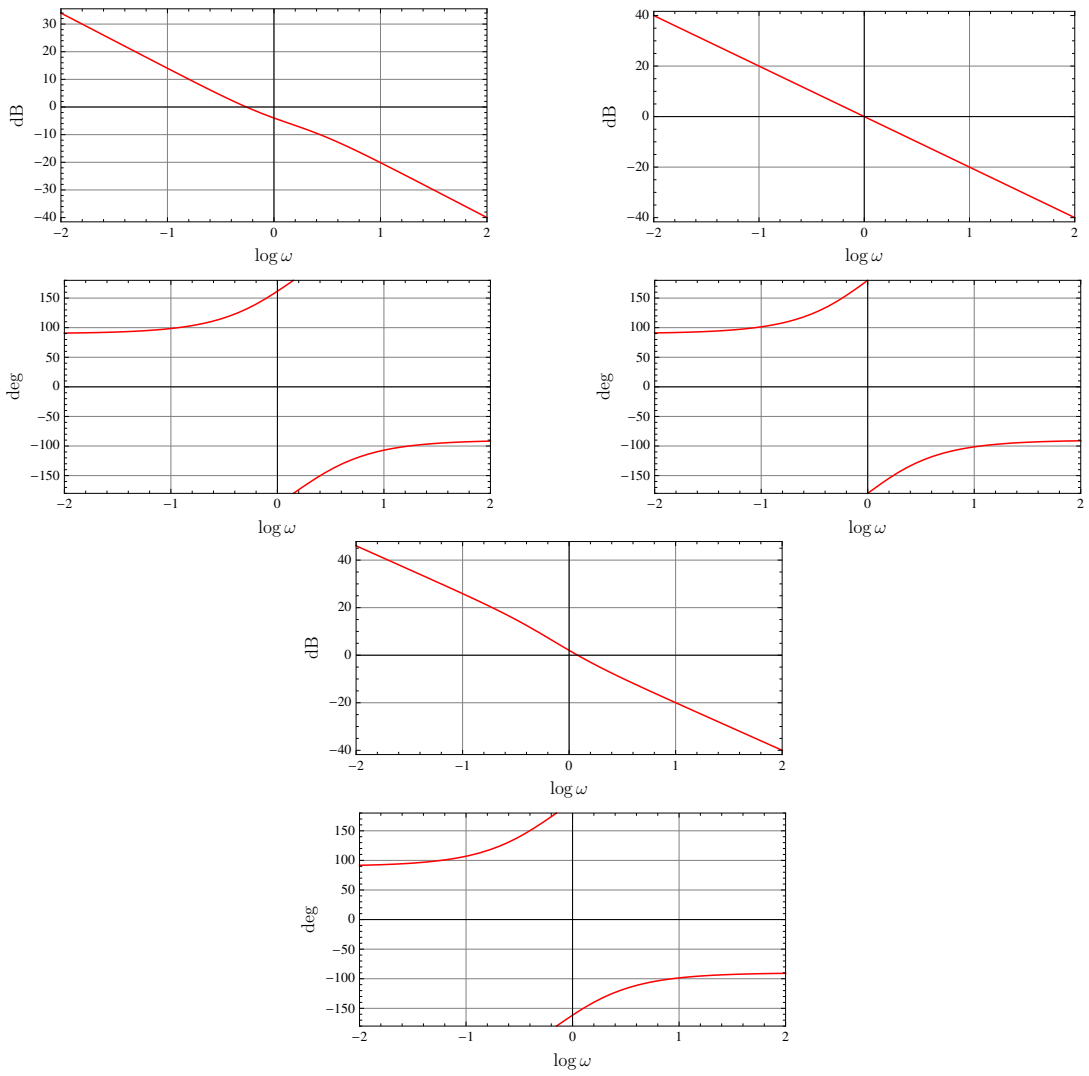


Figure 7.10 Bode plots for $R_L(s) = \frac{s+1}{s(s+a)}$ for $a = -2$ (top left), $a = -1$ (top right), and $a = -\frac{1}{2}$ (bottom)

(ii) The ***lower gain margin*** for R_L defined by

$$K_{\min} = -R_L(i\omega_{\mathrm{pc},k})^{-1} \in (0, 1).$$

If

$$-1 < R_L(i\omega_{\mathrm{pc},1}) < \cdots < R_L(i\omega_{\mathrm{pc},\ell}),$$

then $K_{\min}$ is undefined.

(iii) The ***upper gain margin*** for R_L is defined by

$$K_{\max} = -R_L(i\omega_{\mathrm{pc},k+1})^{-1} \in (1, \infty).$$

If

$$R_L(i\omega_{\mathrm{pc},1}) < \cdots < R_L(i\omega_{\mathrm{pc},\ell}) < -1,$$

then $K_{\max}$ is undefined.

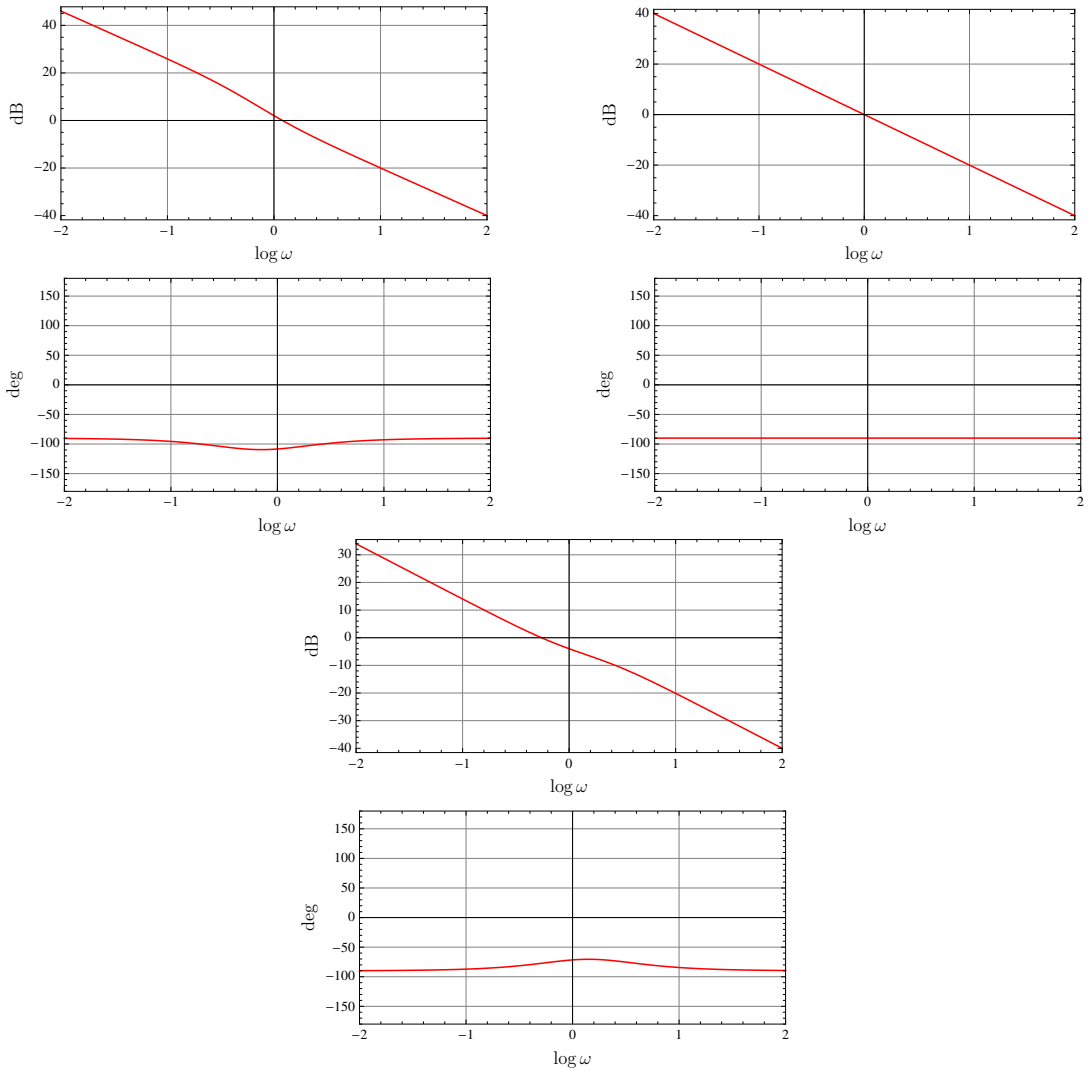


Figure 7.11 Bode plots for $R_L(s) = \frac{s+1}{s(s+a)}$ for $a = \frac{1}{2}$ (top left), $a = 1$ (top right), and $a = 2$ (bottom)

(iv) A ***gain crossover frequency***, $\omega_{\text{gc}} \in [0, \infty)$, for R_L is a frequency for which $|R_L(i\omega_{\text{gc}})| = 1$.

Let $\omega_{\text{gc},1}, \ldots, \omega_{\text{gc},\ell}$ be the gain crossover frequencies for R_L, and assume these are ordered so that

$$\measuredangle(R_L(i\omega_{\text{gc},1})) < \cdots < \measuredangle(R_L(i\omega_{\text{gc},\ell})).$$

(v) The ***lower phase margin*** for R_L is defined by

$$\Phi_{\text{min}} = \begin{cases} 180^\circ - \measuredangle(R_L(i\omega_{\text{gc},\ell})), & \measuredangle(R_L(i\omega_{\text{gc},\ell})) \geq 0 \\ \text{undefined}, & \measuredangle(R_L(i\omega_{\text{gc},\ell})) < 0. \end{cases}$$

(vi) The ***upper phase margin*** for R_L is defined by

$$\Phi_{\text{max}} = \begin{cases} \measuredangle(R_L(i\omega_{\text{gc},1})) + 180^\circ, & \measuredangle(R_L(i\omega_{\text{gc},1})) \leq 0 \\ \text{undefined}, & \measuredangle(R_L(i\omega_{\text{gc},1})) > 0. \end{cases}$$

•

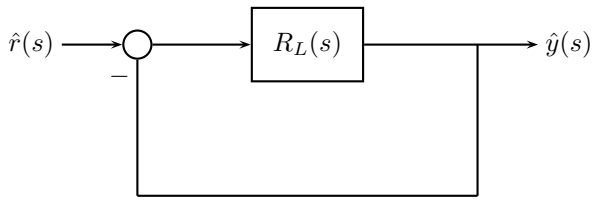


Figure 7.12 Unity gain feedback loop for stability margin discussion

Let us parse these definitions, as they are in actuality quite simple. First of all, we note that it is possible to read the gain and phase margins off the Nyquist plot; this saves one having to compute them directly using the definitions. Rather than try to state in a precise way how to procure the margins from the Nyquist plot, let us simply illustrate the process in Figure 7.13. The basic idea is that for the gain margins, one looks for the positive frequency

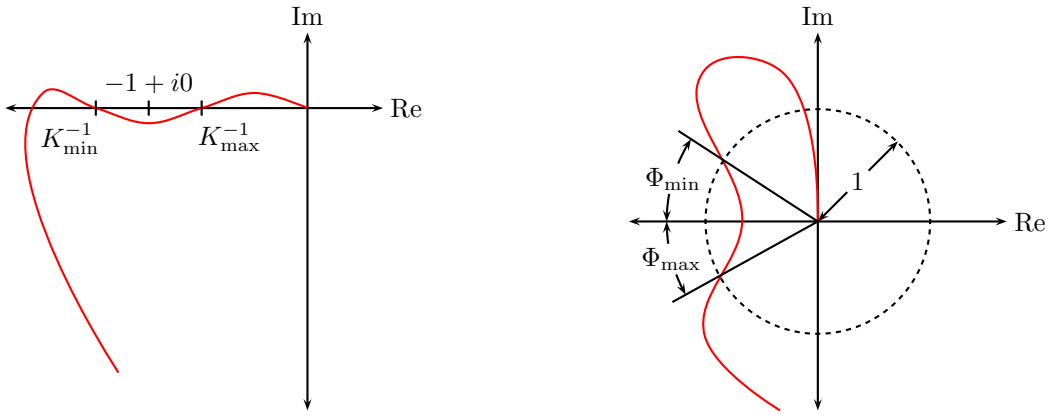


Figure 7.13 Getting gain (left) and phase (right) margins from the Nyquist plot

crossings of the negative real axis closest to $-1+i0$ in each direction. The reciprocal of the distances to the imaginary axis are the gain margins, as indicated in Figure 7.13. For the phase margins, one looks for the positive frequency crossings of the unit circle closest to the point $-1+i0$. The angles to the negative real axis are then the phase margins, again as indicated in Figure 7.13. We shall adopt the convention that when we simply say ***phase margin***, we refer to the smaller of the upper and lower phase margins.

The interpretations of gain crossover and phase crossover frequencies are clear. At a gain crossover frequency, the magnitude on the Bode plot for R_L will be 0dB. At a phase crossover frequency, the graph will cross the upper or lower edge of the phase Bode plot. Note that it is possible that for a given loop gain, some of the margins may not be defined. Let us illustrate this with some examples.

7.11 Examples

1. We consider the loop gain

$$R_L(s) = \frac{10}{s^2+4s+3}.$$

In Figure 7.14 we note that there is one gain crossover frequency, and it is roughly at

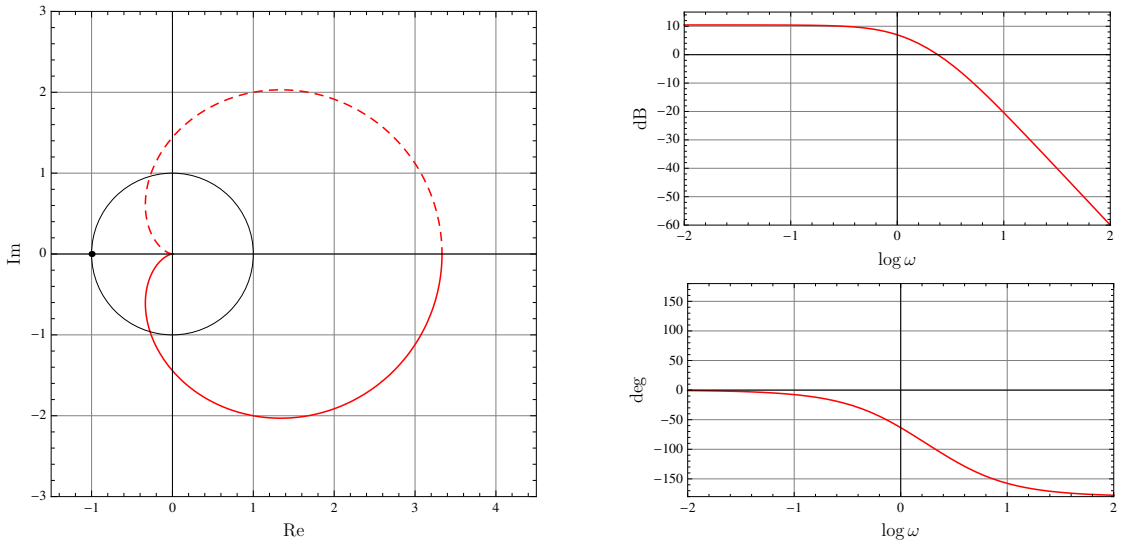


Figure 7.14 Nyquist and Bode plots for $R_L(s) = \frac{10}{s^2+4s+3}$

$\omega_{gc} = 2.1$. The phase at the gain crossover frequency is about $-110°$, which gives the upper phase margin as $\Phi_{max} \approx 70°$. The lower phase margin is not defined. Also, neither of the gain margins are defined.

2. We take as loop gain

$$R_L(s) = -\frac{10}{s^2 + 4s + 3}.$$

The Bode and Nyquist plots are shown in Figure 7.15. From the Bode plot we see

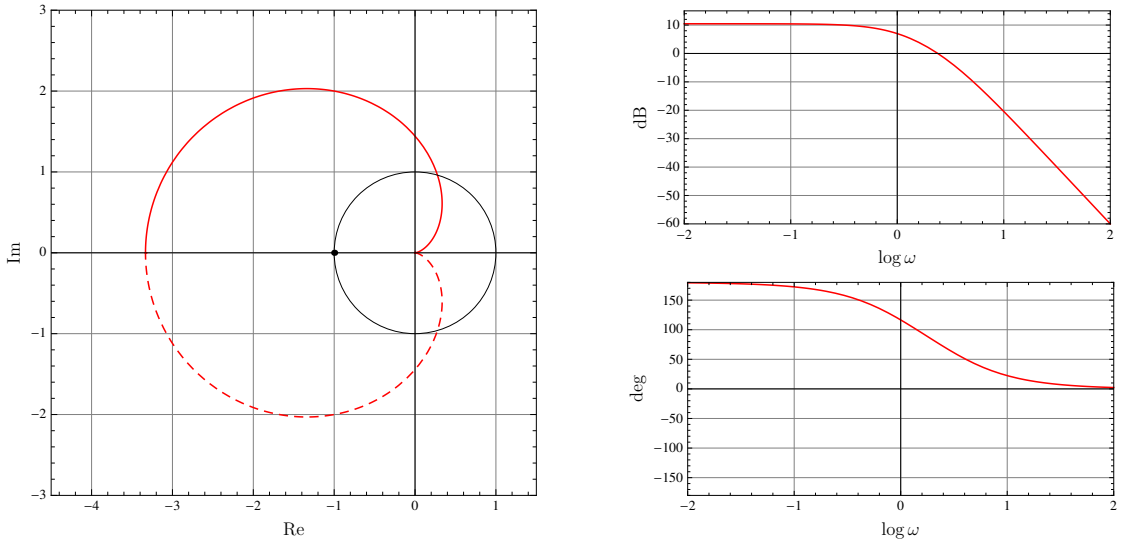


Figure 7.15 Nyquist and Bode plots for $R_L(s) = -\frac{10}{s^2+4s+3}$

that there is one gain crossover frequency and it is approximately at $\omega_{gc} = 2.1$. The phase at the gain crossover frequency it is about $70°$. Thus the lower phase margin is

$\Phi_{\min} \approx 110°$. We note that there the upper phase margin is undefined, as are the phase crossover frequencies.

gain margin

3.

For cases when R_L is BIBO stable, we can offer an interpretation of the gain and phase margins in terms of closed-loop stability.

7.12 Proposition *Let $R_L \in \mathrm{RH}^+_\infty$ be a BIBO stable loop gain, and consider a unity gain feedback block diagram configuration like that of Figure 7.12. Either of the following statements implies IBIBO stability of the closed-loop system:*

(i) $K_{\min}$ *is undefined and either*

(a) $K_{\max} > 1$ *or*

(b) $K_{\max}$ *is undefined.*

(ii) $\Phi_{\min}$ *is undefined and either*

(a) $\Phi_{\max} > 0$.

(b) $\Phi_{\max}$ *is undefined.*

Proof We use the Nyquist criterion for evaluating IBIBO stability. For a stable loop gain R_L, the closed-loop system is IBIBO stable if and only if there are no encirclements of $-1 + i0$. Note that the assumption that $R_L \in \mathrm{RH}^+_\infty$ be BIBO stable implies that there are no poles for R_L on the imaginary axis, so the $(\infty, 0)$-Nyquist contour is well-defined and bounded.

(i) In this case, all crossings of the imaginary axis will occur in the interval $(-1, 0)$. This precludes any encirclements of $-1 + i0$.

(ii) If (ii) holds, then all intersections at positive frequency of the $(\infty, 0)$-Nyquist contour with the unit circle in $\mathbb{C}$ will occur in the lower complex plane. Similarly, those crossings of the unit disk at negative frequencies will take place in the upper complex plane. This clearly precludes any encirclements of $-1 + i0$ which thus implies IBIBO stability. ■

Often when one reads texts on classical control design, one simply sees mentioned "gain margin" and "phase margin" without reference to upper and lower. Typically, a result like Proposition 7.12 is in the back of the minds of the authors, and it is being assumed that the lower margins are undefined. In these cases, there is a more direct link between the stability margins and actual stability—at least when the loop gain is itself BIBO stable—as evidenced by Proposition 7.12. The following examples demonstrate this, and also show that one cannot generally expect the converse of the statements in Proposition 7.12 to hold.

7.13 Examples

1. We consider
$$R_L(s) = \frac{10}{s^2 + 4s + 3},$$
which we looked at in Example 7.11. In this case, the hypotheses of part (ii) of Proposition 7.12 are satisfied, and indeed one can see from the Nyquist criterion that the system is IBIBO stable.

2. Next we consider
$$R_L(s) = -\frac{10}{s^2 + 4s + 3},$$
which we also looked at in Example 7.11. Here, the lower phase margin is defined, so the hypotheses of part (ii) of Proposition 7.12 are not satisfied. In this case, the Nyquist criterion tells us that the system is indeed not BIBO stable.

3. The first of the preceding examples illustrates how one can use the conditions on gain and phase margin of Proposition 7.12 to determine closed-loop stability when R_L is BIBO stable. The second example gives a system which violates the sufficient conditions of Proposition 7.12, and is indeed not IBIBO stable. The question then arises, "Is the condition (ii) of Proposition 7.12 *necessary* for IBIBO stability?" The answer is, "No," and we provide an example which illustrates this.
We take
$$R_L(s) = \frac{(1-s)^2(1+\frac{3s}{25})^3}{2(1+s)^2(1+\frac{s}{100})(1+\frac{s}{50})^3}.$$
This is a BIBO stable loop gain, and its Bode plot is shown in Figure 7.16. From the

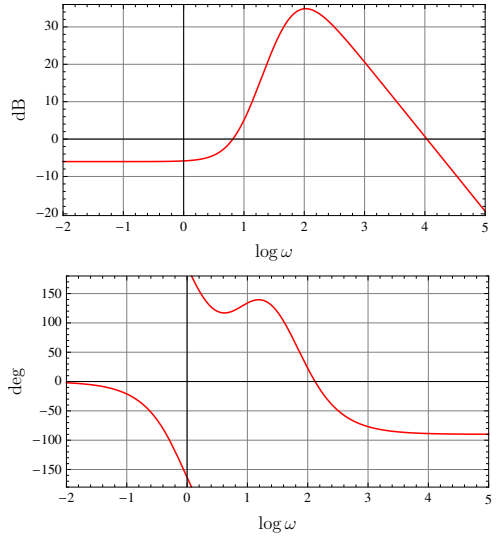


Figure 7.16 Bode plot for BIBO stable transfer function with positive phase margins

Bode plot we can see that there is a gain crossover frequency ω_{gc} satisfying something like $\log \omega_{gc} = 0.8$. The lower phase margin is defined at this frequency, and it is roughly 60°. Thus the lower phase margin is defined, and this loop gain is thus contrary to the hypotheses of part (ii) of Proposition 7.12. Now let us examine the Nyquist contour for the system. In Figure 7.17 we show the Nyquist contour, with the right plot showing a blowup around the origin. This is a pretty messy Nyquist contour, but a careful accounting of what is going on will reveal that there is actually a total of zero encirclements of $-1+i0$. Thus the closed-loop system with loop gain R_L is IBIBO stable. This shows that the converse of Proposition 7.12 is not generally true.

gain margin

4. •

The above examples illustrate that one might wish to make the phase margins large and positive, and make the gain margins large. However, this is only a rough rule of thumb. In Exercise E7.10 the reader can work out an example where look at one stability margin while ignoring the other can be dangerous. The following example from the book of Zhou, Doyle, and Glover [1996] indicates that even when looking at both, they do not necessarily form a useful measure of stability margin.

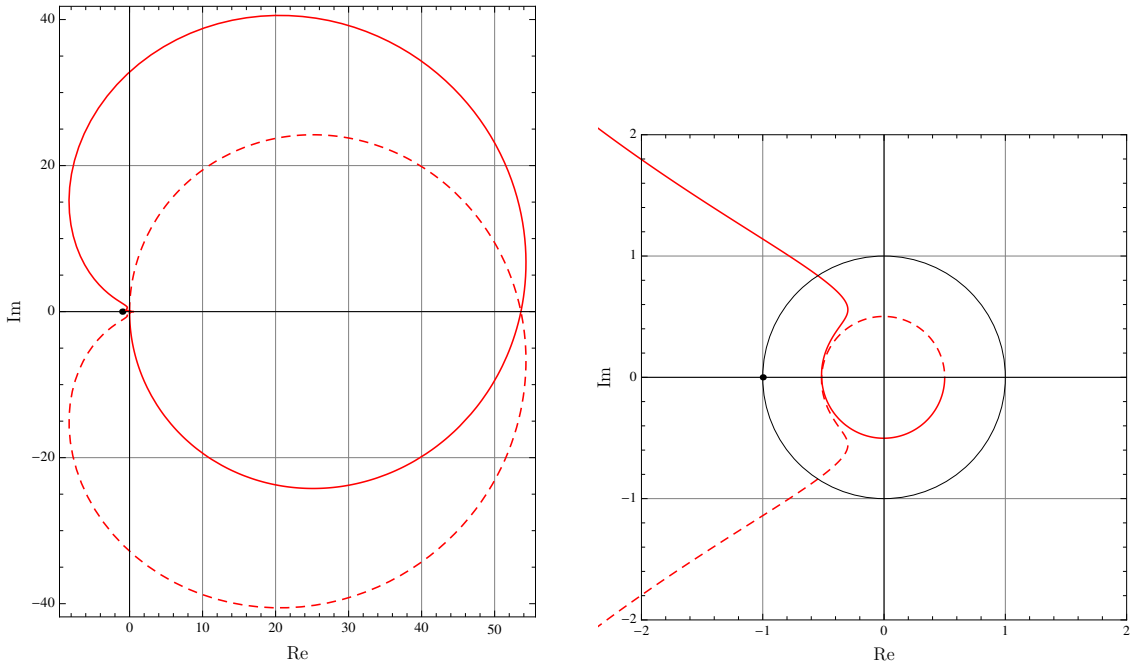


Figure 7.17 Nyquist contour for BIBO stable transfer function with positive phase margins

7.14 Example We take as our plant

$$R_P(s) = \frac{2-s}{2s-1},$$

and we consider two controllers, both of which may be verified to be stabilising:

$$R_{C,1}(s) = 1, \quad R_{C,2}(s) = \frac{s+\frac{33}{10}}{\frac{33}{10}s+1}\frac{s+\frac{11}{20}}{\frac{11}{20}s+1}\frac{\frac{17}{10}s^2+\frac{3}{2}s+1}{s^2+\frac{3}{2}s+\frac{17}{10}}.$$

This second controller is obviously carefully contrived, but let us see what it tells us. In Figure 7.18 are shown the Nyquist plots for the loop gain $R_{C,1}R_P$ and $R_{C,2}R_P$. One can see that the gain and phase margins for the loop gain $R_{C,2}R_P$ are at least as good as those for the loop gain $R_{C,1}R_P$, but that the Nyquist contour for $R_{C,2}R_P$ passes closer to the critical point $-1+i0$. This suggests that gain and phase margin may not be perfect indicators of stability margin. •

With all of the above machinations about gain and phase margins out of the way, let us give perhaps a simpler characterisation of what these notions are trying to capture. If the objective is to stay far way from the point $-1+i0$, then the following result tells us that the sensitivity function is crucial in doing this.

7.15 Proposition $\inf_{\omega>0}|-1-R_L(i\omega)| = \|S_L\|_\infty^{-1}$.

Proof We compute

$$\begin{aligned}\inf_{\omega>0}|-1-R_L(i\omega)| &= \inf_{\omega>0}|1+R_L(i\omega)| \\ &= \Big(\sup_{\omega>0}\Big|\frac{1}{1+R_L(i\omega)}\Big|\Big)^{-1} \\ &= \|S_L\|_\infty^{-1},\end{aligned}$$

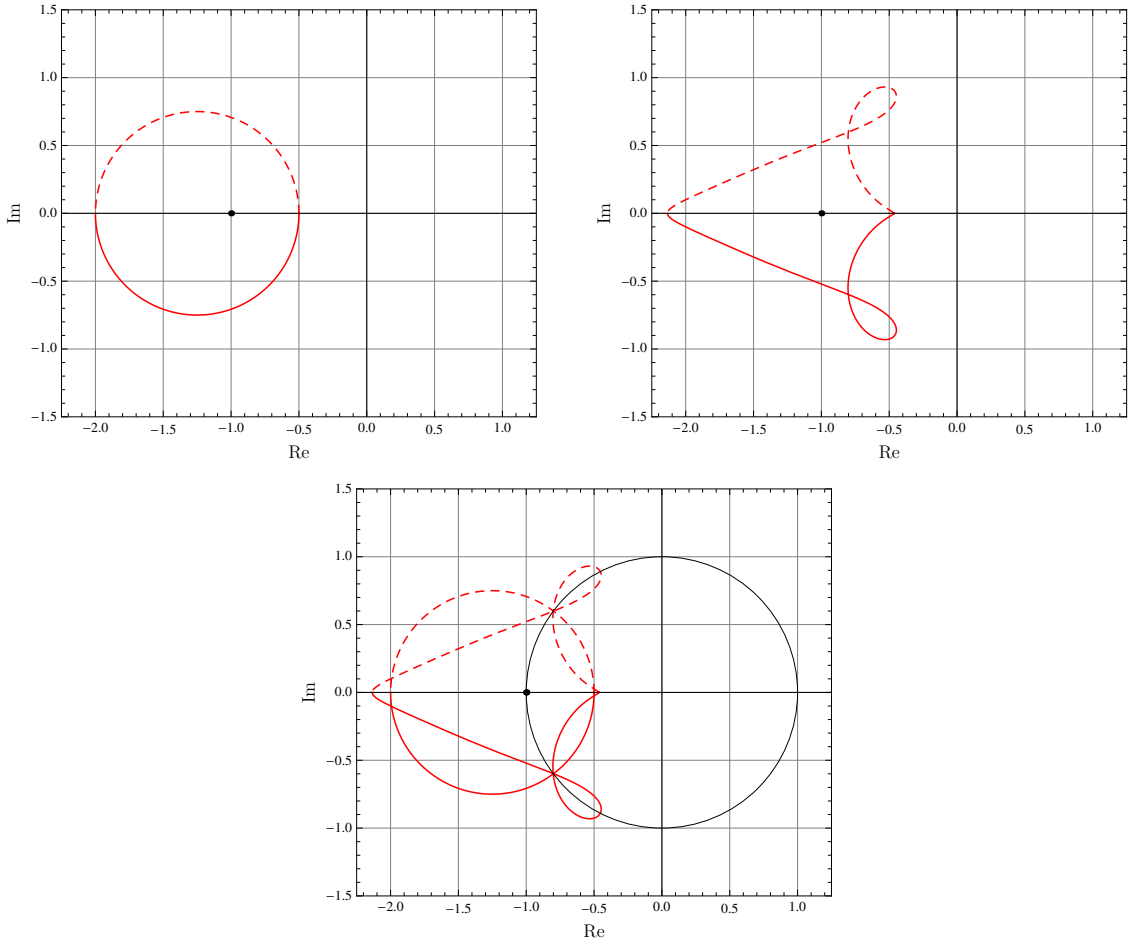


Figure 7.18 Nyquist plots for plant $R_P(s) = \frac{2-s}{2s-1}$ with controller $R_{C,1}$ (top left), with controller $R_{C,2}$ (top right), and both (bottom)

as desired. ■

The results says, simply, that the point on the Nyquist contour which is closest to the point $-1 + i0$ is a distance $\|S_L\|_\infty^{-1}$ away. Thus, to increase the stability margin, one may wish to make the sensitivity function small. This is a reason for minimising the sensitivity function. We shall encounter others in Sections 8.5 and 9.3.

Let us illustrate Proposition 7.15 on the two loop gains of Example 7.14.

7.16 Example **(Example 7.14 cont'd)** We consider the plant transfer function R_P and the two controller transfer functions $R_{C,1}$ and $R_{C,2}$ of Example 7.14. The magnitude Bode plots of the sensitivity function for the two loop gains are shown in Figure 7.19. As expected, the peak magnitude for the sensitivity with the loop gain $R_{C,1}R_P$ is lower than that for $R_{C,2}R_P$, reflecting the fact that the Nyquist contour for the former is further from $-1 + i0$ than for the latter. •

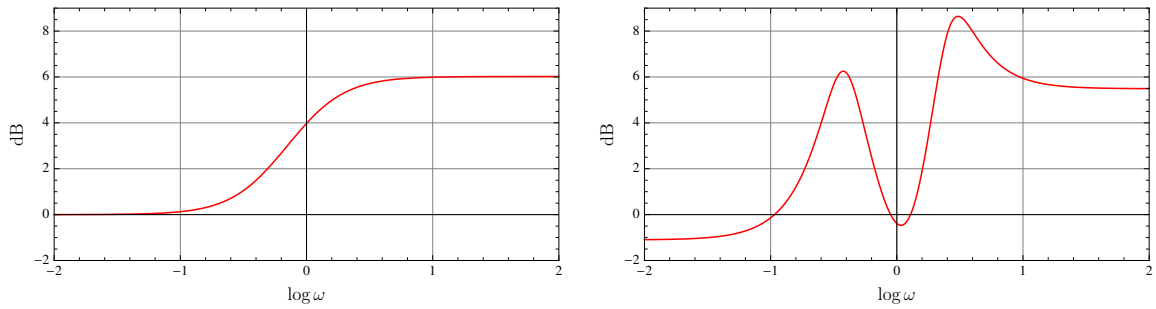


Figure 7.19 Magnitude bode plot for the sensitivity function with loop gain $R_{C,1}R_P$ (left) and $R_{C,2}R_P$ (right)

7.3 Robust stability

We now return to the uncertainty representations of Section 4.5. Let us recall here the basic setup. Given a nominal plant $\bar{R}_P$ and a rational function W_u satisfying $\|W_u\|_\infty < \infty$, we denote by $\mathscr{P}_\times(\bar{R}_P, W_u)$ the set of plants satisfying

$$R_P = (1 + \Delta W_u)\bar{R}_P,$$

where $\|\Delta\|_\infty \leq 1$. We also denote by $\mathscr{P}_+(\bar{R}_P, W_u)$ the set of plants satisfying

$$R_P = \bar{R}_P + \Delta W_u,$$

where $\|\Delta\|_\infty \leq 1$. This is not quite a complete definition, since in Section 4.5 we additionally imposed the requirement that the plants in $\mathscr{P}_\times(\bar{R}_P, W_u)$ and $\mathscr{P}_+(\bar{R}_P, W_u)$ have the same number of unstable poles as does $\bar{R}_P$.

Now we, as usual, consider the unity gain feedback loop of Figure 7.20. We wish to

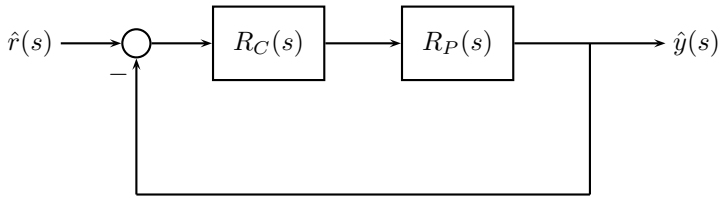


Figure 7.20 Unity gain feedback loop for robust stability

design a controller R_C which stabilises a whole set of plants. We devote this section to a precise formulation of this problem for both uncertainty descriptions, and to providing useful necessary and sufficient conditions for our problem to have a solution.

Let us be precise about this, as it is important that we know what we are saying.

7.17 Definition Let $\bar{R}_P, W_u \in \mathbb{R}(s)$ be a proper rational functions with $\|W_u\|_\infty < \infty$. A controller $R_C \in \mathbb{R}(s)$ provides ***robust stability*** for $\mathscr{P}_\times(\bar{R}_P, W_u)$ (resp. $\mathscr{P}_+(\bar{R}_P, W_u)$) if the feedback interconnection of Figure 7.20 is IBIBO stable for every $R_P \in \mathscr{P}_\times(\bar{R}_P, W)u)$ (resp. for every $R_P \in \mathscr{P}_+(\bar{R}_P, W_u)$). •

Now we provide conditions for determining when a controller is robustly stabilising, breaking our discussion into that for multiplicative, and that for additive uncertainty.

7.3.1 Multiplicative uncertainty

The following important theorem gives simple conditions on when a controller is robustly stabilising. It was first stated by Doyle and Stein [1981] with the first complete proof due to Chen and Desoer [1982].

7.18 Theorem *Let $\bar{R}_P, W_u \in \mathbb{R}(s)$ be a proper rational functions with $\|W_u\|_\infty < \infty$, and suppose that $R_C \in \mathbb{R}(s)$ renders the nominal plant $\bar{R}_P$ IBIBO stable in the interconnection of Figure 7.20. Then R_C provides robust stability for $\mathscr{P}_\times(\bar{R}_P, W_u)$ if and only if $\|W_u\bar{T}_L\|_\infty < 1$, where $\bar{T}_L$ is the complementary sensitivity function, or closed-loop transfer function, for the loop gain $R_C\bar{R}_P$.*

Proof Let $R_C \in \mathbb{R}(s)$. For $R_P \in \mathscr{P}_\times(\bar{R}_P, W_u)$ denote $R_L(R_P) = R_C R_P$. By definition of $\mathscr{P}_\times(\bar{R}_P, W_u)$ it follows that $\bar{R}_P$ and R_P share the same imaginary axis poles for every $R_P \in \mathscr{P}_\times(\bar{R}_P, W_u)$. For $r, R > 0$ and for $R_P \in \mathscr{P}_\times(\bar{R}_P, W_u)$ let $\mathcal{N}_{R,r}(R_P)$ be the (R, r)-Nyquist contour for $R_L(R_P)$.

Now we make a simple computation:

$$\begin{aligned}
& \|W_u\bar{T}_L\|_\infty < 1 \\
\iff \quad & |W_u(i\omega)\bar{T}_L(i\omega)| < 1, \quad \omega \in \mathbb{R} \\
\iff \quad & \Big|\frac{W_u(i\omega)\bar{R}_L(i\omega)}{1+\bar{R}_L(i\omega)}\Big| < 1, \quad \omega \in \mathbb{R} \\
\iff \quad & |W_u(i\omega)\bar{R}_L(i\omega)| < |1+\bar{R}_L(i\omega)|, \quad \omega \in \mathbb{R} \\
\iff \quad & |W_u(i\omega)\bar{R}_L(i\omega)| < |-1-\bar{R}_L(i\omega)|, \quad \omega \in \mathbb{R}.
\end{aligned}$$

This gives a simple interpretation of the condition $\|W_2\bar{T}_L\|_\infty < 1$. We note that $|-1-\bar{R}_L(i\omega)|$ is the distance from the point $-1+i0$ to the point $\bar{R}_L(i\omega)$. Thus the condition $\|W_2\bar{T}_L\|_\infty < 1$ is equivalent to the condition that the open disk of radius $|W_u(i\omega)\bar{R}_L(i\omega)|$ centred at $\bar{R}_L(i\omega)$ not contain the point $-1+i0$. This is depicted in Figure 7.21. It is this interpretation of the condition $\|W_2\bar{T}_L\|_\infty < 1$ we shall employ in the proof.

First suppose that $\|W_u\bar{T}_L\|_\infty < 1$. Note that for $\omega \in \mathbb{R}$ and for $R_P \in \mathscr{P}_\times(\bar{R}_P, W_u)$ we have

$$R_L(R_P)(i\omega) = R_C(i\omega)R_P(i\omega) = (1+\Delta(i\omega)W_u(i\omega))R_C(i\omega)\bar{R}_P(i\omega).$$

Thus the point $R_L(R_P)(i\omega)$ lies in the closed disk of radius $|W_u(i\omega)\bar{R}_L(i\omega)|$ with centre $\bar{R}_L(i\omega)$. From Figure 7.21 we infer that the point of the the Nyquist contour $\mathcal{N}_{R,r}(R_P)$ that are the image of points on the imaginary axis will follow the points on the Nyquist contour $\mathcal{N}_{R,r}(\bar{R}_P)$ while remaining on the same "side" of the point $-1+i0$. Since $W_u \in \mathrm{RH}_\infty^+$ and since Δ is allowable, $R_L(R_P)$ and $\bar{R}_L$ have the same poles on $i\mathbb{R}$ and the same number of poles in $\mathbb{C}_+$. Thus by choosing $r_0 < 0$ sufficiently small and for $R_0 > 0$ sufficiently large, the number of clockwise encirclements of $-1+i0$ by $\mathcal{N}_{R,r}(R_P)$ will equal the same by $\mathcal{N}_{R,r}(\bar{R}_P)$ for all $R > R_0$ and $r < r_0$. From Theorem 7.6 we conclude IBIBO stability of the closed-loop system with loop gain $R_L(R_P)$.

Now suppose that $\|W_u\bar{T}_L\|_\infty \geq 1$. As depicted in Figure 7.21, this implies the existence of $\bar{\omega} \geq 0$ so that the open disk of radius $|W_u(i\bar{\omega})\bar{R}_L(i\bar{\omega})|$ centred at $\bar{R}_L(i\bar{\omega})$ contains the point $-1+i0$. Denote by $\bar{D}(s_0, r)$ be the closed disk of radius r centred at s_0. We claim that for each $\omega > 0$, the map from $\mathscr{P}_\times(\bar{R}_P, W_u)$ to $D\big(\bar{R}_L(i\omega), |W_u(i\omega)\bar{R}_L(i\omega)|\big)$ defined by

$$R_P = (1+\Delta W_u)\bar{R}_P \mapsto R_C(i\omega)R_P(i\omega) \tag{7.4}$$

is surjective. The following lemma helps us to establish this fact.

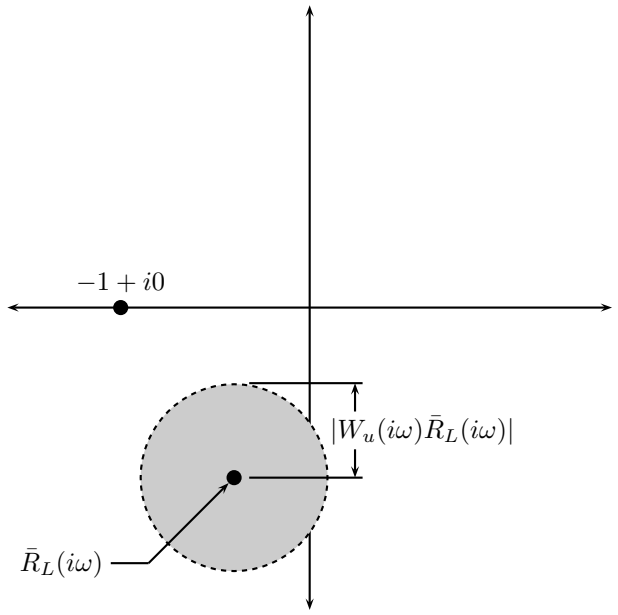


Figure 7.21 Interpretation of robust stability condition for multiplicative uncertainty

1 Lemma *For any $\bar{\omega} \geq 0$ and for any $\theta \in (-\pi, \pi]$ there exists a function $G_\theta \in \mathrm{RH}_\infty^+$ with the properties*

(i) $\angle G_\theta(i\bar{\omega}) = \theta$ and

(ii) the map $\omega \mapsto |G_\theta(i\omega)|$ has a maximum at $\omega = \bar{\omega}$.

Proof Clearly if $\theta = 0$ we may simply define $G_\theta(s) = 1$. Thus suppose that $\theta \neq 0$. Consider the rational function

$$T_{\zeta,\omega_0}(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2},$$

$\zeta, \omega_0 > 0$. In Exercise E4.6 it was shown that for $\frac{1}{\sqrt{2}} < \zeta < 1$ the function $\omega \mapsto |T(i\omega)|$ achieves a unique maximum at $\omega_{\max} = \omega_0\sqrt{1-2\zeta^2}$. Furthermore, the phase at this maximum is given by $\operatorname{atan2}(\zeta, \sqrt{1-2\zeta^2})$. Thus, as ζ varies in the interval $(\frac{1}{\sqrt{2}}, 1)$, the phase at the frequency $\omega_{\max}$ varies between $-\frac{\pi}{4}$ and 0.

Now, given $\theta \in (-\pi, \pi]$, define

$$\tilde{\theta} = \begin{cases} \frac{\theta}{5}, & \theta < 0 \\ -\frac{2\pi-\theta}{10}, & \theta > 0. \end{cases}$$

Thus $\tilde{\theta}$ is guaranteed to live in the interval $(-\frac{\pi}{4}, 0)$. Therefore, there exists $\zeta \in (\frac{1}{\sqrt{2}}, 1)$ so that $\tilde{\theta} = \operatorname{atan2}(\zeta, \sqrt{1-2\zeta^2})$. Now define ω_0 so that $\omega_{\max} = \bar{\omega}$. Now define

$$G_\theta = \begin{cases} T_{\zeta,\omega_0}^5, & \theta < 0 \\ T_{\zeta,\omega_0}^{10}, & \theta > 0. \end{cases}$$

This function G_θ, it is readily verified, does the job. ▼

We now resume showing that the map defined by (7.4) is surjective. It clearly suffices to show that the image of every point on the boundary of $D\big(\bar{R}_L(i\omega), |W_u(i\omega)\bar{R}_L(i\omega)|\big)$ lies in the image of the map, since all other points can then be obtained by scaling Δ. For $\phi \in (-\pi, \pi]$ let

$$s_\phi = \bar{R}_L(i\omega) + |W_u(i\omega)\bar{R}_L(i\omega)|e^{i\phi}$$

be a point on the boundary of $D\big(\bar{R}_L(i\omega), |W_u(i\omega)\bar{R}_L(i\omega)|\big)$. We wish to write

$$s_\phi = R_C(i\omega)R_P(i\omega) = \bar{R}_L(i\omega) + \Delta(i\omega)W_u(i\omega)\bar{R}_L(i\omega)$$

for an appropriate choice of allowable Δ. Thus $\Delta \in \mathrm{RH}_\infty^+$ should necessarily satisfy

$$\Delta(i\omega)W_u(i\omega)\bar{R}_L(i\omega) = |W_u(i\omega)\bar{R}_L(i\omega)|e^{i\phi}.$$

It follows that $|\Delta(i\omega)| = 1$ and that $\measuredangle\Delta(i\omega) + \measuredangle(W_u(i\omega)\bar{R}_L(i\omega)) = \phi$. Therefore, define

$$\theta = \phi - \measuredangle(W_u(i\omega)\bar{R}_L(i\omega)).$$

If G_θ is as defined above, then defining

$$\Delta(s) = \frac{G_\theta(s)}{|G_\theta(i\omega)|}$$

does the job.

The remainder of the proof is now straightforward. Since the map defined by (7.4) is surjective, we conclude that there exists an allowable Δ so that if $R_P = (1 + \Delta W_u)\bar{R}_P$ we have

$$R_L(R_P)(i\bar{\omega}) = R_C(i\bar{\omega})R_P(i\bar{\omega}) = -1 + i0.$$

This implies that the Nyquist contour $\mathcal{N}_{R,r}(R_P)$ for R sufficiently large passes through the point $-1 + i0$. By Theorem 7.6 the closed-loop system is not IBIBO stable. ■

The proof of the above theorem is long-winded, but the idea is, in fact, very simple. Indeed, the essential observation, repeated here outside the confines of the proof, is that the condition $\|W_u T_L\|\infty < 1$ is equivalent to the condition, depicted in Figure 7.21, that, for each frequency ω, the open disk of radius $|W_u(i\omega)\bar{R}_L(i\omega)|$ and centred at $\bar{R}_L(i\omega)$, not contain the point $-1 + j0$.

One may also "reverse engineer" this problem as well. The idea here is as follows. Suppose that we have our nominal plant $\bar{R}_P$ and a controller R_C which IBIBO stabilises the closed-loop system of Figure 7.20. At this point, we have not specified a set of plants over which we want to stabilise. Now we ask, "What is the maximum size of the allowable perturbation to $\bar{R}_P$ if the perturbed plant is to be stabilised by R_C?" The following result makes precise this vague question, and provides its answer.

7.19 Proposition *Let $\bar{R}_P, R_C \in \mathbb{R}(s)$ with $\bar{R}_P$ proper. Also, suppose that the interconnection of Figure 7.20 is IBIBO stable with $R_P = \bar{R}_P$. Let $\bar{T}_L$ be the closed-loop transfer function for the loop gain $\bar{R}_L = R_C\bar{R}_P$. The following statements hold:*

(i) if $W_u \in \mathrm{RH}_\infty^+$ has the property that $\|W_u\|_\infty < \|\bar{T}_L\|_\infty^{-1}$ then R_C provides robust stability for $\mathscr{P}_\times(\bar{R}_P, W_u)$;

(ii) for any $\beta \geq \|\bar{T}_L\|_\infty^{-1}$ there exists $W_u \in \mathrm{RH}_\infty^+$ satisfying $\|W_u\| = \beta$ so that R_C does not robustly stabilise $\mathscr{P}_\times(\bar{R}_P, W_u)$.

Proof (i) This follows directly from Theorem 7.18.

(ii) This part of the result too follows from Theorem 7.18. Indeed, let $\bar{\omega} > 0$ be a frequency for which $|\bar{T}_L(i\bar{\omega})| = \|\bar{T}_L\|_\infty$. One can readily define $W_u \in \mathrm{RH}_\infty^+$ so that $|W_u(i\bar{\omega})| = \|W_u\|_\infty = \beta$. It then follows that $\|W_u\bar{T}_L\|_\infty \geq 1$, so we may apply Theorem 7.18. ■

Roughly, the proposition tells us that as long as we choose the uncertainty weight W_u so that its H_∞-norm is bounded by $\|\bar{T}_L\|_\infty^{-1}$, then we are guaranteed that R_C will provide robust stability. If the H_∞-norm of W_u exceeds $\|\bar{T}_L\|_\infty^{-1}$, then it is possible, but not certain, that R_C will not robustly stabilise. In this latter case, we must check the condition of Theorem 7.18.

Let us illustrate the concept of robust stability for multiplicative uncertainty with a fairly simple example.

7.20 Example We take as our nominal plant the model for a unit mass. Thus

$$\bar{R}_P(s) = \frac{1}{s^2}.$$

The PID control law given by

$$R_C(s) = 1 + 2s + \frac{1}{s}$$

may be verified to stabilise the nominal plant; the Nyquist plot is shown in Figure 7.22. To

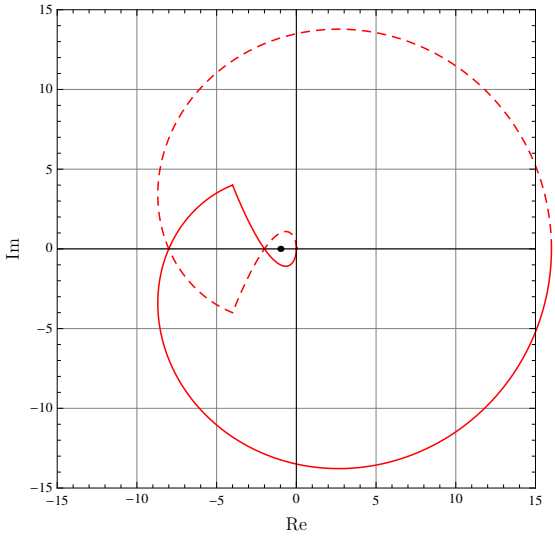


Figure 7.22 Nyquist plot for PID controller and nominal plant

model the plant uncertainty, we choose

$$W_u(s) = \frac{as}{s+1}, \quad a > 0,$$

which has the desirable property of not tailing off to zero as $s \to \infty$; plant uncertainty will generally increase with frequency. We shall determine for what values of a the controller R_C provides robust stability for $\mathscr{P}_\times(\bar{R}_P, W_u)$.

According to Theorem 7.18 we should determine for which values of a the inequality $\|W_u\bar{T}_L\|_\infty < 1$ is satisfied. To determine $\|W_u\bar{T}_L\|_\infty$ we merely need to produce magnitude Bode plots for $W_u\bar{T}_L$ for various values of a, and determine for which values of a the maximum

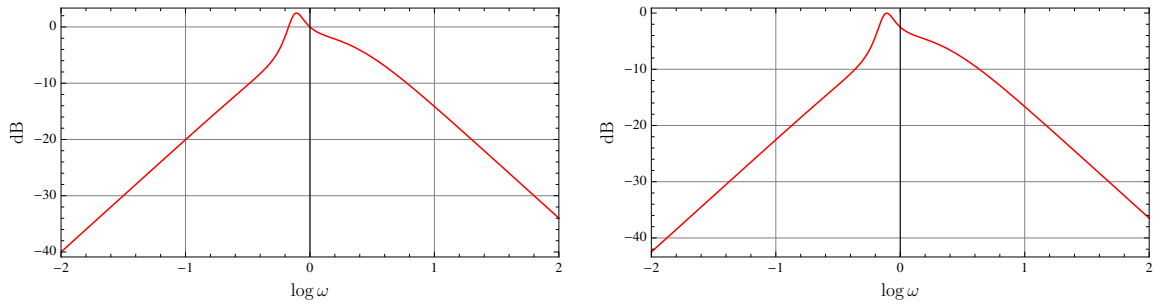


Figure 7.23 The magnitude Bode plot for $W_u\bar{T}_L$ when $a = 1$ (left) and when $a = \frac{3}{4}$ (right)

magnitude does not exceed 0dB. In Figure 7.23 is shown the magnitude Bode plot for $W_u\bar{T}_L$ with $a = 1$. We see that we exceed 0dB by about 2.5dB. Thus we should reduce a by a factor K having the property that $20 \log K = 2.5$ or $K \approx 1.33$. Thus $a \approx 0.75$. Thus let us take $a = \frac{3}{4}$ for which the magnitude Bode plot is produced in Figure 7.23. The magnitude is bounded by 0dB, so we are safe, and all plants of the form

$$R_P(s) = \left(1 + \frac{\frac{3}{4}s\Delta(s)}{s+1}\right)\frac{1}{s^2}$$

will be stabilised by R_C if Δ is allowable. For example, the Nyquist plot for R_P when $\Delta = \frac{s-1}{s+2}$ (which, it can be checked, is allowable), is shown in Figure 7.24. •

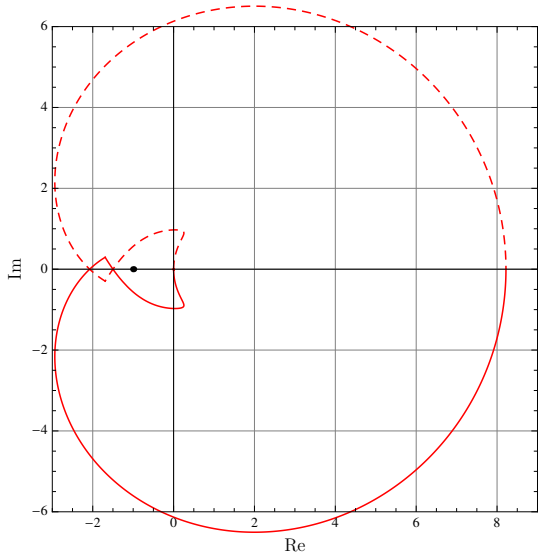


Figure 7.24 Nyquist plot for perturbed plant under multiplicative uncertainty

7.3.2 Additive uncertainty

Now let us state the analogue of Theorem 7.18 for additive uncertainty.

7.21 Theorem *Let $\bar{R}_P, W_u \in \mathbb{R}(s)$ be a proper rational functions with $\|W_u\|_\infty < \infty$, and suppose that $R_C \in \mathbb{R}(s)$ renders the nominal plant $\bar{R}_P$ IBIBO stable in the interconnection of Figure 7.20. Then R_C provides robust stability for $\mathscr{P}_+(\bar{R}_P, W_u)$ if and only if $\|W_u R_C \bar{S}_L\|_\infty < 1$, where $\bar{S}_L$ is the sensitivity function for the loop gain $R_C \bar{R}_P$.*

Proof We adopt the notation of the first paragraph of the proof of Theorem 7.18. The following computation, mirroring the similar one in Theorem 7.18, is the key to the proof:

$$\begin{aligned}
& \|W_u R_C \bar{S}_L\|_\infty < 1 \\
\iff \quad & |W_u(i\omega) R_C(i\omega) \bar{S}_L(i\omega)| < 1, \quad \omega \in \mathbb{R} \\
\iff \quad & \Big| \frac{W_u(i\omega) R_C(i\omega)}{1 + \bar{R}_L(i\omega)} \Big| < 1, \quad \omega \in \mathbb{R} \\
\iff \quad & |W_u(i\omega) R_C(i\omega)| < |1 + \bar{R}_L(i\omega)|, \quad \omega \in \mathbb{R} \\
\iff \quad & |W_u(i\omega) R_C(i\omega)| < |-1 - \bar{R}_L(i\omega)|, \quad \omega \in \mathbb{R}.
\end{aligned}$$

The punchline here is thus that the condition $\|W_u R_C \bar{S}_L\|_\infty < 1$ is equivalent to the condition that the point $-1 + i0$ not be contained, for any $\omega \in \mathbb{R}$, in the open disk of radius $|W_u(i\omega) R_C(i\omega)|$ with centre $\bar{R}_L(i\omega)$. This is depicted in Figure 7.25.

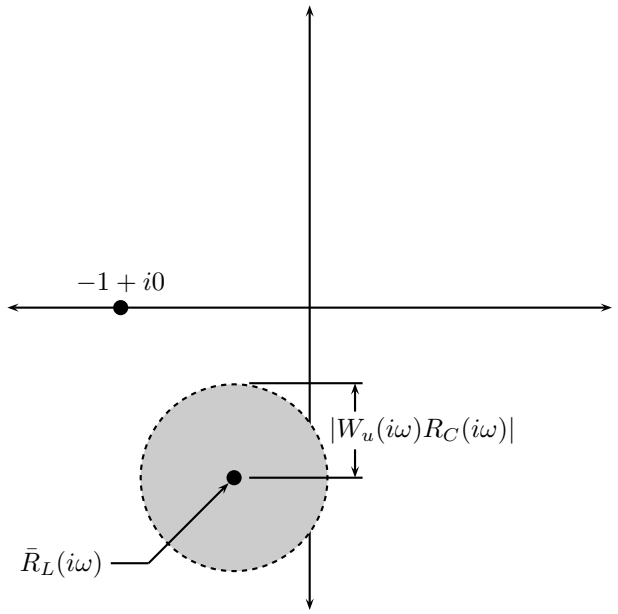


Figure 7.25 Interpretation of robust stability condition for additive uncertainty

The remainder of the proof now very much follows that of Theorem 7.18, so we can safely omit some details. First assume that $\|W_u R_C \bar{S}_L\|_\infty < 1$. For $\omega \in \mathbb{R}$ and for $R_P \in \mathscr{P}_+(\bar{R}_P, W_u)$ we have

$$R_L(R_P)(i\omega) = R_C(i\omega) R_P(i\omega) = \bar{R}_L(i\omega) + \Delta(i\omega) R_C(i\omega) W_u(i\omega).$$

Thus the point $R_L(R_P)(i\omega)$ lies in the closed disk of radius $|W_u(i\omega) R_C(i\omega)|$ with centre $\bar{R}_L(i\omega)$. We may now simply repeat the argument of Theorem 7.18, now using Figure 7.25 rather than Figure 7.21, to conclude that the closed-loop system with loop gain $R_L(R_P)$ is IBIBO stable.

Now suppose that $\|W_uR_C\|_\infty \geq 1$. Thus there exists $\bar{\omega} \geq 0$ so that $|W_u(i\bar{\omega})R_C(i\bar{\omega})| \geq 1$. Thus by Figure 7.25, it follows that $-1 + i0$ is contained in the open disk of radius $|W_u(i\bar{\omega}), R_C(i\bar{\omega})|$ with centre $\bar{R}_L(i\bar{\omega})$. As in the proof of Theorem 7.18, we may employ Lemma 1 of that proof to show that the map from $\mathscr{P}_+(\bar{R}_P, W_u)$ to $D\big(\bar{R}_L(i\omega), |W_u(i\omega)R_C(i\omega)|\big)$ defined by

$$R_P = \bar{R}_P + \Delta W_u \mapsto R_C(i\omega)R_P(i\omega)$$

is surjective for each $\omega > 0$. In particular, it follows that there exists an allowable Δ giving $R_P \in \mathscr{P}_+(\bar{R}_P, W_u)$ so that $R_L(R_P)(i\bar{\omega}) = -1 + i0$. IBIBO instability now follows from Theorem 7.6. ■

Again, the details in the proof are far more complicated than is the essential idea. This essential idea is that for each $\omega \in \mathbb{R}$ the open disk of radius $|W_u(i\omega)R_C(i\omega)|$ and centre $\bar{R}_L(i\omega)$ should not contain the point $-1 + i0$. This is what is depicted in Figure 7.25.

We also have the following, now hopefully obvious, analogue of Proposition 7.19.

7.22 Proposition *Let $\bar{R}_P, R_C \in \mathbb{R}(s)$ with $\bar{R}_P$ proper. Also, suppose that the interconnection of Figure 7.20 is IBIBO stable with $R_P = \bar{R}_P$. Let $\bar{S}_L$ be the sensitivity function for the loop gain $\bar{R}_L = R_C\bar{R}_P$. The following statements hold:*

(i) if $W_u \in \mathrm{RH}_\infty^+$ has the property that $\|W_u\|_\infty < \|R_C\bar{S}_L\|_\infty^{-1}$ then R_C provides robust stability for $\mathscr{P}_+(\bar{R}_P, W_u)$;

(ii) for any $\beta \geq \|R_C\bar{S}_L\|_\infty^{-1}$ there exists $W_u \in \mathrm{RH}_\infty^+$ satisfying $\|W_u\| = \beta$ so that R_C does not robustly stabilise $\mathscr{P}_+(\bar{R}_P, W_u)$.

An example serves to illustrate the ideas for this section.

7.23 Example **(Example 7.20 cont'd)** We carry on look at the nominal plant transfer function

$$\bar{R}_P(s) = \frac{1}{s^2}$$

which is stabilised by the PID controller

$$R_C(s) = 1 + 2s + \frac{1}{s}.$$

Note that we may no longer use our W_u from Example 7.20 as our plant uncertainty model. Indeed, since R_C is improper, W_u is proper but not strictly proper, and $\bar{S}_L$ is proper but not strictly proper (one can readily compute that this is so), it follows that $W_uR_C\bar{S}_L$ is improper. Thus we modify W_u to

$$W_u = \frac{as}{(s+1)^2}$$

to model the plant uncertainty. Again, our objective will be to determine the maximum value of a so that R_C provides robust stability for $\mathscr{P}_+(\bar{R}_P, W_u)$. Thus we should find the maximum value for a so that $\|W_uR_C\bar{S}_L\|_\infty < 1$. In Figure 7.26 is shown the magnitude Bode plot for $W_uR_C\bar{S}_L$ when $a = 1$. From this plot we see that we ought to reduce a by a factor K having the property that $20\log K = 6$ or $K \approx 2.00$. Thus we take $a = \frac{1}{2}$, and in Figure 7.26 we see that with this value of a we remain below the 0dB line. Thus we are guaranteed that all plants of the form

$$R_P(s) = \bar{R}_P(s) + \frac{\frac{1}{2}s\Delta(s)}{(s+1)^2}$$

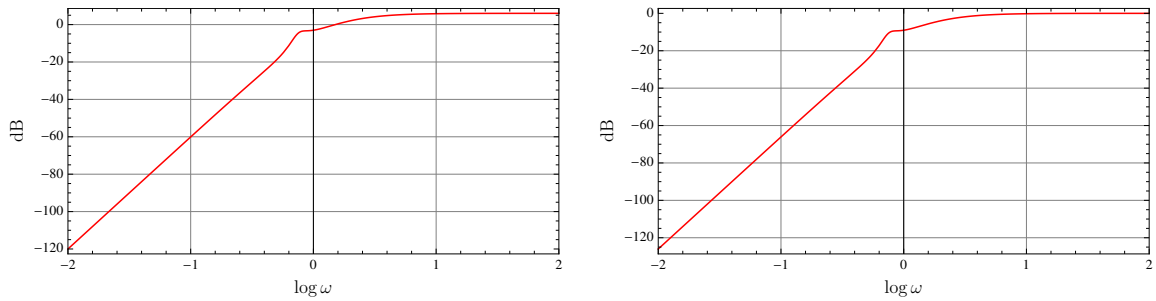


Figure 7.26 The magnitude Bode plot for $W_u R_C \bar{S}_L$ when $a = 1$ (left) and when $a = \frac{1}{2}$ (right)

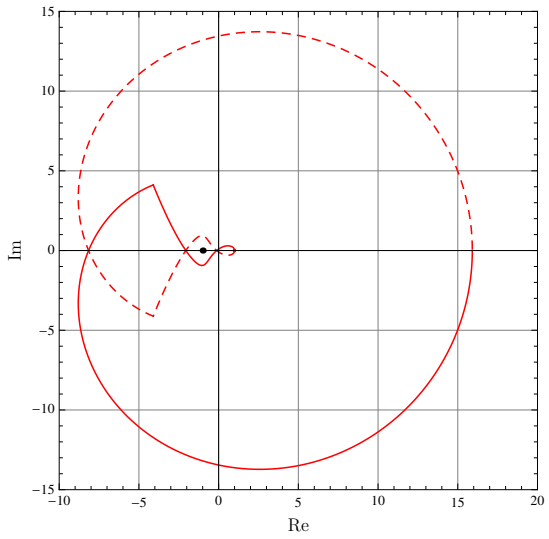


Figure 7.27 Nyquist plot for perturbed plant under additive uncertainty

are stabilised by R_C provided that Δ is allowable. The Nyquist plot for R_P obtained by taking $\Delta = \frac{s-1}{s+2}$ is shown in Figure 7.27. This is a pretty sophisticated Nyquist plot, but nonetheless, it is one for an IBIBO stable system. •

7.4 Summary

In this chapter we have provided a graphical method for analysing the closed-loop stability of a single loop feedback interconnection. You should understand the following things.

1. You need to understand what the Nyquist criterion, in the form of Theorem 7.6, is saying.
2. Drawing Nyquist plots can be a bit tricky. The essential idea is that one should focus on the image of the various parts of the Nyquist contour $\Gamma_{R,r}$. In particular, one should focus on the image of the imaginary axis for very large and very small frequencies. From there one can try to understand of something more subtle is happening.
3. The gain and phase margins are sometimes useful measures of how close a system is to being unstable.

Exercises

E7.1 For the following rational functions R and contours C, explicitly verify the Principle of the Argument by determining the number of encirclements of the origin by the image of C under R.

(a) $R(s) = \frac{1}{s^2}$ and $C = \{e^{i\theta} \mid -\pi < \theta \leq \pi\}$.

(b) $R(s) = s$ and $C = \{e^{i\theta} \mid -\pi < \theta \leq \pi\}$.

(c) $R(s) = \frac{s}{s^2+1}$ and $C = \{2e^{i\theta} \mid -\pi < \theta \leq \pi\}$.

E7.2 Let $F(s) = \ln s$, and consider the contour $C = \{e^{i\theta} \mid -\pi < \theta \leq \pi\}$. Does the Principle of the Argument hold for the image of C under F? Why or why not?

E7.3 For the following loop gains,

(a) $R_L(s) = \frac{\alpha}{s^2(s+1)}$, $\alpha > 0$,

(b) $R_L(s) = \frac{\alpha}{s(s-1)}$, $\alpha > 0$, and

(c) $R_L(s) = \frac{\alpha(s+1)}{s(s-1)}$, $\alpha > 0$,

do the following.

1. Determine the Nyquist contour for the following loop gains which depend on a parameter α satisfying the given conditions. Although the plots you produce may be computer generated, you should make sure you provide analytical explanations for the essential features of the plots as they vary with α.
2. Draw the unity gain feedback block diagram which has R_L as the loop gain.
3. For the three loop gains, use the Nyquist criterion to determine conditions on α for which the closed-loop system is IBIBO stable.
4. Determine IBIBO stability of the three closed-loop systems using the Routh/Hurwitz criterion, and demonstrate that it agrees with your conclusions using the Nyquist criterion.

In this next exercise you will investigate some simple ways of determining how to "close" a Nyquist contour for loop gains with poles on the imaginary axis.

E7.4 Let $R_L \in \mathbb{R}(s)$ be a proper loop gain with a pole at $i\omega_0$ of multiplicity k. For concreteness, suppose that $\omega_0 \geq 0$. If $\omega_0 \neq 0$, this implies that $-i\omega_0$ is also a pole of multiplicity k. Thus we write

$$R_L(s) = \begin{cases} \frac{1}{s^k} R(s), & \omega_0 = 0 \\ \frac{1}{(s^2+\omega_0^2)^k} R(s), & \omega_0 \neq 0, \end{cases}$$

where $i\omega_0$ is neither a pole nor a zero for R.

(a) Show that

$$\lim_{r \to 0} R_L(i\omega_0 + re^{i\theta}) = \frac{a}{r^k} e^{-ik\theta} + O(r^{1-k}).$$

where a is either purely real or purely imaginary. Give the expression for a.

(b) For $\omega_0 = 0$ show that

$$\lim_{\omega \to \omega_{0,-}} \measuredangle R_L(i\omega) \in \left\{-\tfrac{\pi}{2}, 0, \tfrac{\pi}{2}, \pi\right\}.$$

Denote $\theta_0 = \lim_{\omega \to \omega_{0,-}} \measuredangle R_L(i\omega)$.

(c) Determine a relationship between θ_0, and k and a.

(d) Determine the relationship between

$$\lim_{\omega \to \omega_{0,+}} \angle R_L(i\omega)$$

and θ_0. This relationship will depend on k.

(e) Conclude that for a real rational loop gain, the "closing" of the Nyquist contour is always in the clockwise direction and the closing arc subtends an angle of $k\pi$.

E7.5 Let (N, D) be a proper SISO linear system in input/output form, and let $n = \deg(D)$ and $m = \deg(N)$.

(a) Determine $\lim_{\omega\to\infty} \angle T_{N,D}(i\omega)$ and $\lim_{\omega\to\infty} \angle T_{N,D}(i\omega)$.

(b) Comment on part (a) as it bears on the Nyquist plot for the system.

E7.6 Consider the SISO linear system $(N(s), D(s)) = (1, s^2 + 1)$ in input/output form.

(a) Sketch the Nyquist contour for the system, noting that the presence of imaginary axis poles means that the (∞, r)-Nyquist contour is not bounded as $r \to 0$.

Consider the SISO linear system $(N_\epsilon(s), D_\epsilon(s)) = (1, s^2 + \epsilon s + 1)$ which now has a bounded Nyquist contour for $\epsilon > 0$.

(b) Show, using the computer, that the Nyquist contour for (N_ϵ, D_ϵ) approaches that for the system of part (a) as $\epsilon \to 0$.

E7.7 Let (N, D) be a SISO linear system in input/output form with $\deg(N) = 0$ and let

$$\bar{\omega} = \max\{\mathrm{Im}(p) \mid p \text{ is a root of } D\}.$$

Answer the following questions.

(a) Show that $|T_{D,N}(i\omega)|$ is a strictly decreasing function of ω for $|\omega| > \bar{\omega}$.

(b) Comment on part (a) as it bears on the Nyquist plot for the system.

E7.8 Formulate and prove a condition for IBIBO stability for each of the two interconnected systems in Figure E7.1, using the Principle of the Argument along the lines of the

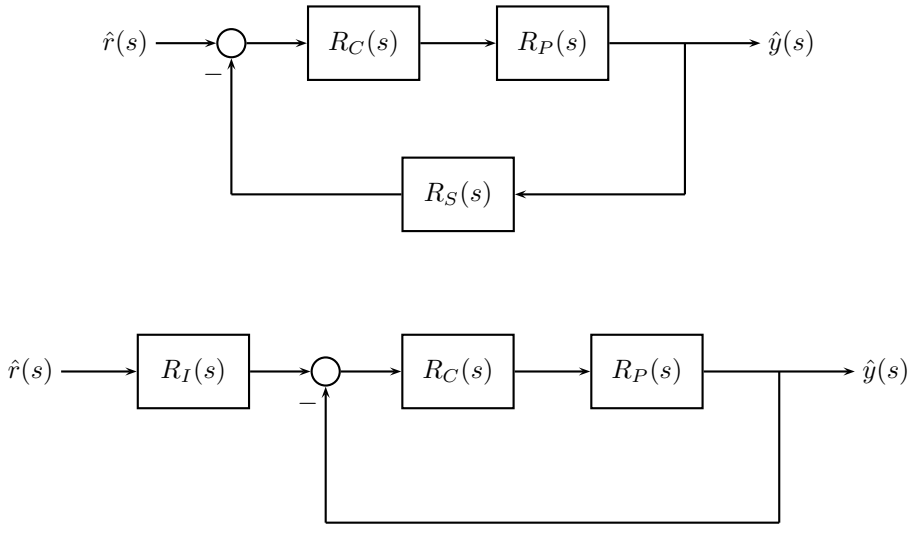


Figure E7.1 Alternate configurations for Nyquist criterion

Nyquist criterion of Theorem 7.6.

E7.9 Let

$$R_P(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s}, \quad R_C(s) = 1.$$

Plot the upper phase margin of the closed-loop system as a function of ζ.

In the next exercise you will investigate gain and phase margins for a simple plant and controller. You will see that it is possible to have one stability margin be large while the other is small. This exercise is taken from the text of [Zhou, Doyle, and Glover 1996].

E7.10 Take

$$R_P(s) = \frac{a - s}{as - 1}, \quad R_C(s) = \frac{s + b}{bs + 1},$$

where $a > 1$ and $b > 0$. Consider these in the standard unity gain feedback loop.

(a) Use the Routh/Hurwitz criterion to show that the closed-loop system is IBIBO stable if $b \in (\frac{1}{a}, a)$.

Take $b = 1$.

(b) Show that $K_{\min} = \frac{1}{a}$, $K_{\max} = a$, $\Phi_{\min}$ is undefined, and $\Phi_{\max} = \arcsin\big(\frac{a^2-1}{a^2+1}\big)$.

(c) Comment on the nature of the stability margins in this case.

Fix a and take $b \in (\frac{1}{a}, a)$.

(d) Show that

$$\lim_{b\to a} K_{\min} = \frac{1}{a^2}, \quad \lim_{b\to a} K_{\max} = a^2, \quad \lim_{b\to a} \Phi_{\max} = 0.$$

(e) Comment on the stability margins in the previous case.

(f) Show that

$$\lim_{b\to \frac{1}{a}} K_{\min} = 1, \quad \lim_{b\to \frac{1}{a}} K_{\max} = 1, \quad \lim_{b\to \frac{1}{a}} \Phi_{\max} = 2\arcsin\Big(\frac{a^2 - 1}{a^2 + 1}\Big).$$

(g) Comment on the stability margins in the previous case.

In this chapter we have used the Nyquist criterion to assess IBIBO stability of input/output feedback systems. It is also possible to use the Nyquist criterion to assess stability of static state feedback, and the following exercise indicates how this is done.

E7.11 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system.

(a) Let $\boldsymbol{f} \in \mathbb{R}^n$. Show that if Σ is controllable then the closed-loop system $\Sigma_{\boldsymbol{f}}$ is internally asymptotically stable if and only if

$$\frac{\boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}}{1 + \boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}} \in \mathrm{RH}_\infty^+.$$

Thus, we may consider closed-loop stability under static state feedback to be equivalent to IBIBO stability of the interconnection in Figure E7.2. The Nyquist criterion can then be applied, as indicated in the following problem.

(b) Using part (a), prove the following result.

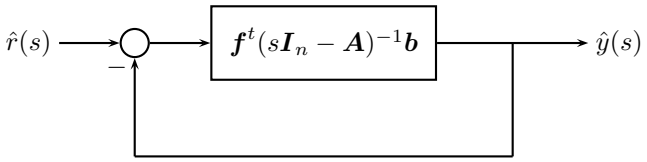


Figure E7.2 A feedback interconnection for static state feedback

Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a controllable SISO linear system and let n_p be the number of eigenvalues of $\boldsymbol{A}$ in $\mathbb{C}_+$. For a state feedback vector $\boldsymbol{f} \in \mathbb{R}^n$ define the loop gain $R_{\boldsymbol{f}}(s) = \boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}$. Assuming that $(\boldsymbol{A}, \boldsymbol{f})$ is observable, the following statements are equivalent:*

(i) the matrix $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ is Hurwitz;

(ii) there exists $r_0, R_0 > 0$ so that the image of $\Gamma_{R,r}$ under $R_{\boldsymbol{f}}$ encircles the point $-1 + i0$ n_p times in the clockwise direction for every $r < r_0$ and $R > R_0$.

To be concrete, consider the situation where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

For this system, answer the following questions.

(c) Show that if we take $\boldsymbol{f} = (0, 2)$ then $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ is Hurwitz. Verify that the result of part (a) of the problem holds.

(d) How many eigenvalues are there for $\boldsymbol{A}$ in $\mathbb{C}_+$?

(e) Plot the Nyquist contour for Figure E7.2 and verify that the Nyquist criterion of part (b) holds.

E7.12 For the plant uncertainty set

$$R_P = \frac{\bar{R}_P}{1 + \Delta W_u \bar{R}_P}, \quad \|\Delta\|_\infty \leq 1,$$

use the idea demonstrated in the proofs of Theorems 7.18 and 7.21 to state and prove a theorem providing conditions for robust stability.

E7.13 For the plant uncertainty set

$$R_P = \frac{\bar{R}_P}{1 + \Delta W_u}, \quad \|\Delta\|_\infty \leq 1,$$

use the idea demonstrated in the proofs of Theorems 7.18 and 7.21 to state and prove a theorem providing conditions for robust stability.

Chapter 8

Performance of control systems

Before we move on to controller design for closed-loop systems, we should carefully investigate to what sort of specifications we should make our design. These appear in various ways, depending on whether we are working with the time-domain, the s-domain, or the frequency-domain. Also, the interplay of performance and stability can be somewhat subtle, and there is often a tradeoff to be made in this regard. Our objective in this chapter is to get the reader familiar with some of the issues that come up when discussing matters of performance. Also, we wish to begin the process of familiarising the reader with the ways in which various system properties can affect the performance measures. In many cases, intuition is one's best friend when such matters come up. Therefore, we take a somewhat intuitive approach in Section 8.2. The objective is to look at tweaking parameters in a couple of simple transfer function to see how they affect the various performance measures. It is hoped that this will give a little insight into how one might, say, glean some things about the step response by looking at the Bode plot. The matter of tracking and steady-state error can be dealt with in a more structured way. Disturbance rejection follows along similar lines. The final matter touched upon in this chapter is the rôle of the sensitivity function in the performance of a unity gain feedback loop. The minimisation of the sensitivity function is often an objective of a successful control design, and in the last section of this chapter we try to indicate why this might be the case.

Contents

8.1 Time-domain performance specifications

We begin with describing what is a common manner for specifying the desired behaviour of a control system. The idea is essentially that the system is to be given a step input, and the response is to have various specified properties. You will recall in Proposition 3.40, it was shown how to determine the step response by solving an initial value problem. In this chapter we will be producing many step responses, and in doing so we have used this initial value problem method. In any case, when one gives a BIBO stable input/output system a step input, it will, after some period of transient response, settle down to some steady-state value. With this in mind, we make some definitions.

8.1 Definition Let (N, D) be a BIBO stable SISO linear system in input/output form, let $1_{N,D}(t)$ be the step response, and suppose that $\lim_{t\to\infty} 1_{N,D}(t) = 1_{\text{ss}} \in \mathbb{R}$. (N, D) is ***steppable*** provided that $1_{\text{ss}} \neq 0$. For a steppable system (N, D), let $y(t)$ be the response to the reference $r(t) = 1(t)(1_{\text{ss}})^{-1}$ so that $\lim_{t\to\infty} y(t) = 1$. We call $y(t)$ the ***normalised step response***.

(i) The ***rise time*** is defined by

$$t_r = \sup_{\delta} \left\{ \delta \;\middle|\; y(t) \leq \tfrac{t}{\delta} \text{ for all } t \in [0, \delta] \right\}.$$

(ii) For $\epsilon \in (0, 1)$ the ***ϵ-settling time*** is defined by

$$t_{s,\epsilon} = \inf_{\delta} \left\{ \delta \;\middle|\; |y(t) - 1| < \epsilon \text{ for all } t \in [\delta, \infty) \right\}.$$

(iii) The ***maximum overshoot*** is defined by $y_{\text{os}} = \sup_t\{y(t) - 1\}$ and the ***maximum percentage overshoot*** is defined by $P_{\text{os}} = y_{\text{os}} \times 100\%$.

(iv) The ***maximum undershoot*** is defined by $y_{\text{us}} = \sup_t\{-y(t)\}$ and the ***maximum percentage undershoot*** is defined by $P_{\text{us}} = y_{\text{us}} \times 100\%$. •

For the most part, these definitions are self-explanatory once you parse the symbolism. A possible exception is the definition of the rise time. The definition we provide is not the usual one. The idea is that rise time measures how quickly the system reaches its steady-state value. A more intuitive way to measure rise time would be to record the smallest time at which the output reaches a certain percentage (say 90%) of its steady-state value. However, the definition we provide still gives a measure of the time to reach the steady-state value, and has the advantage of allowing us to state some useful results. In any event, a typical step response is shown in Figure 8.1 with the relevant quantities labelled.

Notice that in Definition 8.1 we introduce the notion of a steppable input/output system. The following result gives an easy test for when a system is steppable.

8.2 Proposition *A BIBO stable SISO linear system (N, D) in input/output form is steppable if and only if $\lim_{s\to 0} T_{N,D}(s) \neq 0$. Furthermore, in this case $\lim_{t\to\infty} 1_{N,D}(t) = T_{N,D}(0)$.*

Proof By the Final Value Theorem, Proposition E.9(ii), the limiting value for the step response is

$$\lim_{t\to\infty} 1_{N,D}(t) = \lim_{s\to 0} s\hat{1}_{N,D}(s) = \lim_{s\to 0} T_{N,D}(s).$$

Therefore $\lim_{t\to\infty} 1_{N,D}(t) \neq 0$ if and only if $\lim_{s\to 0} \hat{1}_{N,D}(s) \neq 0$, as specified. ■

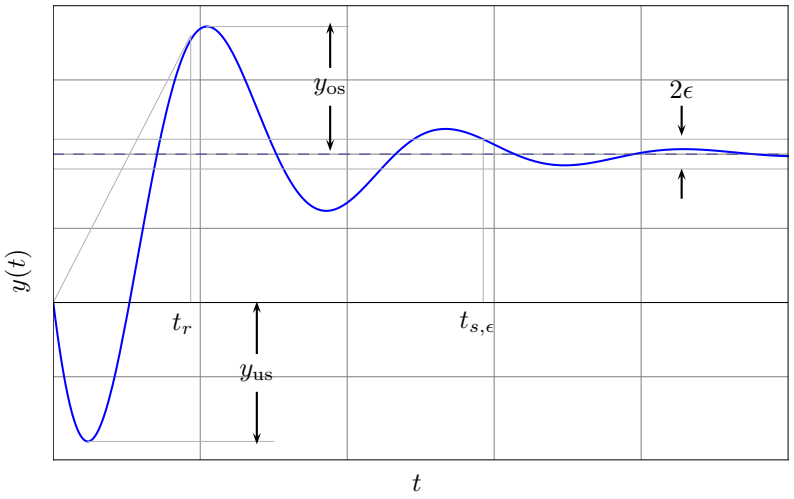


Figure 8.1 Performance parameters in the time-domain

So it is quite simple to identify systems which are not steppable. The idea here is quite simple. When $\lim_{s\to 0} T_{N,D}(s) = 0$ this means that the input only appears in the differential equation after being differentiated at least once. For a step input, this means that the right-hand side of the differential equation forming the initial value problem is zero, and since (N, D) is BIBO stable, the response must decay to zero. The following examples illustrate both sides of the story.

8.3 Examples

1. We first take $(N(s), D(s)) = (1, s^2+2s+2)$. By Proposition 3.40 the initial value problem to solve for the step response is

$$\ddot{1}_{N,D}(t) + 2\,\dot{1}_{N,D}(t) + 2\,1_{N,D}(t) = 1, \quad y(0) = 0,\ \dot{y}(0) = 0.$$

   The step response here is computed to be

$$1_{N,D}(t) = \tfrac{1}{2} - \tfrac{1}{2}e^{-t}(\cos t + \sin t).$$

   Note that $\lim_{t\to\infty} 1_{N,D}(t) = \frac{1}{2}$, and so the system is steppable. Therefore, we can compute the normalised step response, and it is

$$y(t) = 1 - e^{-t}(\cos t + \sin t).$$

2. Next we take $(N(s), D(s)) = (s, s^2 + s2 + 2)$. Again by Proposition 3.40, the initial value problem to be solved is

$$\ddot{1}_{N,D}(t) + 2\,\dot{1}_{N,D}(t) + 2\,1_{N,D}(t) = 0, \quad y(0) = 0,\ \dot{y}(0) = 1.$$

   The solution is $1_{N,D}(t) = e^{-t}\sin t$ which decays to zero, so the system is not steppable. This is consistent with Proposition 8.2 since $\lim_{s\to 0} T_{N,D}(s) = 0$. •

8.2 Performance for some classes of transfer functions

It is quite natural to *specify* performance criteria in the time-domain. However, in order to perform design, one needs to see how system parameters affect the time-domain performance,

and how various performance measures get reflected in the other representations of control systems—the Laplace transform domain and the frequency-domain. To get a feel for this, in this section we look at some concrete classes of transfer functions.

8.2.1 Simple first-order systems

The issues surrounding performance of control systems can be difficult to grasp, so our approach will be to begin at the beginning, describing what happens with simple systems, then moving on to talk about things more general. In this, the first of the three sections devoted to rather specific transfer functions, we will be considering the situation when the transfer function

$$T_L(s) = \frac{R_L(s)}{1 + R_L(s)}$$

has a certain form, without giving consideration to the precise form of the loop gain R_L. The simplest case one can deal with is one where the system transfer function is first order. If we normalise the transfer function so that it will have a steady-state response of 1 to a unit step input, the transfer functions under consideration look like

$$T_\tau(s) = \frac{1}{\tau s + 1},$$

and so essentially depend upon a single parameter τ. The step response of such a system is depicted in Figure 8.2. The parameter τ is exactly the rise time for a first-order system, at

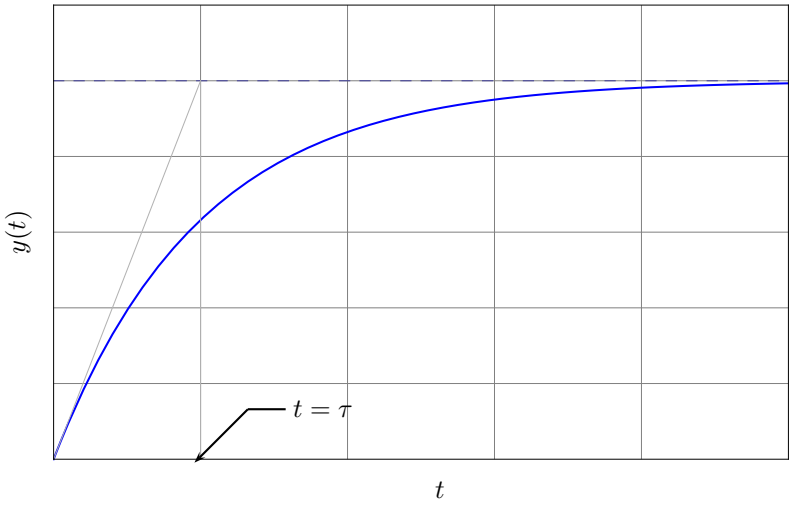


Figure 8.2 Step-response for typical first-order system

least by our definition of rise time. Thus, by making τ smaller, we ensure the system will more quickly reach its steady-state value.

Let's see how this is reflected in the behaviour of the poles of the transfer function, and in the Bode plot for the transfer function. The behaviour of the poles in this simple first-order case is straightforward. There is one pole at $s = -\frac{1}{\tau}$. Thus making τ smaller moves this pole further from the origin, and further into the negative complex plane. Thus the issues in designing a first-order transfer function are one: move the pole as far to the left as possible. Of course, not many transfer functions that one will obtain are first-order.

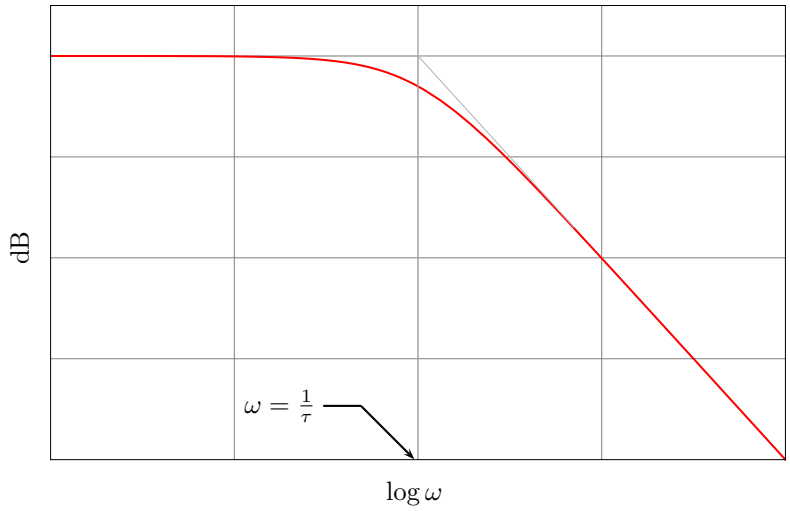


Figure 8.3 Bode plot for typical first-order system

The Bode plot for the transfer function T_τ is very simple, of course: it has a single break frequency at $\omega = \frac{1}{\tau}$. A typical such plot is shown in Figure 8.3. Obviously, the greater values of break frequency correspond to quicker response times. Thus, for simple first-order systems, the name of the game can be seen as pushing the break frequency as far as possible to the right. This is, in fact, a general strategy, and we begin an exploration of this by stating an obvious result.

8.4 Proposition *The magnitude of $T_\tau(s)$ at $s = \frac{i}{\tau}$ is $\frac{1}{\sqrt{2}}$.*

Proof This is a simple calculation since

$$T_\tau\left(\tfrac{i}{\tau}\right) = \frac{1}{i+1},$$

whose magnitude we readily compute to be $\frac{1}{\sqrt{2}}$. ■

We also compute $20 \log \frac{1}{\sqrt{2}} \approx -3.01$dB. Thus the magnitude of $T_\tau(\frac{i}{\tau})$ is approximately -3dB which one can easily pick off from a Bode plot. We call $\frac{1}{\tau}$ the ***bandwidth*** for the first-order transfer function. Note that for first-order systems, larger bandwidth translates to better performance, at least in terms of rise time.

8.2.2 Simple second-order systems

First-order systems are not capable of sustaining much rich behaviour, so let's see how second-order systems look. After again requiring a transfer function which produces a steady-state of 1 to a unit step input, the typical second-order transfer function we look at is

$$T_{\zeta,\omega_0}(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2},$$

which depends on the parameters ζ and ω_0. If we are interested in BIBO stable transfer functions, the Routh/Hurwitz criterion, and Example 5.35 in particular, then we can without loss of generality suppose that both ζ and ω_0 are positive.

The nature of the closed form expression for the step response depends on the value of ζ. In any case, using Proposition 3.40(ii), we ascertain that the step response is given by

$$y(t) = \begin{cases} 1 + \left(\frac{\zeta}{\sqrt{\zeta^2-1}} - 1\right)e^{(-\zeta-\sqrt{\zeta^2-1})\omega_0 t} - \left(\frac{\zeta}{\sqrt{\zeta^2-1}}\right)e^{(-\zeta+\sqrt{\zeta^2-1})\omega_0 t}, & \zeta > 1 \\ 1 - e^{\omega_0 t}(1+\omega_0 t), & \zeta = 1 \\ 1 - e^{-\zeta\omega_0 t}\left(\cos(\sqrt{1-\zeta^2}\omega_0 t) + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin(\sqrt{1-\zeta^2}\omega_0 t)\right), & \zeta < 1. \end{cases}$$

This is plotted in Figure 8.4 for various values of ζ. Note that for large ζ there is no

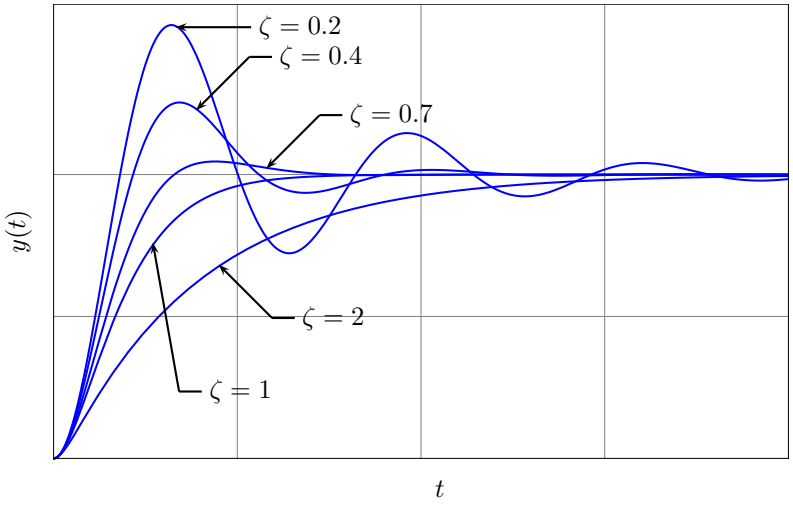


Figure 8.4 Step response for typical second-order system for various values of ζ

overshoot, but that as ζ decreases, the response develops an overshoot. Let us make some precise statements about the character of the step response of a second-order system.

8.5 Proposition *Let $y(t)$ be the step response for the SISO linear system (N, D) with transfer function T_{ζ,ω_0}. The following statements hold:*

(i) when $\zeta \geq 1$ there is no overshoot;

(ii) when $\zeta < 1$ there is overshoot, and it has magnitude $y_{\text{os}} = e^{-\pi\zeta/\sqrt{1-\zeta^2}}$ and occurs at time $t_{\text{os}} = \frac{\pi}{\sqrt{1-\zeta^2}\omega_0}$;

(iii) if $\zeta < 1$ the rise time satisfies

$$-(e^{\omega_0\zeta t_r}\sqrt{1-\zeta^2}) + \sqrt{1-\zeta^2}\cos(\omega_0\sqrt{1-\zeta^2}t_r) + (\omega_0 t_r + \zeta)\sin(\omega_0\sqrt{1-\zeta^2}t_r) = e^{-\omega_0\zeta t_r}t_r\sqrt{1-\zeta^2}.$$

Proof (i) and (ii) The proof here consists of differentiating the step response and setting it to zero to obtain those times at which it is zero. When $\zeta \geq 1$ the derivative is zero only when $t = 0$. When $\zeta < 1$ the derivative is zero when $\omega_0\sqrt{1-\zeta^2}t = n\pi$, $n \in \mathbb{Z}_+$. The smallest positive time will be the time of maximum overshoot, so we take $n = 1$, and from this the stated formulae follow.

(iii) The rise time t_r is the smallest positive time which satisfies

$$\dot{y}(t_0) = \frac{y(t_0)}{t_0},$$

since the slope at the rise time should equal the slope of the line through the points $(0, 0)$ and $(t_0, y(t_0))$. After some manipulation this relation has the form stated in the proposition. ■

In Figure 8.5 we plot the maximum overshoot and the time at which it occurs as a function of

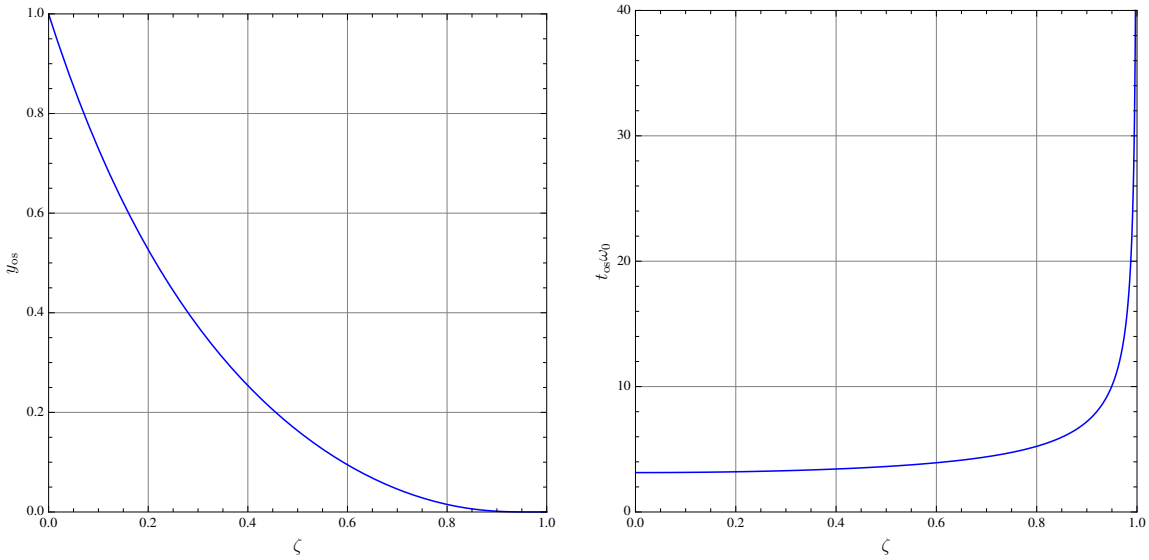


Figure 8.5 Maximum overshoot (left) and time of maximum overshoot (right) as functions of ζ

ζ. Note that the overshoot decreases as ζ increases, and that the time of overshoot increases as ζ increases.

The rise time t_r is not easily understood via an analytical formula. Nevertheless, we may numerically determine the rise time for various ζ and ω_0 by solving the equation in part (iii) of Proposition 8.5 and the results are shown in Figure 8.6. In that same figure we also plot the curve $t_r = \frac{9}{5\omega_0}$, and we see that it is a very good approximation—indeed it is indistinguishable in our plot from the curve for $\zeta = 0.9$. Thus, for second-order transfer functions $T_{\zeta,\omega_0}(s)$ with $\zeta < 1$, it is fair to use $t_r \approx \frac{9\omega_0}{5}$. Of course, for more general transfer functions, even more general second-order transfer functions (say, with nontrivial numerators), this relationship can no longer be expected to hold.

From the above discussion we see that overshoot is controlled by the damping factor ζ and rise time is essentially controlled by the natural frequency ω_0. However, we note that the time for maximum overshoot, t_{os}, depends only upon ζ, and indeed increases as ζ increases. Thus, there is a possible tradeoff to make when selecting ζ if one chooses a small ϵ in the ϵ-settling time specification. A commonly used rule of thumb is that one should choose $\zeta = \frac{1}{\sqrt{2}}$ which leads to $y_{\text{os}} = e^{-\pi} \approx 0.043$ and $t_{\text{os}}\omega_0 = \pi\sqrt{2} \approx 4.44$.

Let's look now to see how the poles of T_{ζ,ω_0} depend upon the parameters ζ and ω_0. One readily determines that the poles are $\omega_0(-\zeta \pm \sqrt{1-\zeta^2})$. As we have just seen, good response dictates that we should have poles with nonzero imaginary part, so we consider the situation when $\zeta < 1$. In this case, the poles are located as in Figure 8.7. Thus the poles lie on a circle

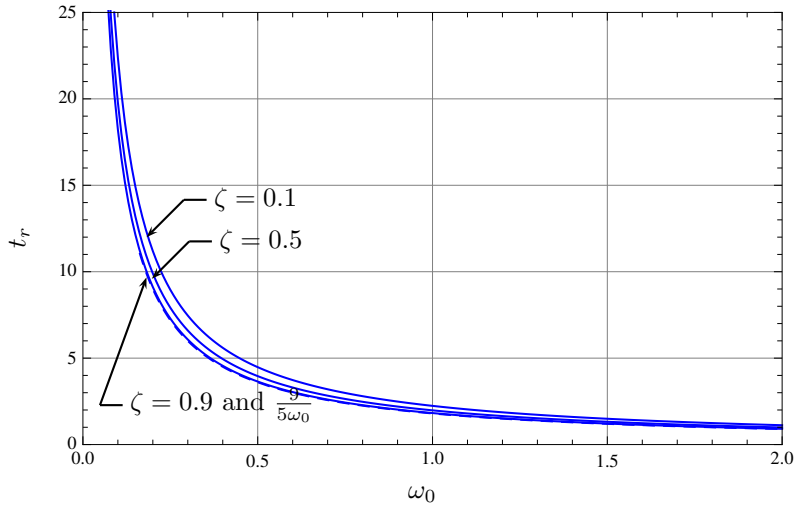


Figure 8.6 Behaviour of rise time versus ω_0 for various $\zeta < 1$

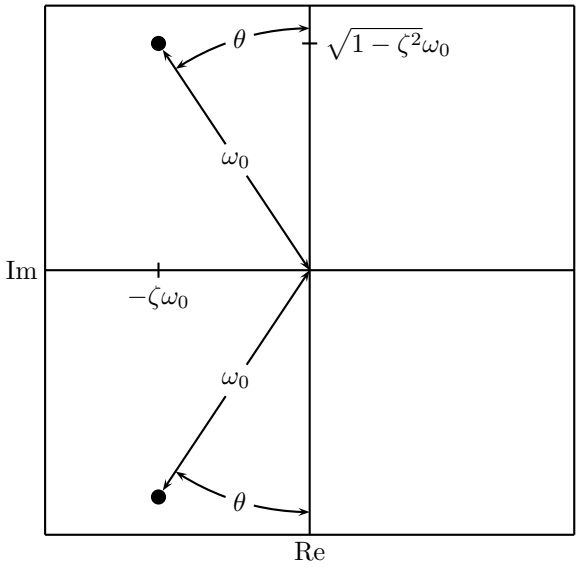


Figure 8.7 Pole locations for T_{ζ,ω_0} when $0 < \zeta < 1$

of radius ω_0, and the angle they make with the imaginary axis is $\sin^{-1}\zeta$. Our rule of thumb of taking $\zeta = \frac{1}{\sqrt{2}}$ then specifies that the poles lie on the rays emanating from the origin into $\mathbb{C}_-$ at an angle of 45° from the imaginary axis.

Finally, for second-order systems, let's see how the parameters ζ and ω_0 affect the Bode plot for the system. In Exercise E4.6 the essential character of the frequency response for T_{ζ,ω_0} is investigated, and let us just record the outcome of this.

8.6 Proposition *If* $H_{\zeta,\omega_0}(\omega) = T_{\zeta,\omega_0}(i\omega)$, *then the following statements hold:*

(i) $|H_{\zeta,\omega_0}(\omega)| = \dfrac{\omega_0^2}{\sqrt{(\omega_0^2-\omega^2)^2 + 4\zeta^2\omega_0^2\omega^2}}$;

(ii) $\angle H_{\zeta,\omega_0}(\omega) = \arctan\dfrac{-2\zeta\omega_0\omega}{\omega_0^2-\omega^2}$;

(iii) *for* $\zeta < \frac{1}{\sqrt{2}}$, *the function* $\omega \mapsto |H_{\zeta,\omega_0}(\omega)|$ *has a maximum of* $\frac{1}{2\zeta\sqrt{1-\zeta^2}}$ *at the frequency* $\omega_m = \omega_0\sqrt{1-2\zeta^2}$;

(iv) $\angle H_{\zeta,\omega_0}(\omega_m) = \arctan\dfrac{-\sqrt{1-2\zeta^2}}{\zeta}$.

In Figure 8.8 we label the typical points on the Bode plot for the transfer function T_{ζ,ω_0}

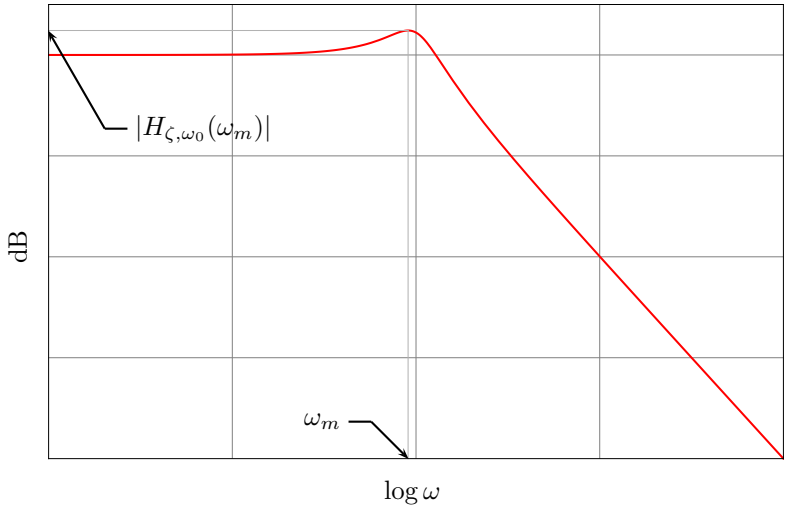


Figure 8.8 Typical Bode plot for second-order system with $\zeta < \frac{1}{\sqrt{2}}$

when $\zeta < \frac{1}{\sqrt{2}}$. As we decrease ζ the peak becomes larger, and shifted to the left. The phase, as we decrease ζ, tends to $-90°$ at the peak frequency ω_m.

Based on the discussion with first-order systems, for a second-order system with transfer function T_{ζ,ω_0}, the ***bandwidth*** is that frequency $\omega_{\zeta,\omega_0} > 0$ for which $|T_{\zeta,\omega_0}(i\omega_{\zeta,\omega_0})| = \frac{1}{\sqrt{2}}$. The following result gives an explicit expression for bandwidth of second-order transfer functions. Its proof is via direct calculation.

8.7 Proposition *When* $\zeta < \frac{1}{\sqrt{2}}$ *the bandwidth for* T_{ζ,ω_0} *satisfies*

$$\frac{\omega_{\zeta,\omega_0}}{\omega_0} = \sqrt{1-2\zeta^2+\sqrt{2(1-2\zeta^2)+4\zeta^4}}.$$

Thus we see that the bandwidth is directly proportional to the natural frequency ω_0. The dependence on ζ is shown in Figure 8.9. Thus we duplicate our observation for first-order systems that one should maximise the bandwidth to minimise the rise time. This is one of the general themes in control synthesis, namely that, all other things being constant, one should maximise bandwidth.

8.2.3 The addition of zeros and more poles to second-order systems

In our general buildup, the next thing we look at is the effect of adding to a second-order transfer function either a zero, i.e., making a numerator which has a root, or an additional pole. The idea is that we will investigate the effect that these have on the nature of the second-order response. This is carried out by Mulligan Jr. [1949]. detail

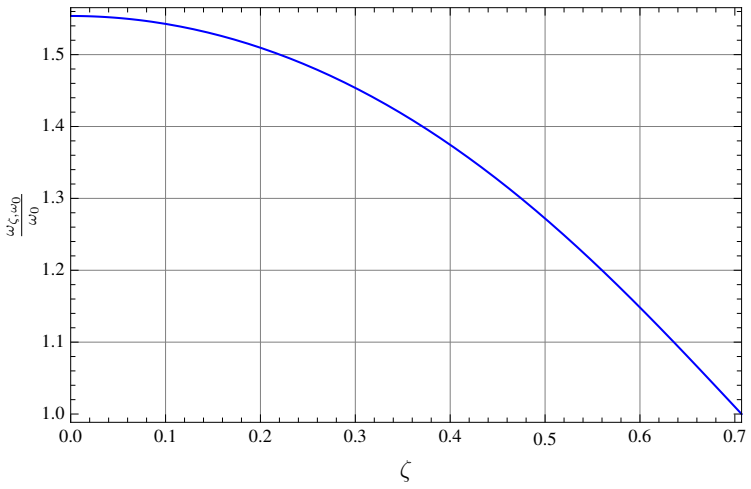


Figure 8.9 Dependence of bandwidth on ζ for second-order systems

Adding a zero

To investigate what happens to the time signal when we add a zero to a second-order system with a zero placed at $-\alpha\zeta\omega_0$. If we normalise the transfer function to that it has unit value at $s = 0$, we get

$$T(s) = \frac{\omega_0^2}{\alpha\zeta\omega_0} \frac{s + \alpha\zeta\omega_0}{s^2 + 2\zeta\omega_0 s + \omega_0^2}.$$

For concreteness we take $\zeta = \frac{1}{2}$ and $\omega_0 = 1$. The step responses and magnitude Bode plots are shown in Figure 8.10 for various values of α. We note a couple of things.

8.8 Remarks

1. The addition of a zero increases the overshoot for $\alpha < 3$, and dramatically so for $\alpha < 1$.
2. If the added zero is nonminimum phase, i.e., when $\alpha < 0$, the step response exhibits undershoot. Thus nonminimum phase systems have this property of reacting in a manner contrary to what we want, at least initially. This phenomenon will be explored further in Section 9.1.
3. The magnitude Bode plot is the same for $\alpha = -\frac{1}{2}$ as it is for $\alpha = \frac{1}{2}$. Where the Bode plots will differ is in the phase, as in the former case, the system is nonminimum phase.
4. When comparing the step response and the magnitude Bode plots for positive α's, one sees that the general tendency of larger bandwidths[1] to produce shorter rise times is preserved. •

Adding a pole

Now we look at the effect of an additional pole at $-\alpha\zeta\omega_0$. The normalised transfer function is

$$T(s) = \frac{\alpha\zeta\omega_0^3}{(s + \alpha\zeta\omega_0)(s^2 + 2\zeta\omega_0 s + \omega_0^2)}.$$

[1] We have not yet defined bandwidth for general transfer functions, although it is clear how to do so. The bandwidth, roughly, is the smallest frequency above which the magnitude of the frequency response remains below $\frac{1}{\sqrt{2}}$ times its zero frequency value. This is made precise in Definition 8.25.

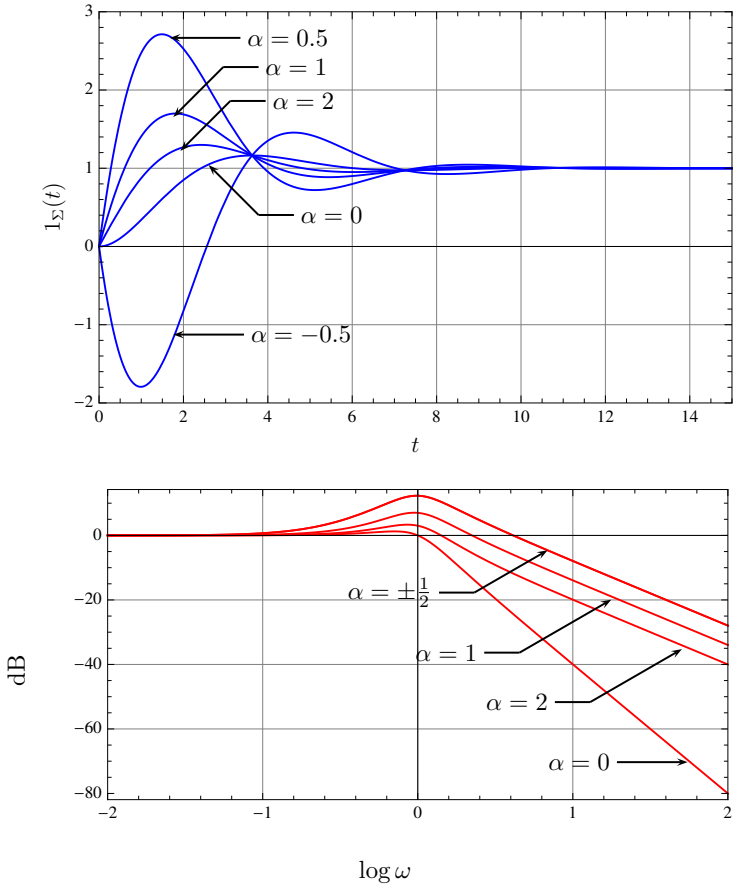


Figure 8.10 The effect of an added zero for $\alpha = -\frac{1}{2}, 0, \frac{1}{2}, 1, 2$

We once again fix $\zeta = \frac{1}{2}$ and $\omega_0 = 1$, and plot the step response for varying α in Figure 8.11. We make the following observations.

8.9 Remarks

1. If a pole is added with $\alpha < 3$, the rise time will be dramatically increased. This is also reflected in the bandwidth of the system increasing with α.
2. The larger bandwidths in this case are accompanied by a more pronounced peak in the Bode plot. As with second-order systems where this is a consequence of a smaller damping factor, we see that there is more overshoot. •

8.2.4 Summary

This section has been something of a mixed bag of examples and informal observations. We do not try to make it more than that at this point. Some of the things covered here have a more general and rigorous treatment in Chapter 9. However, it is worth summarising the gist of what we have said in an informal way. These are not theorems...

1. Increased bandwidth can mean shorter rise times.
2. In terms of poles, larger bandwidth sometimes means closed-loop poles that are far from the imaginary axis.

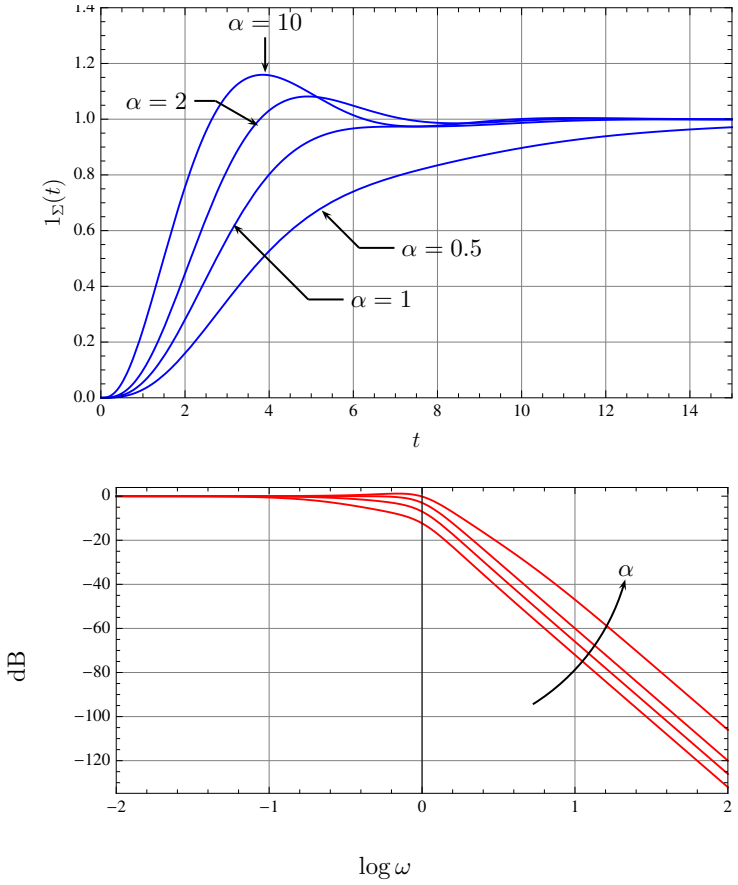


Figure 8.11 The effect of an added pole for $\alpha = \frac{1}{2}, 1, 2, 10$

3. Large overshoot can arise when the Bode plot exhibits a large "peak" at some frequency. This is readily seen for second-order systems, but can happen for other systems.
4. In terms of poles of the closed-loop transfer function, large overshoot can arise when poles are close to the imaginary axis, as compared to their distance from the real axis.
5. Zeros of the closed-loop transfer function lying in $\mathbb{C}_+$ can lead to undershoot in the step response, this having a deleterious effect on the system's performance.

These rough guidelines can be useful in predicting the behaviour of a system based upon the location of its poles, or on the shape of its frequency response. The former connection forms the basis for root-locus design which is covered in Chapter 11. The frequency response ideas we shall make much use of, as they form the basis for the design methodology of Chapters 12 and 15. It is existence of the rigorous mathematical ideas for control design in Chapter 15 that motivate the use of frequency response methods in design.

8.3 Steady-state error

An important consideration is that the difference between the reference signal and the output should be as small a possible. When we studied the PID controller in Section 6.5 we noticed that with an integrator it was possible to make at least certain types of transfer function have no steady-state error. Here we look at this in a slightly more systematic

manner. The first few subsections deal with descriptive matters.

8.3.1 System type for SISO linear system in input/output form

For a SISO system (N, D) in input/output form, a ***controlled output*** is a pair $(r(t), y(t))$ defined for $t \in [0, \infty)$ with the property that

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)r(t)$$

The ***error*** for a controlled output $(r(t), y(t))$ is $e(t) = r(t) - y(t)$, and the ***steady-state error*** is

$$\lim_{t\to\infty}(r(t) - y(t)),$$

and we allow the possibility that this limit may not exist. With this language, we have the following definition of system "type."

8.10 Definition Let $k \geq 0$. A signal $r \in \mathscr{U}$ defined on $[0, \infty)$ is of ***type k*** if $r(t) = Ct^k$ for some $C > 0$. A BIBO stable SISO linear system (N, D) is of ***type k*** if $\lim_{t\to\infty}(r(t) - y(t))$ exists and is nonzero for every controlled output $(r(t), y(t))$ with $r(t)$ of type k. •

In the definition of the type of a SISO linear system in input/output form, it is essential that the system be BIBO stable, i.e., that the roots of D all lie in $\mathbb{C}_-$. If this is not the case, then one might expect the output to grow exponentially, and so error bounds like those in the definition are not possible.

The following result attempts to flush out all the implications of system type.

8.11 Proposition *Let (N, D) be a BIBO stable SISO linear system in input/output form. The following statements are equivalent:*

(i) *(N, D) is of type k;*

(ii) *$\lim_{t\to\infty}(r(t) - y(t))$ exists and is nonzero for some controlled output $(r(t), y(t))$ with $r(t)$ of type k.*

(iii) *$\lim_{s\to 0} \dfrac{1}{s^k}\big(1 - T_{N,D}(s)\big)$ exists and is nonzero.*

The preceding three equivalent statements imply the following:

(iv) *$\lim_{t\to\infty}(r(t) - y(t)) = 0$ for every controlled output $(r(t), y(t))$ with $r(t)$ of type ℓ with $\ell < k$.*

Proof (i) $\iff$ (ii) That (i) implies (ii) is clear. To show the converse, suppose that $\lim_{t\to\infty}(\bar{r}(t) - \bar{y}(t)) = K$ for some nonzero K and for some controlled output $(\bar{r}(t), \bar{y}(t))$ with $\bar{r}(t)$ of type k. Suppose that $\deg(D) = n$ and let $y_1(t), \dots, y_n(t)$ be the n linearly independent solutions to $D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = 0$. If $\bar{y}_p(t)$ is a particular solution to $D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = \bar{r}(t)$ we must have

$$\bar{y}(t) = \bar{c}_1 y_1(t) + \cdots + \bar{c}_n y_n(t) + \bar{y}_p(t)$$

for some $\bar{c}_1, \dots, \bar{c}_n \in \mathbb{R}$. By hypothesis we then have

$$\lim_{t\to\infty}\big(\bar{r}(t) - \bar{c}_1 y_1(t) - \cdots - \bar{c}_n y_n(t) - \bar{y}_p(t)\big) = \lim_{t\to\infty}\big(\bar{r}(t) - \bar{y}_p(t)\big) = K.$$

Here we have used the fact that the roots of D are in $\mathbb{C}_-$ so the solutions $y_1(t), \dots, y_n(t)$ all decay to zero.

Now let $(r(t), y(t))$ be a controlled output with $r(t)$ be an arbitrary signal of type k. Note that we must have $r(t) = A\bar{r}(t)$ for some $A > 0$. We may take $y_p(t) = A\bar{y}_p(t)$ as a

particular solution to $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)r(t)$ by linearity of the differential equation. This means that we must have

$$y(t) = c_1 y_1(t) + \cdots + c_n y_n(t) + A\bar{y}_p(t)$$

for some $c_1, \ldots, c_n \in \mathbb{R}$. Thus we have

$$\begin{aligned}\lim_{t\to\infty}\big(r(t) - y(t)\big) &= \lim_{t\to\infty}\big(A\bar{r}(t) - c_1 y_1(t) - \cdots - c_n y_n(t) - A\bar{y}_p(t)\big)\\ &= \lim_{t\to\infty} A\big(\bar{r}(t) - \bar{y}_p(t)\big) = AK.\end{aligned}$$

This completes this part of the proof.

(ii) $\iff$ (iii) Let $(r(t), y(t))$ be a controlled output with $r(t) = \frac{t^k}{k!}$, and suppose that $y(0) = 0, y^{(1)}(0) = 0 \ldots, y^{(n-1)}(0) = 0$. Note that $\hat{r}(s) = \frac{1}{s^{k+1}}$. Taking the Laplace transform of the differential equation $D\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)y(t) = N\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)r(t)$ gives $D(s)\hat{y}(s) = N(s)\hat{r}(s)$. By Proposition E.9(ii) we have

$$\begin{aligned}\lim_{t\to\infty}\big(r(t) - y(t)\big) &= \lim_{s\to 0} s\big(\hat{r}(s) - \hat{y}(s)\big)\\ &= \lim_{s\to 0} s\big(\hat{r}(s) - T_{N,D}(s)\hat{r}(s)\big)\\ &= \lim_{s\to 0} s\frac{1 - T_{N,D}(s)}{s^{k+1}}\end{aligned}$$

from which we ascertain that

$$\lim_{t\to\infty}\big(r(t) - y(t)\big) = \lim_{s\to 0}\frac{1 - T_{N,D}(s)}{s^k}. \tag{8.1}$$

From this the result clearly follows.

(iii) $\implies$ (iv) Suppose that

$$\frac{1 - T_{N,D}(s)}{s^k} = K$$

for some nonzero constant K and let $\ell \in \{0, 1, \ldots, k-1\}$. Let $(r(t), y(t))$ be a controlled output with $r(t)$ a signal of type ℓ. Since the roots of D are in $\mathbb{C}_-$, we can without loss of generality suppose that $y(0) = 0, y^{(1)}(0) = 0, \ldots, y^{(n-1)}(0) = 0$. We then have

$$\begin{aligned}\lim_{t\to\infty}\big(r(t) - y(t)\big) &= \lim_{s\to 0}\frac{1 - T_{N,D}(s)}{s^\ell}\\ &= \lim_{s\to 0} s^{k-\ell}\frac{1 - T_{N,D}(s)}{s^k}\\ &= K\lim_{s\to 0} s^{k-\ell} = 0.\end{aligned}$$

This completes the proof. ■

Let us examine the consequences of this result by making a few observations.

8.12 Remarks

1. Although we state the definition for systems in input/output form, it obviously applies to SISO linear systems and to interconnected SISO linear systems since these give rise to systems in input/output form after simplification of their transfer functions.
2. The idea is that a system of type k can track up to a constant error a reference signal which is a polynomial of degree k. Thus, for example, a system of type 0 can track a step input up to a constant error. A system of type 1 can track a ramp input up to a constant error, and can exactly track a step input for large time. •

Let's see how this plays out for some examples.

8.13 Examples In each of these examples we look at a transfer function, decide what is its type, and plot its response to inputs of various types.

1. We take
$$T_{N,D}(s) = \frac{1}{s^2 + 3s + 2}.$$
This transfer function is type 0, as may be determine by checking the limit of part (iii) of Proposition 8.11. For a step reference, ramp reference, and parabolic reference,
$$r_1(t) = \begin{cases} 1, & t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$
$$r_2(t) = \begin{cases} t, & t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$
$$r_3(t) = \begin{cases} t^2, & t \geq 0 \\ 0, & \text{otherwise,} \end{cases}$$
respectively, we may ascertain using Proposition 3.40, that the step, ramp, and parabolic responses are
$$y_1(t) = \tfrac{1}{2} + \tfrac{1}{2}e^{-2t} - e^{-t}, \quad y_2(t) = \tfrac{1}{4}(2t-3) - \tfrac{1}{4}e^{-2t} + e^{-t},$$
$$y_3(t) = \tfrac{1}{4}(2t^2 - 6t + 7) + \tfrac{1}{4}e^{-2t} - 2e^{-t},$$
and the errors, $e_i(t) = r_i(t) - y_i(t)$, $i = 1, 2, 3$, are plotted in Figure 8.12. Notice that the step error response has a nonzero limit as $t \to \infty$, but that the ramp and parabolic responses grow without limit. This is what we expect from a type 0 system.
2. We take
$$T_{N,D}(s) = \frac{2}{s^2 + 3s + 2}$$
which has type 1, using Proposition 8.11(iii). The step, ramp, and parabolic responses are
$$y_1(t) = 1 + e^{-2t} - 2e^{-t}, \quad y_2(t) = \tfrac{1}{2}(2t-3) - \tfrac{1}{2}e^{-2t} + 2e^{-t},$$
$$y_3(t) = \tfrac{1}{2}(2t^2 - 6t + 7) + \tfrac{1}{2}e^{-2t} - 4e^{-t},$$
and the errors are shown in Figure 8.13. Since the system is type 1, the step response gives zero steady-state error and the ramp response has constant steady-state error. The parabolic input gives a linearly growing steady-state error.

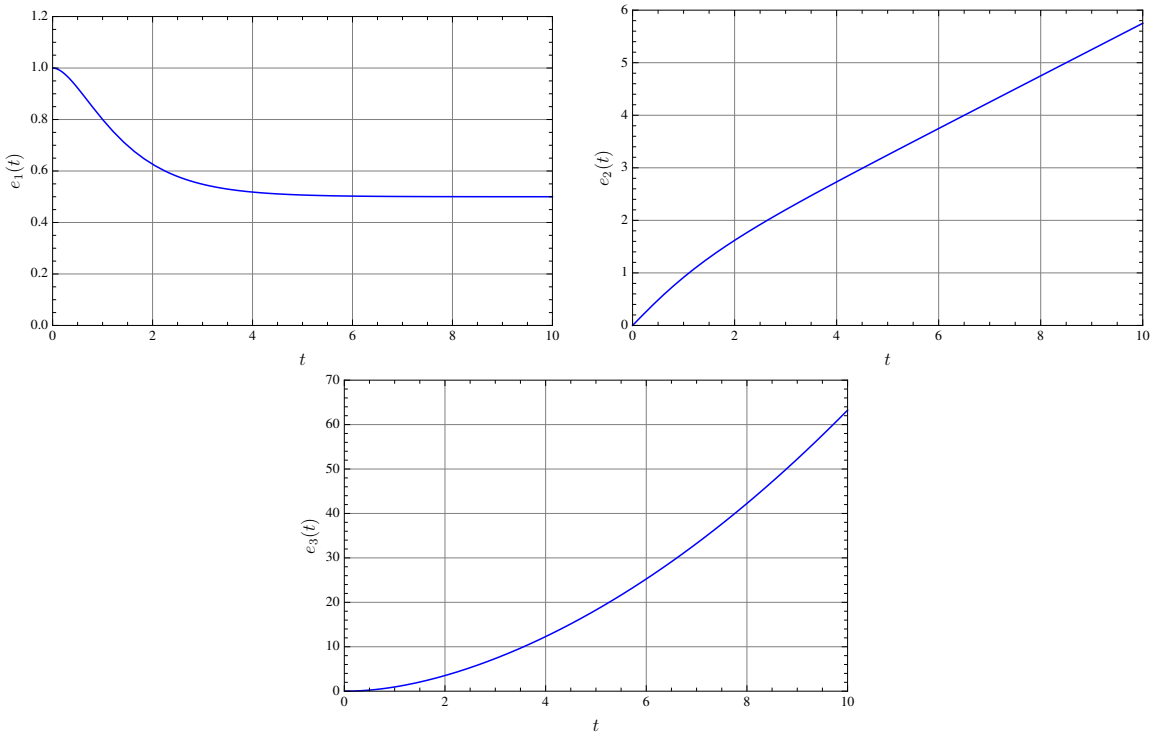


Figure 8.12 Step (top left), ramp (top right), and parabolic (bottom) error responses for a system of type 0

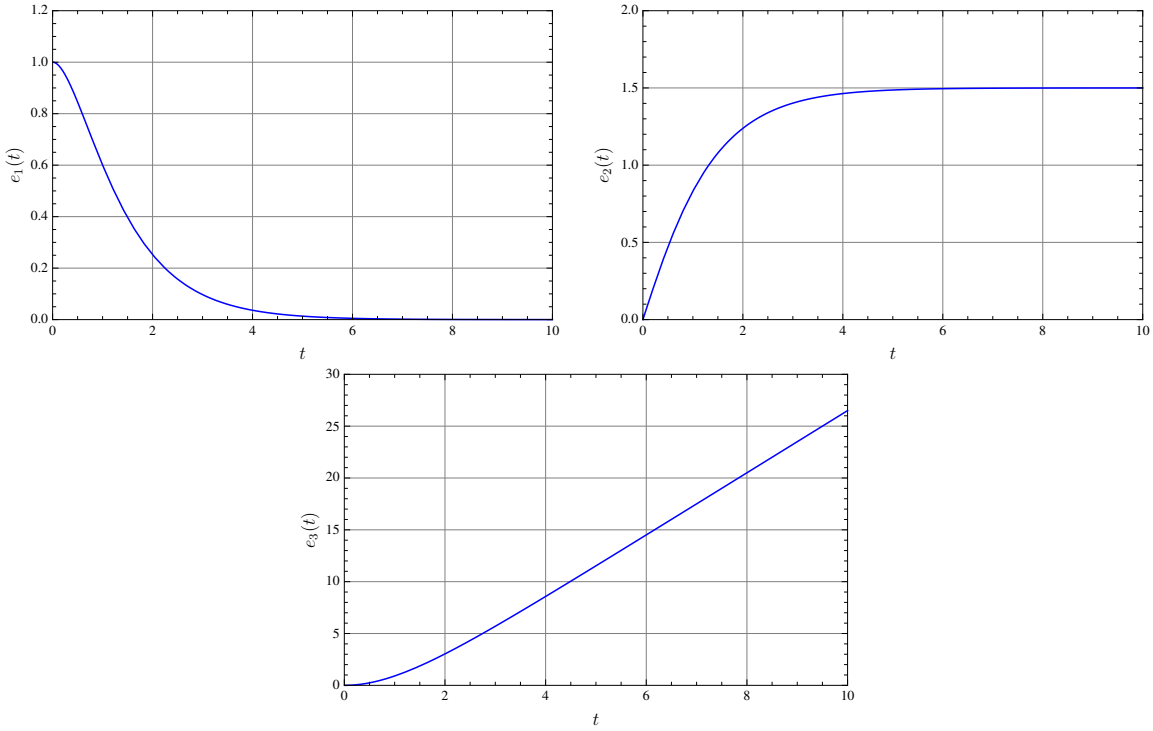


Figure 8.13 Step (top left), ramp (top right), and parabolic (bottom) responses for a system of type 1

3. The last example we look at is that with transfer function

$$T_{N,D}(s) = \frac{3s+2}{s^2+3s+2}.$$

This transfer function is type 2. The step, ramp, and parabolic responses are

$$y_1(t) = 1 - 2e^{-2t} + e^{-t}, \quad y_2(t) = t + e^{-2t} - e^{-t}, \quad y_3(t) = t^2 - 1 - e^{-2t} + 2e^{-2t}.$$

The errors are plotted in Figure 8.14. Note that the step and ramp steady-state errors

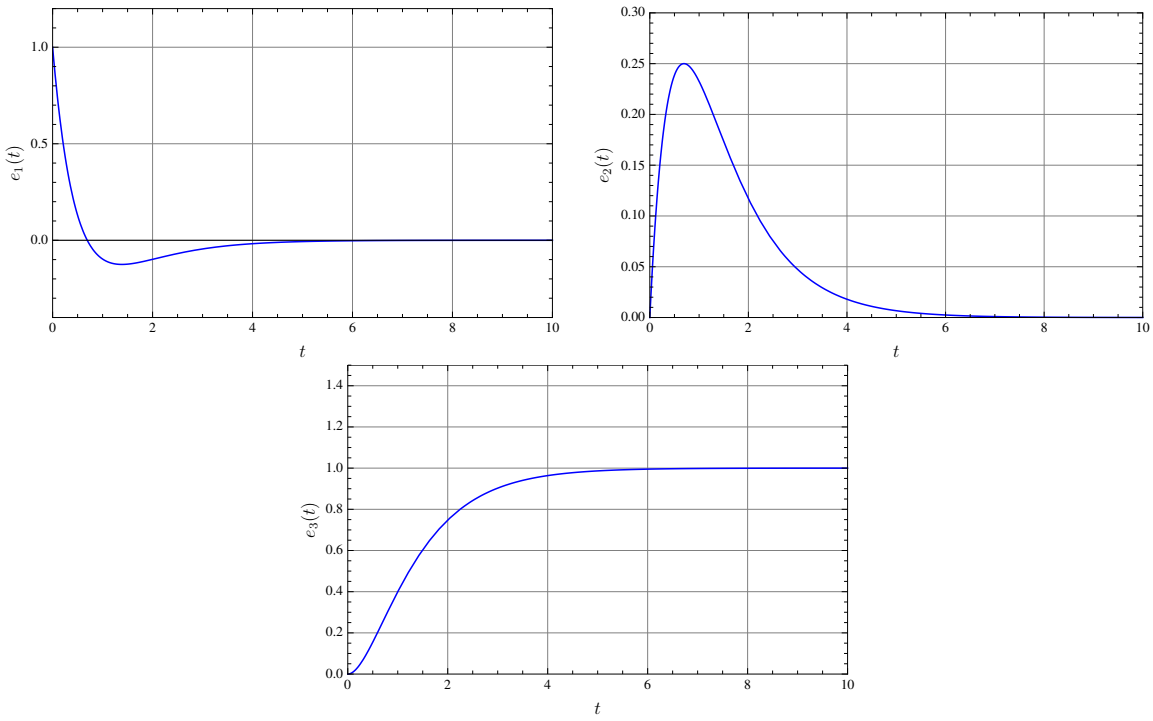


Figure 8.14 Step (top left), ramp (top right), and parabolic (bottom) responses for a system of type 2

shrink to zero, but the parabolic response has a constant error. •

8.3.2 System type for unity feedback closed-loop systems

To see how the steady-state error is reflected in a simple closed-loop setting, let us look at the situation depicted originally in Figure 6.25, and reproduced in Figure 8.15. Thus we are not thinking here so much about having a controller and a plant, but as combining these to get the transfer function R_L which is the loop gain in this case. In any event, we may directly give conditions on the transfer function R_L to determine the type of the closed-loop transfer function

$$T_L(s) = \frac{R_L(s)}{1 + R_L(s)}.$$

These conditions are as follows.

8.14 Proposition *Let $R_L \in \mathbb{R}(s)$ be proper and define*

$$T_L(s) = \frac{R_L(s)}{1 + R_L(s)}.$$

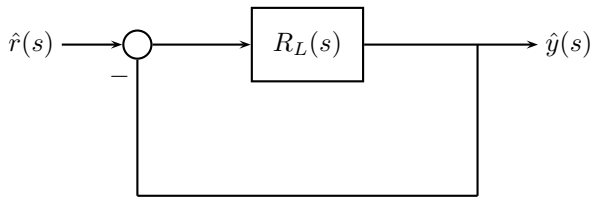


Figure 8.15 Unity gain feedback loop for investigating steady-state error

If (N, D) denotes the c.f.r. of T_L, then (N, D) is of type $k > 0$ if and only if $\lim_{s\to 0} s^k R_L(s)$ exists and is nonzero. (N, D) is of type 0 if and only if $\lim_{s\to 0} R_L(s)$ exists and is not equal to -1.

Proof We compute

$$1 - T_L(s) = \frac{1}{1 + R_L(s)}.$$

Thus, by Proposition 8.11(iii), (N, D) is of type k if and only if

$$\lim_{s\to 0} \frac{1}{s^k(1 + R_L(s))}$$

exists and is nonzero. For $k > 0$ we have

$$\lim_{s\to 0} \frac{1}{s^k(1 + R_L(s))} = \frac{1}{\lim_{s\to 0} s^k R_L(s)},$$

and the result follows directly in this case. For $k = 0$ we have

$$\lim_{s\to 0} \frac{1}{s^k(1 + R_L(s))} = \frac{1}{1 + \lim_{s\to 0} R_L(s)}.$$

Thus, provided that $R_L(0) \neq -1$ as hypothesised, the system is of type 0 if and only if $\lim_{s\to 0} R_L(s)$ exists. ■

The situation here, then, is quite simple. If $R_L(s)$ is proper, for some $k \geq 0$ we can write

$$R_L(s) = \frac{N_L(s)}{s^k D_L(s)}$$

with D_L monic, D_L and N_L coprime, and $D_L(0) \neq 0$. Thus we factor from the denominator as many factors of s as we can. Each such factor is an integrator. The situation is depicted in Figure 8.16. One can see, for example, why often the implementation of a PID control law (with integration as part of the implementation) will give a type 1 closed-loop system. To be precise, we can state the following.

8.15 Corollary *For the unity gain feedback loop of Figure 8.15, the closed-loop system is of type $k > 0$ if and only if there exists $R \in \mathbb{R}(s)$ with the properties*

(i) $R(0) \neq 0$ *and*

(ii) $R_L(s) = \frac{1}{s^k} R(s)$.

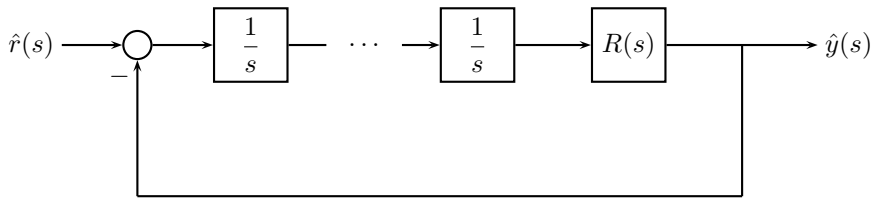


Figure 8.16 A unity feedback loop with k integrators and $R(0) \neq 0$

Furthermore, if R_L is the product of a plant transfer function R_P with a controller transfer function $R_C(s) = K(1 + T_D s + \frac{1}{T_I s})$, then the closed-loop system will be of type 1 provided that $\lim_{s\to 0} R_P(s)$ exists and is nonzero.

Proof We need only prove the second part of the corollary as the first is a direct consequence of Proposition 8.14. The closed-loop transfer function T_L satisfies

$$1 - T_L(s) = \frac{1}{1 + K(1 + T_D s + \frac{1}{T_I s})R_P(s)},$$

from which we determine that

$$\lim_{s\to 0} \frac{1 - T_L(s)}{s} = \frac{1}{\lim_{s\to 0} K(s + T_D s^2 + \frac{1}{T_I})R_P(s)},$$

and from this the result follows, since if $\lim_{s\to 0} R_P(s)$ exists and is nonzero, then $\lim_{s\to 0} s^\ell R_P(s) = 0$ for $\ell > 0$. ■

Interestingly, there is more we can say about type k systems when $k \geq 1$. The following result gives more a detailed description of the steady-state error in these cases.

8.16 Proposition *Let $y(t)$ be the normalised step response for the closed-loop system depicted in Figure 8.15. The following statements hold:*

(i) if the closed-loop system is type 1 with $\lim_{s\to 0} sR_L(s) = C$ with C a nonzero constant, then

$$\int_0^\infty e(t)\,\mathrm{d}t = \tfrac{1}{C};$$

(ii) if the closed-loop system is type k with $k \geq 2$ then

$$\int_0^\infty e(t)\,\mathrm{d}t = 0.$$

Proof (i) Since $\lim_{t\to\infty} e(t) = 0$, $\hat{e}(t)$ must be strictly proper. Therefore, by Proposition E.10, taking $s_0 = 0$, we have

$$\int_0^\infty e(t)\,\mathrm{d}t = \lim_{s\to 0} \hat{e}(s).$$

Since

$$\hat{e}(s) = (1 - T_L(s))\hat{r}(s) = \frac{s}{s + sR_L(s)}\frac{1}{s},$$

we have $\lim_{s\to 0} \hat{e}(s) = \frac{1}{C}$.

(ii) The idea here is the same as that in the previous part of the result except that we have

$$\hat{e}(s) = (1 - T_L(s))\hat{r}(s) = \frac{s^k}{s^k + s^k R_L(s)} \frac{1}{s},$$

with $k \geq 2$, and so $\lim_{s \to 0} \hat{e}(s) = 0$. ■

The essential point is that the proposition will hold for *any* loop gain R_L of type 1 (for part (i)) or type 2 (for (ii)). An interesting consequence of the second part of the proposition is the following.

8.17 Corollary *Let $y(t)$ be the normalised step response for the closed-loop system depicted in Figure 8.15. If the closed-loop transfer system is type k for $k \geq 2$, then $y(t)$ exhibits overshoot.*

Proof Since the error starts at $e(0) = 1$, in order that

$$\int_0^\infty e(t)\, \mathrm{d}t = 0,$$

the error must at some time be negative. However, negative error means overshoot. ■

This issue of determining the behaviour of a closed-loop system which depends only on properties of the loop gain is given further attention in Section 9.1. Here the effect of unstable poles and nonminimum phase zeros for the plant is flushed out in a general setting.

8.3.3 Error indices

finish From standard texts (see Truxal, page 83).

8.3.4 The internal model principle

To this point, the discussion has centred around tracking type k signals, i.e., those that are powers of t. However, one often wishes to track more "exotic" signals. The manner for doing so is suggested, upon reflection, by the discussion to this point, and is loosely called finish the "internal model principle."

8.4 Disturbance rejection

Our goal in this section is to do something rather similar to what we did in the last section, except that we wish to look at the relationships between disturbances and the output. Since a disturbance may enter a system in any of the signals, the appropriate setting here is that of an interconnected SISO linear system $(\mathcal{S}, \mathcal{G})$. We will suppose that we are dealing with such a system and that there are m nodes with the input being node 1 and the output being node m.

We need a notion of output like that introduced in the previous section when talking about transfer function types. We let $i \in \{2, \ldots, m\}$ and let $(\mathcal{S}_i, \mathcal{G}_i)$ be the ith-appended system. A ***j*-disturbed output with input at node *i*** is a pair $(d(t), y(t))$ defined on $[0, \infty)$ where $y(t)$ is the response at node j for the input $d(t)$ at node i for the interconnected SISO linear system $(\mathcal{S}_i, \mathcal{G}_i)$, where the input at node 1 is taken to be zero.

8.18 Definition Let $(\mathcal{S}, \mathcal{G})$ be an IBIBO stable interconnected SISO system as above and let $j \in \{2, \dots, m\}$. The system is of ***disturbance type k at node j with input at node i*** if $\lim_{t\to\infty} y(t)$ exists and is nonzero for every j-disturbed output $(d(t), y(t))$ with input at node i, where $d(t)$ is of type k. •

The idea here is very simple. We wish to allow a disturbance at any node which may enter the system at any other node. Just where the disturbance enters a system and where it is measured can drastically change the behaviour of the system. Typically, however, one measures the disturbance at the output of the interconnected system. That is, typically in the above discussion j is taken to be m. Before getting to an illustration by example, let us provide a result which gives some obvious consequences of the notion of disturbance type. The following result can be proved along the same lines as Proposition 8.11 and by application of the Final Value Theorem.

8.19 Proposition *Let $(\mathcal{S}, \mathcal{G})$ be an IBIBO stable interconnected SISO linear system with input in node 1 and output in node m. For $j \in \{2, \dots, m\}$ the following statements are equivalent:*

(i) *$(\mathcal{S}, \mathcal{G})$ is of j-disturbance type k with input at node i;*

(ii) *$\lim_{t\to\infty} y(t)$ exists and is nonzero for some j-disturbed output $(d(t), y(t))$ with input at node i, where $d(t)$ is of type k;*

(iii) *$\lim_{s\to 0} \frac{1}{s^k} T_{ji}(s)$ exists and is nonzero.*

The preceding three equivalent statements imply the following:

(iv) *$\lim_{t\to\infty} y(t) = 0$ for every j-disturbed output $(d(t), y(t))$ with $d(t)$ of type ℓ with $\ell < k$.*

Now let us consider a simple example.

8.20 Example Let us consider the interconnected SISO linear system whose signal flow graph is shown in Figure 8.17. The system has three places where disturbances ought to naturally

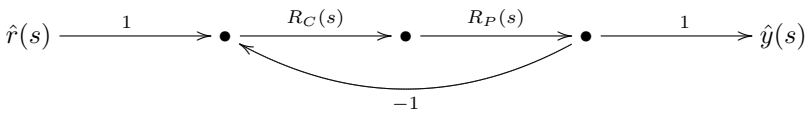


Figure 8.17 A system ready to be disturbed

be considered, as shown in Figure 8.18. Note that one should strictly add another node to

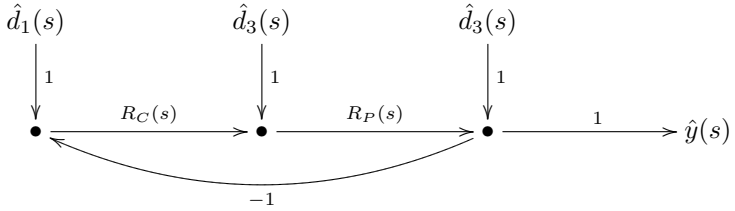


Figure 8.18 Disturbances

the output node $\hat{y}$, but this does not change anything, so it can be safely omitted. This will

generally be the case for an interconnected SISO linear system. In any case, the transfer function for the signal flow graph is

$$T_{\mathcal{S},\mathcal{G}}(s) = \frac{\hat{y}(s)}{\hat{r}(s)} = \frac{R_C(s)R_P(s)}{1 + R_C(s)(s)R_P(s)},$$

the transfer function from $\hat{d}_1$ to $\hat{y}$ is

$$T_1(s) = \frac{\hat{y}(s)}{\hat{d}_1(s)} = \frac{R_C(s)R_P(s)}{1 + R_C(s)R_P(s)},$$

the transfer function from $\hat{d}_2$ to $\hat{y}$ is

$$T_2(s) = \frac{\hat{y}(s)}{\hat{d}_2(s)} = \frac{R_P(s)}{1 + R_{C_1}(s)R_{C_2}(s)R_P(s)},$$

and the transfer function from $\hat{d}_3$ to $\hat{y}$ is

$$T_3(s) = \frac{\hat{y}(s)}{\hat{d}_3(s)} = \frac{1}{1 + R_C(s)R_P(s)}.$$

Let's attach some concrete transfer functions at this point. Our transfer functions are pretty artificial, but serve to illustrate the point. We take

$$R_C(s) = \frac{1}{s}, \quad R_P(s) = \frac{s}{s+2}.$$

Check We compute the corresponding transfer functions to be

$$T_{\mathcal{S},\mathcal{G}}(s) = \frac{s^2}{s^2+s+2}, \quad T_1(s) = \frac{s^2}{s^2+s+2}, \quad T_2(s) = \frac{s}{s^2+s+2}, \quad T_3(s) = \frac{s+2}{s^2+s+2}.$$

A straightforward application of Proposition 8.11(iii) indicates that $T_{\mathcal{S},\mathcal{G}}$ is of type 0. It is also easy to see from Proposition 8.19(iii) that the system is of disturbance type 2 with respect to the disturbance d_1, of disturbance type 1 with respect to the disturbance d_2, and of disturbance type 0 with respect to d_3. •

As we see in this example, the ability of a system to reject disturbances may differ from the ability of a system to represent a reference signal. Furthermore, for disturbance rejection, one typically wants the transfer function from the disturbance to the output to have the property that the expected disturbance type gives zero steady-state error. Thus, if step disturbances are what is expected, it is sufficient that the system be of disturbance type 0.

Let us look at our falling mass example thinking of the gravitational force as a disturbance.

8.21 Example **(Example 6.60 cont'd)** The block diagram for the falling mass with the gravitational force is shown in Figure 8.19. We think of the input $-mg$ as a disturbance input and so write $d(t) = -mg1(t)$. This input is then a step input if we think of holding the mass then letting it go. In fact, in our analysis, let's agree to take $y(0) = 0$ and $\dot{y}(0) = 0$ and see

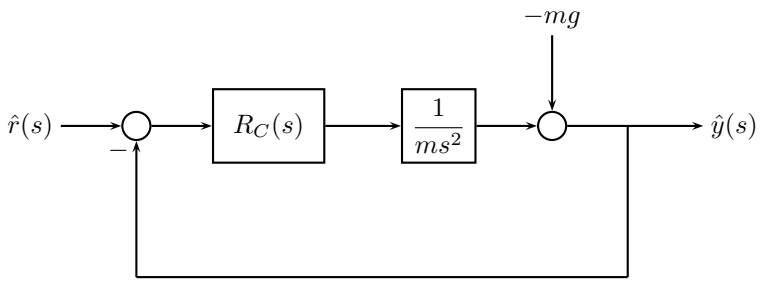


Figure 8.19 Block diagram for falling mass with gravitational force

if the controller can return the mass to $y = 0$ after we let it fall under gravity. The transfer function from the disturbance to the output is

$$T_d(s) = \frac{\hat{y}(s)}{\hat{d}(s)} = \frac{ms^2}{ms^2 + R_C(s)}.$$

Let us look at the form of this transfer function for the various controllers we employed in Example 6.60.

1. With proportional control we have $R_C(s) = K$ and so

$$T_d(s) = \frac{ms^2}{ms^2 + K}.$$

   This transfer function has poles on the imaginary axis, and so is not a candidate for having its type defined. Nonetheless, we can compute the time response using Proposition 3.40 given $\hat{d}(s) = -\frac{mg}{s}$. We ascertain that the output to this disturbance is

$$y_d(t) = -mg \cos \sqrt{\frac{K}{m}} t.$$

   One can see that the system does not respond very nicely in this case to the step disturbance, and to further illustrate the point, we plot this step response in Figure 8.20 for $m = 1$, $g = 9.81$, and $K = 28$.

2. For derivative control we take $R_C(s) = KT_D s$ which gives

$$T_d(s) = \frac{ms}{ms + KT_D}.$$

   This disturbance transfer function is type 1 and so the steady-state disturbance to the step input $d(t) = -mg$ is zero. Using Proposition 3.40 we determine that the response is

$$y_d(t) = -mge^{-KT_D t/m}$$

   and we plot this response in Figure 8.21 for $m = 1$, $g = 9.81$, $K = 28$, and $T_D = \frac{9}{28}$. Note that this response decays to zero which is consistent with our observation that the error to a step disturbance should decay to zero in the steady-state.

3. If we use an integral controller we take $R_C(s) = \frac{1}{T_I s}$ from which we compute

$$T_d(s) = \frac{mT_I s^3}{mT_I s^3 + K}.$$

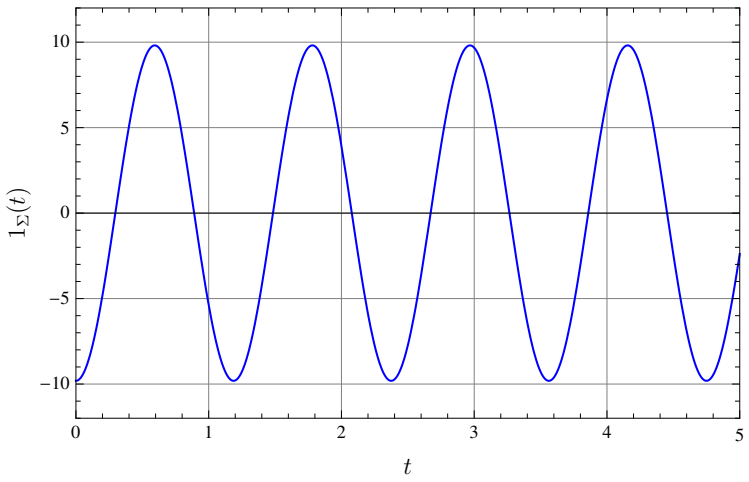


Figure 8.20 Response of falling mass to step disturbance with proportional control

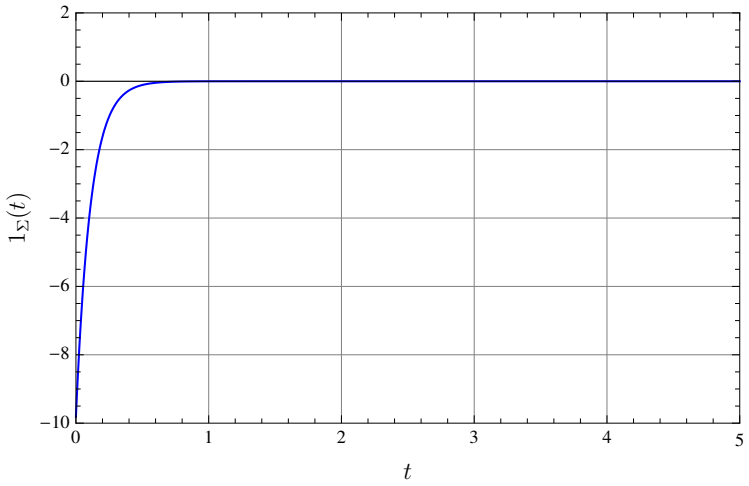


Figure 8.21 Response of falling mass to step disturbance with derivative control

To examine this transfer function for type, let us be concrete and choose $m = 1$, $g = 9.81$, $K = 28$, $T_D = \frac{9}{28}$, and $T_I = \frac{7}{10}$. One determines (Mathematica®!) that the roots of $mT_I s^3 + K$ are then

$$\sqrt[3]{5} \pm \sqrt{3}\sqrt[3]{5}i, \quad -2\sqrt[3]{5}.$$

Since the transfer function has poles in $\mathbb{C}_+$, the notion of type is not applicable. We plot the response to the step gravitational disturbance in Figure 8.22, and we see that it is badly behaved. Thus the given integral controller will magnify the disturbance.

4. Finally we combine the three controllers into a nice PID control law: $R_C(s) = K + KT_D s + \frac{K}{T_I s}$. The disturbance to output transfer function is then

$$T_d(s) = \frac{mT_I s^3}{mT_I s^3 + KT_D T_I s^2 + KT_I s + K}.$$

For our parameters $m = 1$, $g = 9.81$, $K = 28$, $T_D = \frac{9}{28}$, and $T_I = \frac{7}{10}$, we determine that the roots of the denominator polynomial are $-5, -2 \pm 2i$ (recall that we had chosen

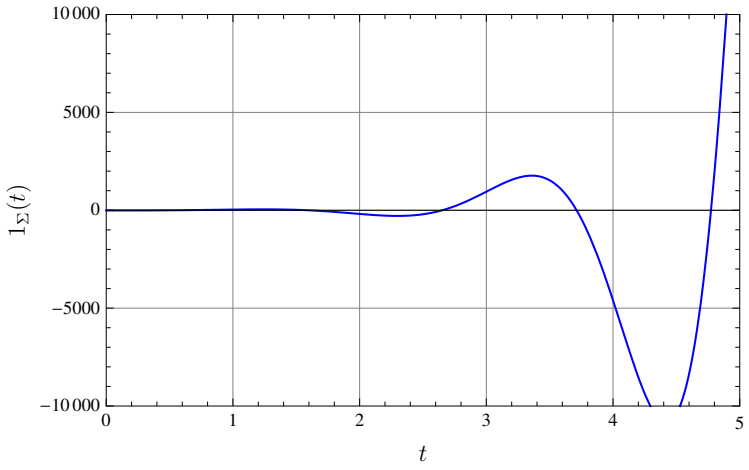


Figure 8.22 Response of falling mass to step disturbance with integral control

the PID parameters in just this way). Since these roots are all in $\mathbb{C}_-$, we may make the observation that the disturbance type is 3. Therefore the steady-state error to a step disturbance should be zero. The response of this transfer function to the gravitational step disturbance is shown in Figure 8.23 for $m = 1$, $g = 9.81$, $K = 28$, $T_D = \frac{9}{28}$, and

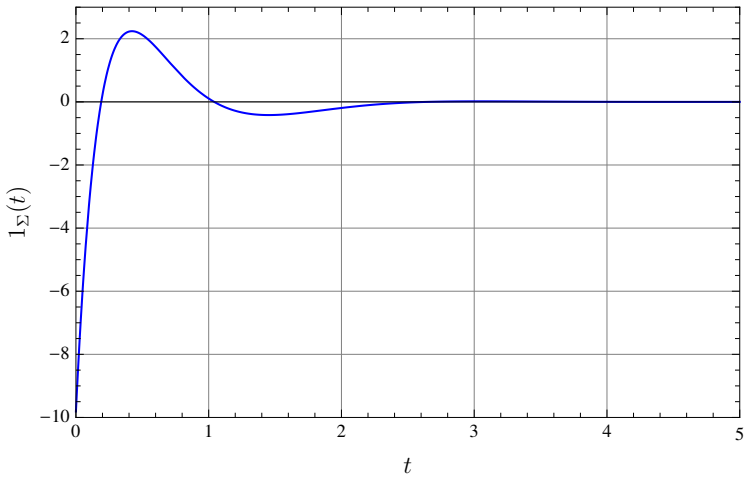


Figure 8.23 Response of falling mass to step disturbance with PID control

$T_I = \frac{7}{10}$. This response decays to zero as it ought to.

Let us address the question of how to interpret our computations for the gravitational disturbance to the falling mass. One needs to be careful not to misinterpret the disturbance response for the system response. The input/output response of the system for the various controllers was already investigated in Example 6.60. When examining the system including the effects of the disturbance, one must take into account both the input/output response and the response to the disturbance.

For example, with the PID controller we had chosen, our analysis of both these responses suggests that the system with the chosen parameters should behave nicely with the gravitational force. To verify this, let us look at the situation where we hold the mass still at $y = 0$

and at time $t = 0$ let it go. Our PID controller is then charged with bringing the mass back to $y = 0$. The initial value problem governing this situation is

$$m\ddot{y} + KT_D\dot{y}(t) + Ky(t) + \frac{K}{T_I}\int_0^t y(\tau)\,\mathrm{d}\tau = -mg, \quad y(0) = 0,\ \dot{y}(0) = 0.$$

To solve this equation we differentiate once to get the initial value problem

$$m\dddot{y} + KT_D\ddot{y}(t) + K\dot{y}(t) + \frac{K}{T_I}y(t) = 0, \quad y(0) = 0,\ \dot{y}(0) = 0,\ \ddot{y}(0) = -g.$$

The solution is plotted in Figure 8.24 for the chosen PID parameters. As should be the case,

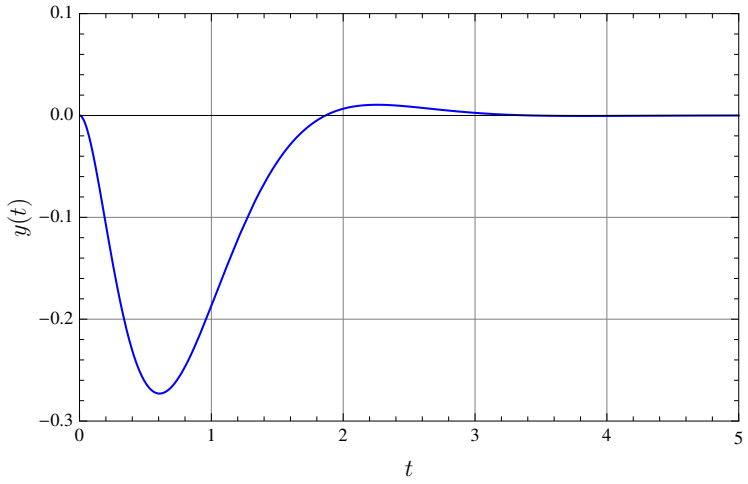


Figure 8.24 Response of falling mass with PID controller

the response is nice with the PID controller bringing the mass back to its initial height in relatively short order.

This example is one which contains a lot of information, so you would be well served to spend some time thinking about it. •

8.5 The sensitivity function

In this section we introduce and study a transfer function which is, at least for certain block diagram configurations, "complementary" to the transfer function. The block diagram configuration we look at is the one we first looked at in Section 6.3.2, and the configuration is reproduced in Figure 8.25. One may extend the discussion here to more general block diagram configurations, but this unity gain feedback setup is one which can be handled nicely. In this scenario, we are typically supposing that the loop gain R_L is the product of a plant transfer function R_P and a controller transfer function R_C. But unless we say otherwise, we just take R_L as a transfer function in its own right, and do not concern ourselves with from where it comes.

8.5.1 Basic properties of the sensitivity function

smooth this

Let us first make some elementary observations concerning the closed-loop transfer function and the sensitivity function. The following result follows immediately from the definition of T_L and S_L.

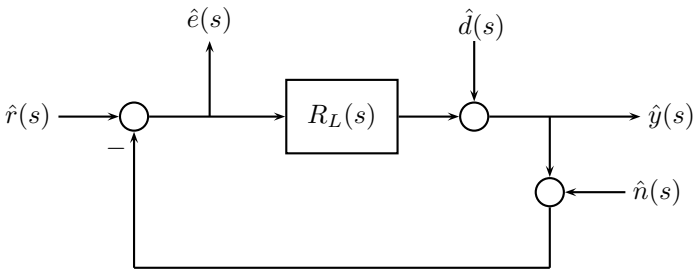


Figure 8.25 Block diagram for investigating the sensitivity function

8.22 Proposition *For the feedback interconnection of Figure 9.3, the following statements hold:*

(i) $p \in \mathbb{C}$ *is a pole for* R_L *if and only if* $S_L(p) = 0$ *and* $T_L(p) = 1$*;*

(ii) $z \in \mathbb{C}$ *is a zero for* R_L *if and only if* $S_L(z) = 1$ *and* $T_L(z) = 0$.

The import of this simple result is that it holds for *any* loop gain R_L. Thus the zeros and poles for R_L are immediately reflected in the closed-loop transfer function and the sensitivity function. Let us introduce the following notation:

$$\begin{aligned} Z(S_L) &= \{s \in \mathbb{C} \mid S_L(s) = 0\} \\ Z(T_L) &= \{s \in \mathbb{C} \mid T_L(s) = 0\}. \end{aligned} \tag{8.2}$$

This notation will be picked up again in Section 9.2.1.

Let us illustrate this with an example some of the simpler features of the sensitivity function.

8.23 Example **(Example 6.60 cont'd)** Let us carry on with our falling mass example. What we shall do is plot the frequency response of the closed-loop transfer function and the sensitivity function on the same Bode plot. We have $R_L(s) = R_C(s)\frac{1}{ms^2}$, and we shall use the four controller transfer functions of Example 6.60. In all cases, we take $m = 1$.

1. We take $R_C(s) = K$ with $K = 28$. In Figure 8.26 we give the Bode plots for the closed-loop transfer function (the solid line) and the sensitivity function (the dashed line).
2. Next we take $R_C(s) = KT_Ds$ with $K = 28$ and $T_D = \frac{9}{28}$. In Figure 8.26 we give the Bode plots for the closed-loop transfer function and the sensitivity function.
3. Now we consider pure integral control with $R_C(s) = \frac{1}{T_Is}$ with $T_I = \frac{7}{10}$. In Figure 8.26 we give the Bode plots for the transfer function (the solid line) and the sensitivity function (the dashed line).
4. Finally, we look at a PID controller so that $R_C(s) = K + KT_Ds + \frac{1}{T_Is}$, and we use the same numerical values as above. As expected, in Figure 8.27 you will find the Bode plots for the closed-loop transfer function and the sensitivity function.

In all cases, note that the gain attenuation at high frequencies leads to high sensitivities at these frequencies, and all the associated disadvantages. •

8.5.2 Quantitative performance measures

Our emphasis to this point has been essentially on *descriptive* measures of performance. However, it is very helpful to have on hand *quantitative* measures of performance. Indeed, such performance measures form the backbone of modern control theory, as from these

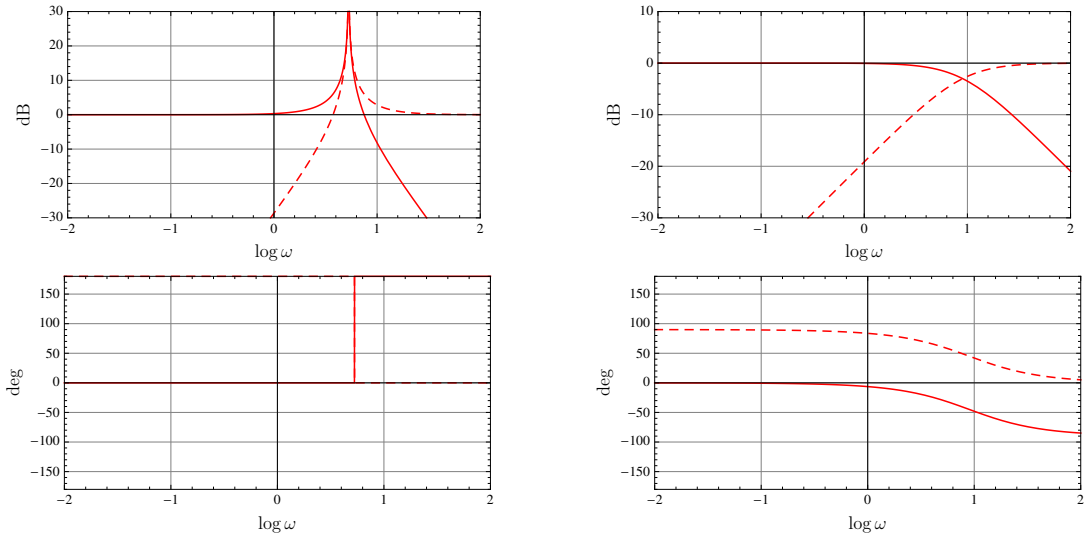


Figure 8.26 Bode plots for the closed-loop transfer function and sensitivity function for the falling mass with proportional controller (left) and derivative controller (right). In each case, the solid line represents the closed-loop transfer function and the dashed line the sensitivity function.

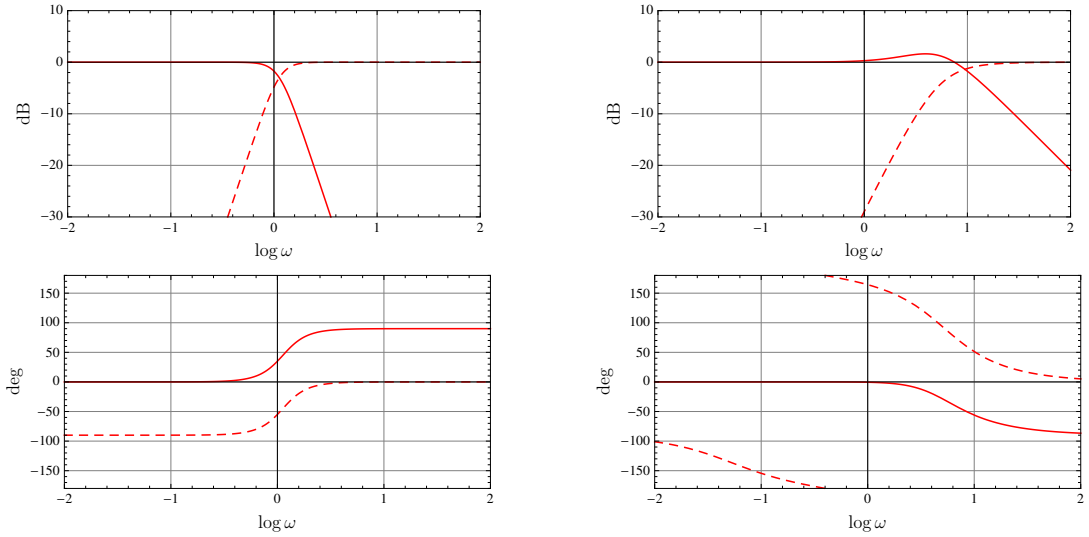


Figure 8.27 Bode plots for the closed-loop transfer function and sensitivity function for the falling mass with integral controller (left) and PID controller (right). In each case, the solid line represents the closed-loop transfer function and the dashed line the sensitivity function.

precise performance characteristics spring useful design methodologies. The reader will wish to recall from Section 5.3.1 the definitions of the L^2 and L^∞-norms.

The quantitative measures we will provide are those on the error of a system to a step input. Thus we work with the block diagram of Figure 6.25 which we reproduce in Figure 8.28. We shall assume that the closed-loop system is type 1 so that the steady-state error to a step input is zero. In order to make meaningful comparisons, we deal with the

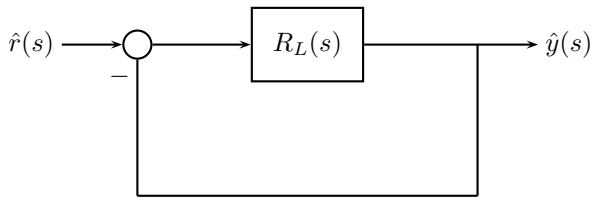


Figure 8.28 Block diagram for quantitative performance measures

unit step response to the reference $1(t)$.

For design methodologies with a mathematical basis, the following two measures of performance are often the most useful:

$$\|e\|_2 = \Big(\int_0^\infty e^2(t)\,\mathrm{d}t\Big)^{1/2}$$
$$\|e\|_\infty = \sup_{\alpha\geq 0}\{|e(t)| \leq \alpha \text{ for almost every } t\}.$$

The first is of course the L^2-norm of the error, and the second the L^∞-norm of the error. The error measure

$$\|e\|_1 = \int_0^\infty |e(t)|\,\mathrm{d}t$$

is also used. For example, you will recall from Section 13.1 that the criterion used by Ziegler and Nicols in their PID tuning was the minimisation of the L^1-norm. In practice, one will sometimes wish to consider alternative performance measures. For example, the following two, related to the L^2 and L^1-norms, will serve to penalise errors which occur for large times, but deemphasise large initial errors:

$$\|e\|_{t,2} = \Big(\int_0^\infty te^2(t)\,\mathrm{d}t\Big)^{1/2}$$
$$\|e\|_{t,1} = \int_0^\infty t|e(t)|\,\mathrm{d}t.$$

One can imagine choosing other weighting functions of time by which to multiply the error to suit the particular nature of the problem with which one is dealing.

One of the reasons for the usefulness of the L^2 and L^∞ norms is that they make it possible to formulate statements in terms of transfer functions. For example, we have the following result which was derived in the course of the proof of Theorem 5.22, if we recall that the transfer function from the input to the error in Figure 8.28 is the sensitivity function.

8.24 Proposition *Consider the block diagram of Figure 8.28 and suppose that the closed-loop system is type k for $k \geq 1$. If $e(t)$ is the error to an input $u \in \mathrm{L}^2[0,\infty)$, then*

$$\|e\|_2 \leq \|S_L\|_\infty \|u\|_2.$$

where S_L is the sensitivity function of the closed-loop system.

This proposition tells us that if a design objective is to minimise the transfer of energy from the input signal to the error signal, then one should aim to minimise the H^∞-norm of S_L.

8.6 Frequency-domain performance specifications

The time-domain performance specifications of Section 8.1 are traditionally what one encounters as these are the easiest to achieve a feeling for. However, there are powerful controller design methods which rely on specifying the performance requirements in the frequency-domain, not the time-domain. In this section we address this in two ways. First we look at some specifications that are actually most naturally given in the frequency-domain. These are followed by an explanation of how time-domain specifications can be approximately turned into frequency-domain specifications.

8.6.1 Natural frequency-domain specifications

There are a significant number of performance specifications that are very easily presented in the frequency-domain. In this section we list some of these. In a given design problem, one may wish to invoke only some of these constraints and not others. Also, the exact way in which the specifications are best presented can vary in detail, depending on the particular application. However, it can be hoped that what we say here is a helpful guide in getting started in a given problem.

Before we begin, we remark that all of the specifications we give will be are what Helton and Merrino [1998] call "disk constraints." This means that the closed-loop transfer function T_L will be specified to satisfy a constraint of the form

$$|M(i\omega) - T_L(i\omega)| \leq R(i\omega), \quad \omega \in [\omega_1, \omega_2]$$

R+ defined?

for functions $M\colon i[\omega_1, \omega_2] \to \mathbb{C}$ $M\colon i[\omega_1, \omega_2] \to \mathbb{R}_+$, and for $\omega_1 < \omega_2 > 0$. Thus at each frequency ω in the interval $[\omega_1, \omega_2]$, the value of T_L at $i\omega$ must lie in the disk of radius $R(i\omega)$ with centre $M(i\omega)$.

Stability margin lower bound

Recall from Proposition 7.15 that the point on the Nyquist contour closest to $-1 + i0$ is a distance $\|S_L\|_\infty$ away. Thus, to improve stability margins, one should provide a lower bound for this distance. This means specifying an upper bound on the H_∞-norm of the sensitivity function. In terms of the transfer function this gives

$$|1 - T_L(i\omega)| \leq \rho_{\mathrm{sm}}, \quad \omega > 0. \tag{8.3}$$

It is possible that one may wish to enforce this constraint more or less at various frequencies. For example, if one knows that a system will be operating in a certain frequency range, then stability margins in this frequency range will be more important that in others. Thus one may relax (8.3) by introducing a weighting function W and asking that

$$|W(i\omega)(1 - T_L(i\omega))| \leq \rho_{\mathrm{sm}}, \quad \omega > 0, \quad \omega > 0.$$

As a disk constraint this reads.

$$|1 - T_L(i\omega)| \leq \frac{\rho_{\mathrm{sm}}}{|W(i\omega)|}, \quad \omega > 0. \tag{8.4}$$

Tracking specifications

Recall that the tracking error is minimised as in Proposition 8.24 by minimising $\|S_L\|_\infty$. However, the kind of tracking error minimisation demanded by Proposition 8.24 is extremely

stringent. Indeed, it asks that for *any* sort of input, we minimise the L_2-norm of the error. However, in practice one will wish to minimise the tracking error for inputs having a certain frequency response. To this end, we may specify various sorts of specifications that are less restrictive than minimising $\|S_L\|_\infty$.

A first case we consider is a constraint

$$|1 - T_L(i\omega)| < R_{\text{tr}}, \quad \omega \in [\omega_1, \omega_2]. \tag{8.5}$$

This corresponds to tracking well signals whose frequency response is predominantly supported in the given range.

Another approach that may be taken occurs when one knows, or approximately knows, the transfer function for the reference one wishes to track. That is, suppose that we wish to track the reference $r(t)$ whose Laplace transform is $\hat{r}(s)$. Generalising the analysis of Section 8.3 in a straightforward way from type k reference signals to reference signals with a general Laplace transform, we see that to track r well we should require

$$|(1 - T_L(i\omega))\hat{r}(i\omega)| \leq \rho_{\text{tr}}, \quad \omega \in [\omega_1, \omega_2].$$

This is then made into the disk constraint

$$|1 - T_L(i\omega)| \leq \frac{\rho_{\text{tr}}}{|\hat{r}(i\omega)|}, \quad \omega \in [\omega_1, \omega_2]. \tag{8.6}$$

Bandwidth constraints

One of the more important features of a closed-loop transfer function is its bandwidth. As we saw in Section 8.2, larger bandwidth generally means quicker response. However, one wishes to limit the bandwidth since it is often destructive to a system's physical components to have the response to high-frequency signals not be adequately attenuated.

Let us take this opportunity to define bandwidth for fairly general systems. Motivated by our observations for first and second-order transfer functions, we make the following definition.

8.25 Definition Let (N, D) be a proper SISO linear system in input/output form and let

$$\|T_{N,D}\|_\infty = \sup_\omega \{|H_{N,D}(\omega)|\},$$

as usual. When $\lim_{\omega \to 0} |H_{N,D}(\omega)| < \infty$, we consider the following five cases:

(i) (N, D) *is strictly proper and steppable:* the ***bandwidth*** of (N, D) is defined by

$$\omega_{N,D} = \inf_{\bar{\omega}} \left\{ \bar{\omega} \;\middle|\; \tfrac{|H_{N,D}(\omega)|}{|H_{N,D}(0)|} \leq \tfrac{1}{\sqrt{2}} \text{ for all } \omega > \bar{\omega} \right\};$$

(ii) (N, D) *is not strictly proper and not steppable:* the ***bandwidth*** of (N, D) is defined by

$$\omega_{N,D} = \sup_{\bar{\omega}} \left\{ \bar{\omega} \;\middle|\; \tfrac{|H_{N,D}(\omega)|}{|H_{N,D}(\infty)|} \leq \tfrac{1}{\sqrt{2}} \text{ for all } \omega < \bar{\omega} \right\};$$

(iii) (N, D) *is strictly proper, not steppable, and* $\|T_{N,D}\|_\infty < \infty$*:* the ***lower cutoff frequency*** of (N, D) is defined by

$$\omega_{N,D}^{\text{lower}} = \sup_{\bar{\omega}} \left\{ \bar{\omega} \;\middle|\; \tfrac{|H_{N,D}(\omega)|}{\|T_{N,D}\|_\infty} \leq \tfrac{1}{\sqrt{2}} \text{ for all } \omega < \bar{\omega} \right\}$$

and the ***upper cutoff frequency*** of (N, D) is defined by

$$\omega_{N,D}^{\text{upper}} = \inf_{\bar{\omega}} \left\{ \bar{\omega} \;\middle|\; \tfrac{|H_{N,D}(\omega)|}{\|T_{N,D}\|_\infty} \leq \tfrac{1}{\sqrt{2}} \text{ for all } \omega > \bar{\omega} \right\},$$

and the ***bandwidth*** is given by $\omega_{N,D} = \omega_{N,D}^{\text{upper}} - \omega_{N,D}^{\text{lower}}$;

(iv) (N, D) *is strictly proper, not steppable, and* $\|T_{N,D}\|_\infty = \infty$*:* the bandwidth of (N, D) is not defined;

(v) (N, D) *is not strictly proper and steppable:* the bandwidth of (N, D) is not defined.

When $\lim_{\omega \to 0} |H_{N,D}(\omega)| = \infty$, bandwidth is undefined. •

Thus $\omega_{N,D}$ is, for "typical" systems, simply the frequency at which the magnitude part of the Bode plot drops below, and stays below, the DC value minus -3dB. For most other transfer functions, definition of bandwidth is still possible, but must be modified to suit the characteristics of the system. Note that the bandwidth is not defined for systems with imaginary axis poles. However, this is not an issue since such systems are not BIBO stable, and we are typically interested in bandwidth for the closed-loop transfer function, where BIBO stability is essential.

bandwidth and gain crossover frequency

With is general notion of bandwidth, the typical bandwidth constraint one may pose might take the form

$$|T_L(i\omega)| < \rho_{\text{bw}}, \quad \omega > \omega_{\text{bw}}. \tag{8.7}$$

High-frequency roll-off

In practise one does not want the closed-loop transfer function to be proper, but strictly proper. This will ensure that the system will not be overly susceptible to high-frequency disturbances in the reference. Furthermore, one will often want the closed-loop transfer function to not only be proper, but to have a certain relative degree, so that it falls off at a prescribed rate as $\omega \to \infty$. Note that

$$T_L(i\omega) = \frac{R_C(i\omega)R_P(i\omega)}{1 + R_C(i\omega)R_P(i\omega)}.$$

Therefore, assuming that $R_C R_P$ is proper, for large frequencies $T_L(i\omega)$ behaves like $R_C(i\omega)R_P(i\omega)$. To render the desired behaviour as a disk inequality, we first assume that we insist on controllers that satisfy an inequality of the form

$$|R_C(i\omega)| \leq \frac{\rho_C}{|\omega|^{n_C}}$$

for some $\rho_C > 0$ and $n_C \in \mathbb{N}$. This is tantamount to making a relative degree constraint on the controller. Since we are primarily concerned with high-frequency behaviour, the choice of ρ_C is not critical. Assuming that such a constraint has been enforced, the high-frequency approximation

$$T_L(i\omega) \approx R_C(i\omega)R_P(i\omega)$$

gives rise to the inequality

$$|T_L(i\omega)| \leq \frac{\rho_C |R_P(i\omega)|}{|\omega|^{n_C}}, \quad [\omega_1, \omega_2]. \tag{8.8}$$

Controller output constraints

The transfer function

$$R_C S_L = \frac{R_C}{1 + R_C R_P}$$

is sometimes called the ***closed-loop controller***. It is the transfer function from the error to the output from the plant. One would wish to minimise this transfer function in order to ensure that the controller is not excessively aggressive. In practise, an excessively aggressive controller might cause saturation, i.e., the controller may not physically be able to supply the output needed. Saturation is a nonlinear effect, and should be avoided, unless its effects are explicitly accounted for in the modelling.

In any event, the constraint we consider to limit the output of the closed-loop controller is

$$|R_C(i\omega)(1 - T_L(i\omega))| \leq \rho_{\text{clc}},$$

where we leave the frequency unspecified for the moment. Using the relation

$$T_L = \frac{R_C R_P}{1 + R_C R_P} \quad \Longrightarrow \quad R_C R_P = \frac{T_L}{1 - T_L},$$

this gives the disk constraint

$$|T_L(i\omega)| \leq \rho_{\text{clc}} |R_P(i\omega)|.$$

Now let us turn our attention to the reasonable ranges of frequency for the invocation of such a constraint. Clearly the constraint will have the greatest effect for frequencies that are zeros for R_P that are on the imaginary axis. This we let $iz_1, \ldots, iz_k$ be the imaginary axis zeros for R_P, and take as our disk constraint

$$|T_L(i\omega)| \leq \rho_{\text{clc}} |R_P(i\omega)|, \quad \omega \in [z_1 - r_1, z_1 + r_1] \cup \cdots \cup [z_k - r_k, z_k + r_k], \tag{8.9}$$

for some $r_1, \ldots, r_k > 0$.

Plant output constraints

The idea here is quite similar to that for the closed-loop controller. To wit, the name ***closed-loop plant*** is often given to the transfer function

$$R_P S_L = \frac{R_P}{1 + R_C R_P}.$$

This is the transfer function from the output of the controller to the output of the plant. That this should be kept from being too large is a reflection of the desire to not have the plant "overreact" to the controller.

The details of the computation of the ensuing disk constraint go very much like that for the closed-loop controller. We begin by considering the constraint

$$|R_P(i\omega)(1 - T_L(i\omega))| \leq \rho_{\text{clp}},$$

immediately giving rise to the disk constraint

$$|1 - T_L(i\omega)| \leq \frac{\rho_{\text{clp}}}{|R_P(i\omega)|}.$$

To determine when one should invoke this constraint, note that it most restrictive for frequencies near imaginary axis poles for R_P. Thus we let $p_1, \ldots, p_k$ be the imaginary axis poles for R_P and take the constraint

$$|1 - T_L(i\omega)| \leq \frac{\rho_{\text{clp}}}{|R_P(i\omega)|}, \quad \omega \in [p_1 - r_1, p_1 + r_1] \cup \cdots \cup [p_k - r_k, p_k + r_k], \tag{8.10}$$

for given $r_1, \ldots, r_k > 0$.

Overshoot attenuation

In Section 8.2.2 we saw that there is a relationship, at least for second-order transfer functions, between a peak in the frequency response for the closed-loop transfer function and a large overshoot. While no theorem to this effect is known to the author, it may be desirable to enforce a constraint on $\|T_L\|_\infty$ of the type

$$|T_L(i\omega)| \leq \rho_{\text{os}}, \quad \omega \geq 0. \tag{8.11}$$

Summary

We have given a list of "typical" frequency-domain constraints. In practise one rarely enforce all of these simultaneously. Nevertheless, in a given application, any of the constraints (8.3), (8.4), (8.5), (8.6), (8.8), (8.7), (8.9), (8.10), or (8.11) may prove useful. Note that specifying the constants and various weighting functions will be a process of trial and error in order to ensure that one specifies a problem that is solvable, and still has satisfactory performance. In Chapter 9 we discuss in detail the idea that not all types of frequency-domain performance specification should be expected to be achievable, particularly for unstable and/or nonminimum phase plants. The methods for obtaining controllers satisfying the type of frequency-domain constraints we have discussed in this section may be found in the book [Helton and Merrino 1998], where a software is demonstrated for doing such design. This is given further discussion in Chapter 15 in terms of "H_∞ methods."

8.6.2 Turning time-domain specifications into frequency-domain specifications

It is generally not possible to make a given time-domain performance specification and produce an exact, tractable frequency-domain specification. Nonetheless, one can make some progress in producing reasonably effective and manageable frequency-domain specifications from many types of time-domain specifications. The idea is to produce a transfer function having the desired time-domain behaviour, then using this as the basis for forming the frequency-domain specifications.

finish

8.7 Summary

When designing a controller the first step is typically to determine acceptable performance criterion. In this chapter, we have come up with a variety of performance measures. Let us review what we have said.

1. Based upon a system's step response, we defined various performance features (rise time, overshoot, etc.). These features should be understood at least in that one should be able to compare two signals and determine which is the better with respect to certain of these performance features.

2. Some of the character of system response are exhibited by a simple second-order transfer function. In particular, the tradeoffs one has to make in controller design begin to show up for such systems in that one cannot perfectly satisfy all performance measures.
3. For simple systems, often one can obtain an adequate understanding of the problems to be encountered in system performance by using the observations seen when additional poles and zeros are added to a second-order transfer function.
4. The effects of the existence of unstable poles and nonminimum phase zeros on the step response should be understood.
5. System type is an easily understood concept. Particularly, one should be able to readily determine the system type of a unity gain feedback loop with ease.
6. For disturbance rejection, one should understand that a disturbance may affect a system's dynamics in a variety of ways, and that the way these are quantified are via appended systems, and systems types for these appended systems.
7. For design, the sensitivity function is an important consideration as concerns designing controllers which are in some sense robust. The reader should be aware of some of the tradeoffs which are necessitated by the need to have a good closed-loop response along with desirable robustness and disturbance rejection characteristics.

Exercises

E8.1 Consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ with

$$\boldsymbol{A} = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1,$$

where $\sigma \leq 0$ and $\omega > 0$. For $\omega = 1$ and $\sigma \in \{0, -0.1, -0.7, -1\}$, do the following.

(a) Determine the steady-state value of the output for a unit step input.
(b) For a unit step input and zero initial conditions, plot the output response (you do not have to provide an analytical expression for it).
(c) From your plot, determine the rise time, t_r.
(d) From your plot, determine the ϵ-settling time, $t_{s,\epsilon}$, for $\epsilon = 0.1$.
(e) Determine the percentage overshoot P_{os}.
(f) Plot the location of the poles of the transfer function T_Σ in the complex plane.
(g) Produce the magnitude Bode plot for $H_\Sigma(\omega)$, and from it determine the bandwidth of the system.

E8.2 Denote by Σ_1 the SISO linear system of Exercise E8.1. Add in series with the system Σ_1 the first-order SISO system $\Sigma_2 = (\boldsymbol{A}_2, \boldsymbol{b}_2, \boldsymbol{c}_2^t, \boldsymbol{D}_2)$ with

$$\boldsymbol{A}_2 = \begin{bmatrix} -\alpha \end{bmatrix}, \quad \boldsymbol{b}_2 = \begin{bmatrix} 1 \end{bmatrix}, \quad \boldsymbol{c}_2 = \begin{bmatrix} 1 \end{bmatrix}, \quad \boldsymbol{D}_2 = \boldsymbol{0}_1.$$

That is, consider a SISO linear system having as its input the input to Σ_2 and as its output the output from Σ_1 (see Exercise E2.1).

(a) Determine the transfer function for the interconnected system, and from this ascertain the steady-state output arising from a unit step input.

Fix $\sigma = -1$ and $\omega = 1$, and for $\alpha \in \{0, 0.1, 1, 5\}$, do the following.

(b) For a unit step input and zero initial conditions, plot the output response (you do not have to provide an analytical expression for it).
(c) From your plot, determine the rise time, t_r.
(d) From your plot, determine the ϵ-settling time, $t_{s,\epsilon}$, for $\epsilon = 0.1$.
(e) From your plot, determine the percentage overshoot P_{os}.
(f) Plot the location of the poles of the transfer function in the complex plane.
(g) Produce the magnitude Bode plot for the interconnected system, and from it determine the bandwidth of the system.

E8.3 Denote by Σ_1 the SISO linear system of Exercise E8.1. Interconnect Σ_1 with blocks containing s and $\alpha \in \mathbb{R}$ as shown in the block diagram Figure E8.1. Thus the input gets fed into the parallel block, whose output becomes the input into Σ_1.

(a) Determine the transfer function for the interconnected system, and from this ascertain the steady-state output arising from a unit step input.

Fix $\sigma = -1$ and $\omega = 1$, and for $\alpha \in \{-1, 0, 1, 5\}$, do the following.

(b) For a unit step input and zero initial conditions, plot the output response (you do not have to provide an analytical expression for it).
(c) From your plot, determine the rise time, t_r.
(d) From your plot, determine the ϵ-settling time, $t_{s,\epsilon}$, for $\epsilon = 0.1$.
(e) From your plot, determine the percentage overshoot P_{os}.

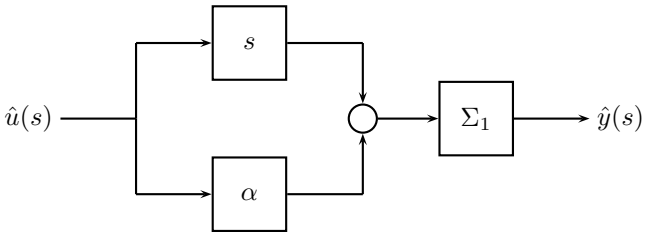


Figure E8.1 Block diagram for system interconnection

(f) Plot the location of the poles and zeros of the transfer function in the complex plane.

(g) Produce the magnitude Bode plot for the interconnected system, and from it determine the bandwidth of the system.

(h) Produce the phase Bode plot for the interconnected system and make some comments on what you observe.

E8.4 Consider the closed-loop interconnection in Figure E8.2, and assume that it is IBIBO

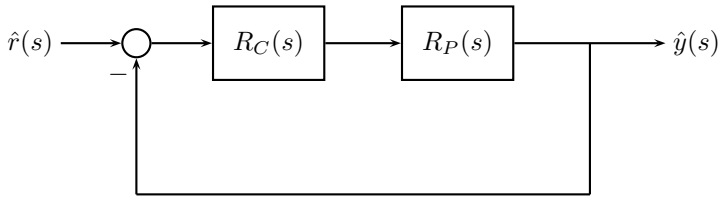


Figure E8.2 A closed-loop system

stable. Suppose that the loop gain $R_C R_P$ satisfies $\lim_{s\to 0} R_P(s)R_C(s) = K$ where $K \neq -1$.

(a) Show that the closed-loop transfer function is type 0. If one desires zero steady-state error to a step input, will the closed-loop system be satisfactory?

(b) Design a rational function $\tilde{R}_C$ which has the property that, when put into the block diagram of Figure E8.3, the resulting closed-loop transfer function will be

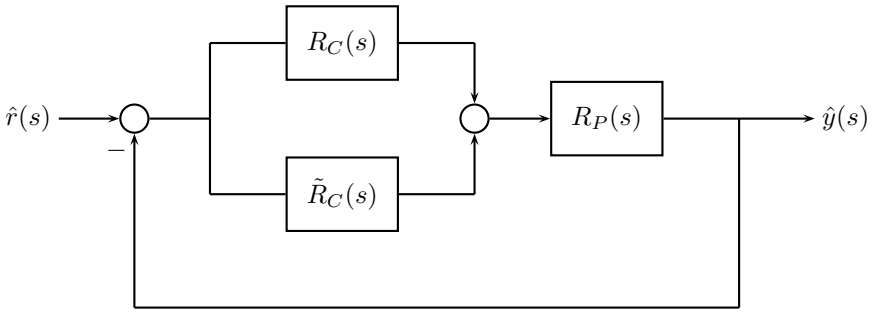


Figure E8.3 A modified closed-loop system

of type 1. For the closed-loop system you have just designed, what will be the error of the system to a step input?

Now suppose that you wish to consider the effect of a disturbance which enters the system as in Figure E8.4.

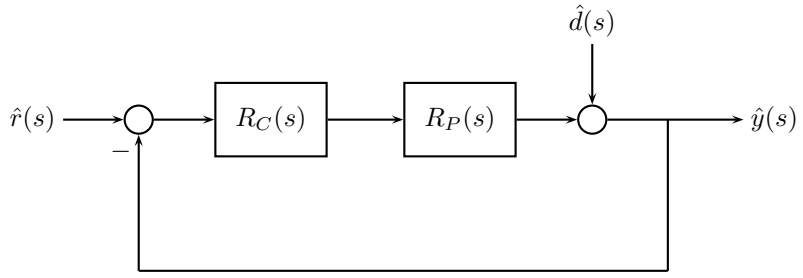


Figure E8.4 Adding a disturbance

(c) What is the system type with respect to this disturbance for the controller depicted in the block diagram Figure E8.2?

(d) What is the system type with respect to this disturbance for the modified controller depicted in the block diagram Figure E8.3 (i.e., that obtained by replacing the R_C block with the parallel blocks containing your $\tilde{R}_C$ and R_C)?

(e) Comment on the effectiveness of the modified controller as concerns rejection of step disturbances in this case.

E8.5 Let $R_P \in \mathbb{R}(s)$ be a strictly proper plant with $R_C \in \mathscr{S}(R_P)$ a proper controller.

(a) Show that if the Nyquist plot for $R_L = R_C R_P$ lies in $\overline{\mathbb{C}}_+$ then $\|S_L\|_\infty < 1$. What can you say about the performance of the resulting closed-loop system?

Now consider the two plants

$$R_{P,1} = \frac{1}{s+1}, \quad R_{P,2} = \frac{s-1}{s^2+2s+1}.$$

(b) Produce the Bode plots for both plants with the controller $R_C(s) = \frac{1}{2} \in \mathscr{S}(R_{P,1}) \cap \mathscr{S}(R_{P,2})$. How do they differ?

(c) Produce the Nyquist plots for both loop gains $R_{L,1} = R_C R_{P,1}$ and $R_{L,2} = R_C R_{P,2}$.

(d) Comment on the comparative performance of the closed-loop systems in light of your work done in part (a).

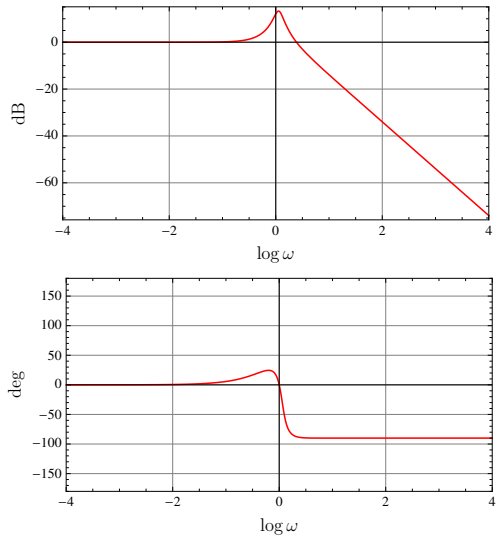
(e) Produce the step response for both closed-loop systems and comment on their comparative behaviour.

E8.6 In problems (a)–(d) there are two SISO linear systems (N_1, D_1) and (N_2, D_2) in input/output form, but you are not told what they are. Instead, you are given a pair of Bode plots, and a pair of step responses, one each for the pair of transfer functions T_{N_1,D_1} and T_{N_2,D_2}. You are not told which Bode plot and which step response come from the same transfer function.

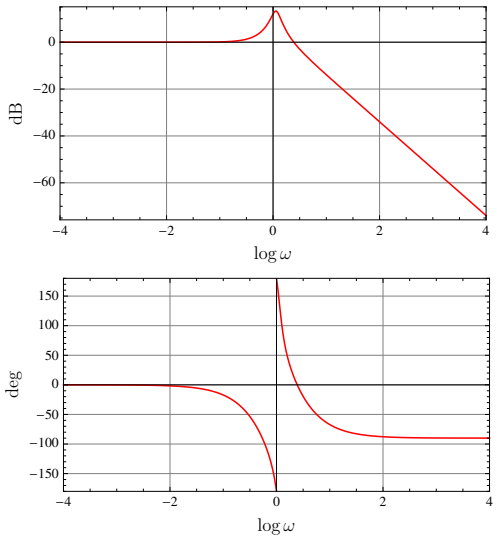
In each case, do the following:

1. Indicate which Bode plot goes with which step response. correctly.
2. Indicate clearly (but not necessarily at great length) the features of the plots that justify your choice in step 1.

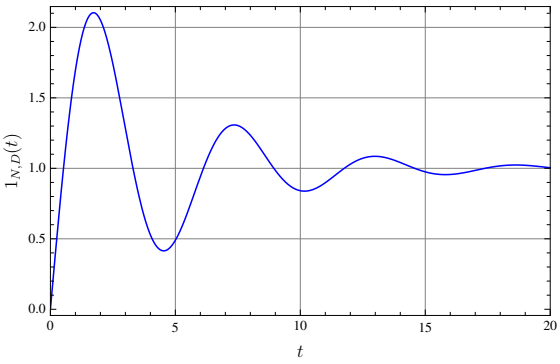
(a)



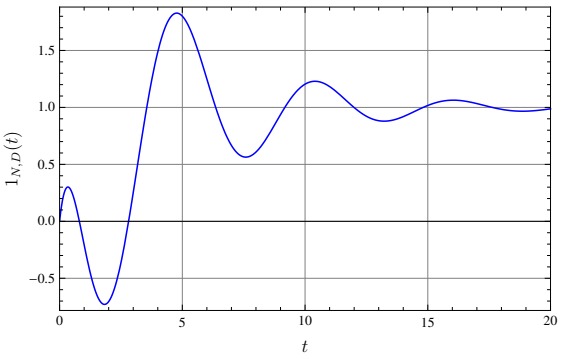
Bode plot 1



Bode plot 2

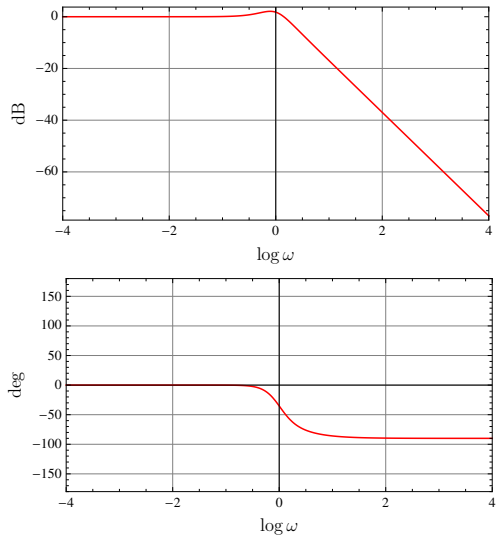


Step response 1

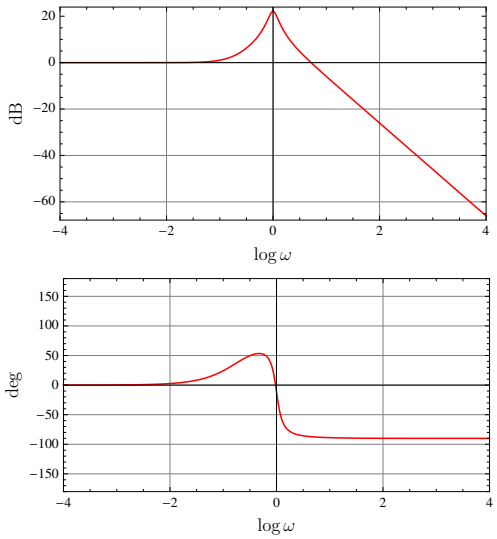


Step response 2

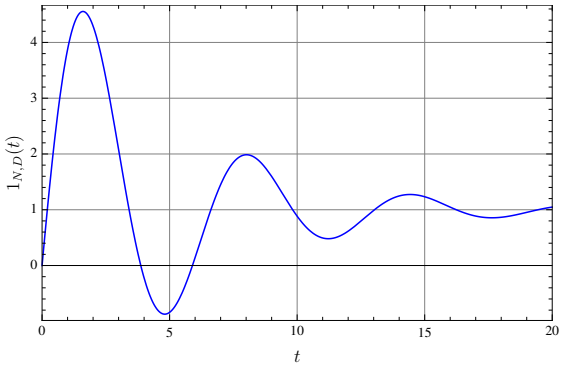
(b)



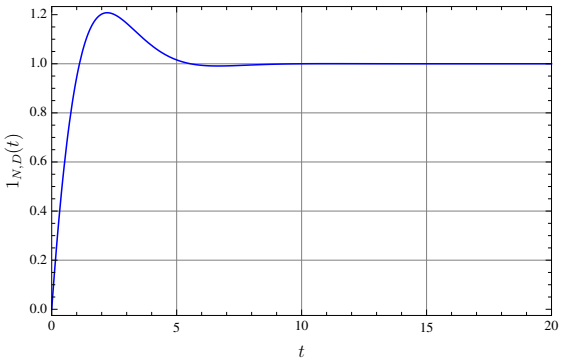
Bode plot 1



Bode plot 2

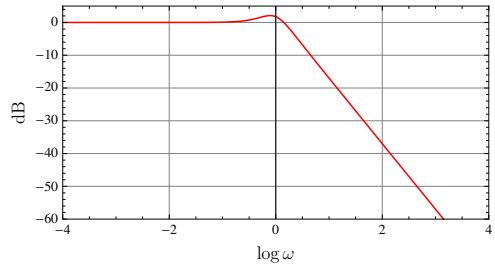


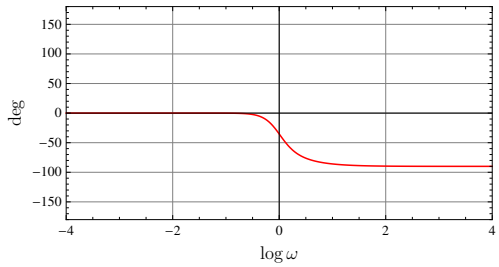
Step response 1



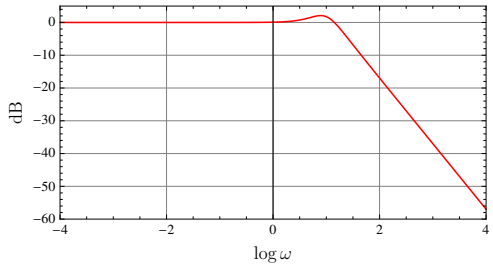
Step response 2

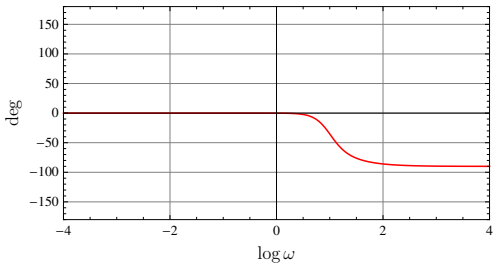
(c)



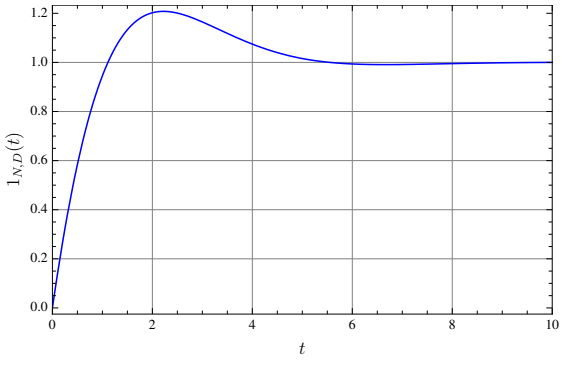


Bode plot 1

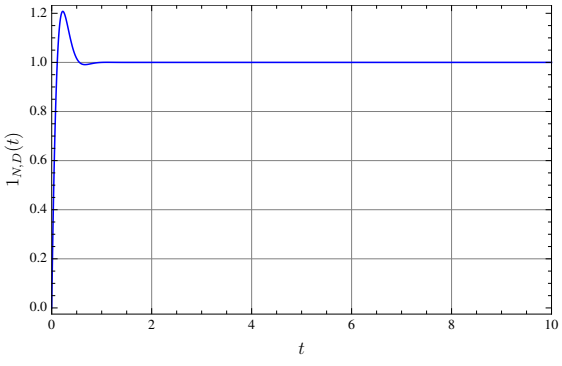




Bode plot 2

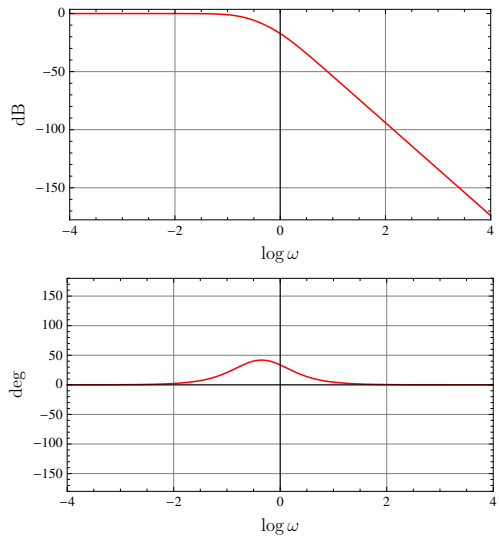


Step response 1

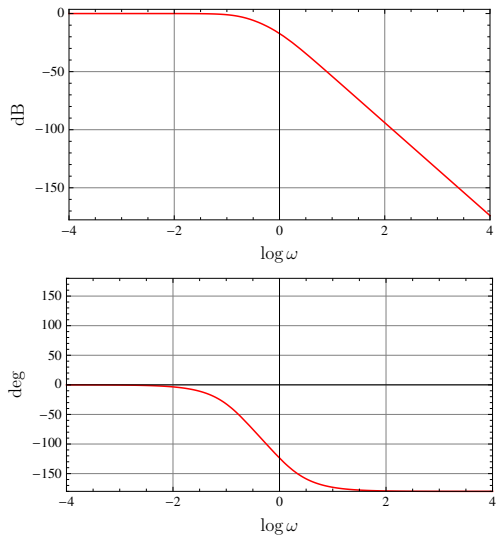


Step response 2

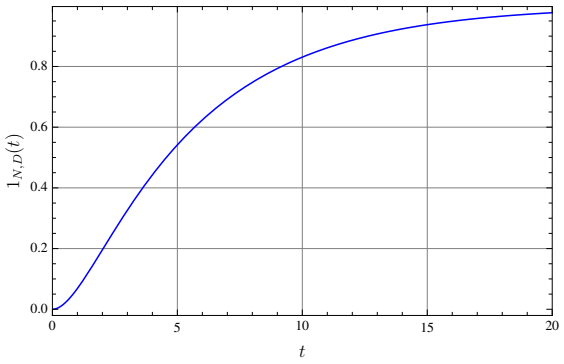
(d)



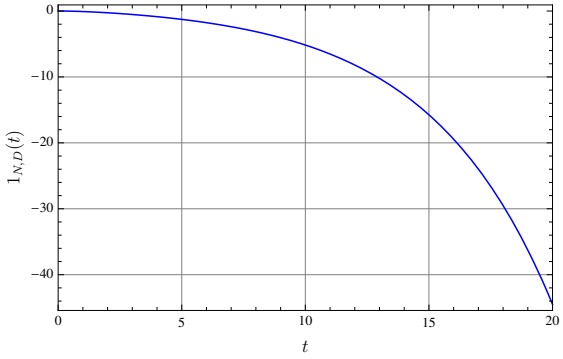
Bode plot 1



Bode plot 2



Step response 1



Step response 2

The material in this chapter has focused upon the unity gain feedback loop and its relation to the solution of Problem 6.41 concerning design for input/output systems. In the next two problems, you will investigate a few aspects of performance for Problem 6.48 where static state feedback is considered, and for Problem 6.53 where static output feedback is considered. Recall that the block diagram representation for static state feedback is as in Figure 8.5, and that the block diagram representation for static output feedback is as in

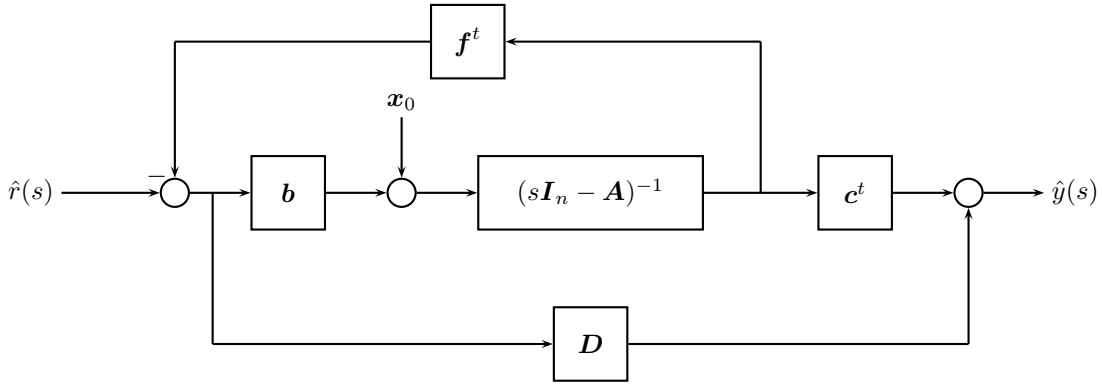


Figure 8.5 Block diagram for static state feedback

Figure 8.6. You may also wish to refer to Theorems 6.49 and 6.54 concerning the form of

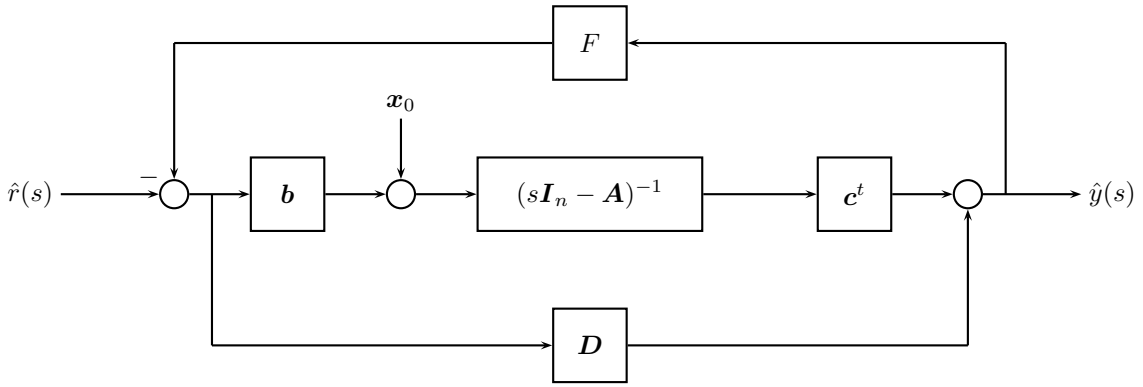


Figure 8.6 Block diagram for static output feedback

the transfer function under static state feedback and static output feedback.

E8.7 In this exercise we consider a controllable SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ and a state feedback vector $\boldsymbol{f}$. You may suppose that $(\boldsymbol{A}, \boldsymbol{b})$ is in controller canonical form, and that $\boldsymbol{f} \in \mathscr{S}_{\mathrm{s}}(\Sigma)$.

(a) Compute the transfer function from the reference $\hat{r}(s)$ to the error $\hat{r}(s) - \hat{y}(s)$ for the closed-loop system of Figure 8.5.

(b) Determine the system type for the closed-loop system of Figure 8.5. Note that the system type will depend on the relationship between the state feedback vector $\boldsymbol{f}$ and the system Σ.

Now we will consider a concrete example of the above situation by taking

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \boldsymbol{f} = \begin{bmatrix} f_0 \\ f_1 \end{bmatrix}.$$

(c) What are the possible values for the system type for the closed-loop system in this case?

(d) For what values of f_0 and f_1 does the system type achieve its maximum possible value?

(e) Let $\boldsymbol{f}$ be a state feedback vector from part (d), i.e., so that the system type of the closed-loop system is maximal. Plot the step response of the closed-loop system. What is the steady-state error?

(f) Let $\boldsymbol{f}$ be a state feedback vector that is not of the type which answers part (d), i.e., so that the system type of the closed-loop system is not maximal. Plot the step response of the closed-loop system. What is the steady-state error?

E8.8 In this exercise we consider a controllable SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ and an output feedback constant F. You may suppose that $(\boldsymbol{A}, \boldsymbol{b})$ is in controller canonical form.

(a) Compute the transfer function from the reference $\hat{r}(s)$ to the error $\hat{r}(s) - \hat{y}(s)$ for the closed-loop system of Figure 8.6.

(b) Determine the system type for the closed-loop system of Figure 8.6. Note that the system type will depend on the relationship between the output feedback constant F and the system Σ.

Now we will consider a concrete example of the above situation by taking

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

(c) What are the possible values for the system type for the closed-loop system in this case?

(d) For what values of F does the system type achieve its maximum possible value?

(e) Let F be an output feedback constant from part (d), i.e., so that the system type of the closed-loop system is maximal. Plot the step response of the closed-loop system. What is the steady-state error?

(f) Let F be a output feedback constant that is not of the type which answers part (d), i.e., so that the system type of the closed-loop system is not maximal. Plot the step response of the closed-loop system. What is the steady-state error?

Chapter 9

Design limitations for feedback control

In this chapter we shall follow up on some of the discussion of Chapter 8 with details concerning performance limitations in both the frequency and time-domains. The objective of this chapter is to clarify the way in which certain plant features can impinge upon the attainability of certain performance specifications. That such matters can arise should be clear from parts of our discussion in the preceding chapter. There we saw that even for simple second-order systems there is a tradeoff to be made when simultaneously optimising, for example, rise time and overshoot. In this chapter such matters will be brought into clearer focus, and presented in a general context.

Many of the ideas we discuss here have been known for some time, but a very nice current summary of results of the type we present may be found in the book of Seron, Braslavsky, and Goodwin [1997]. The starting point for the discussion in this chapter might be thought of as Bode's Gain/Phase Theorem stated in Section 4.4.2.

Contents

9.1 Performance restrictions in the time-domain for general systems

The objective when studying feedback control systems is to see how behaviour of certain transfer functions affects the response. To this end, we concentrate on a certain simple feedback configuration, namely the unity gain feedback loop of Figure 6.25 that we again reproduce, this time in Figure 9.1. While in Section 8.3 we investigated the transfer function T_L of the closed-loop system of Figure 9.1, in this section we focus on how the character of the loop gain R_L itself affects the system performance. Note that it is possible to have R_L *not* be BIBO stable, but for the closed-loop system to be IBIBO stable. Thus, if $R_L = R_C R_P$ for a controller transfer function R_C and a plant transfer function R_P, it is possible to stabilise an unstable plant using feedback. However, we shall see in this section that unstable plants, along with nonminimum phase plants, can impose on the system certain performance

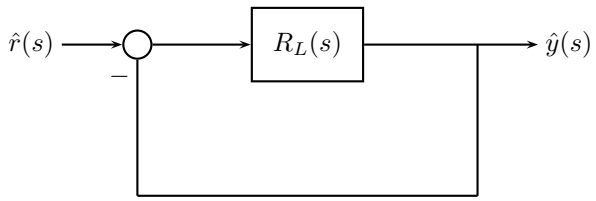


Figure 9.1 Unity gain feedback loop for studying time-domain behaviour

limitations. Our treatment follows that of Seron, Braslavsky, and Goodwin [1997].

The integral constraints of the following result form the backbone for a part of what will follow. The crucial thing to note here is that for *any* plant that has poles or zeros in $\mathbb{C}_+$, there will be performance restrictions, regardless of what kind of controller one employs.

9.1 Theorem *Suppose the unity gain feedback loop of Figure 9.1 is IBIBO stable and that the closed-loop transfer function is steppable. Let $y(t)$ be the normalised step response and let $e(t) = r(t) - y(t)$ be the error. Let*

$$\alpha = \inf_{s \in \mathbb{C}} \{ \mathrm{Re}(s) > \mathrm{Re}(p) \text{ when } p \text{ is a pole of } T_L \}$$

be the largest value of the real parts of the pole of T_L. The following statements hold.

(i) If $p \in \mathbb{C}_+$ is a pole for R_L then

$$\int_0^\infty e^{-pt} e(t)\,\mathrm{d}t = 0 \quad \text{and} \quad \int_0^\infty e^{-pt} y(t)\,\mathrm{d}t = \frac{1}{T_L(0)p}.$$

(ii) If $z \in \mathbb{C}_+$ is a zero for R_L then

$$\int_0^\infty e^{-zt} e(t)\,\mathrm{d}t = \frac{1}{T_L(0)z} \quad \text{and} \quad \int_0^\infty e^{-zt} y(t)\,\mathrm{d}t = 0.$$

(iii) If R_L has a pole at 0 then the following statements hold:

(a) if $p \in \overline{\mathbb{C}}_-$ is a pole for R_L with $\mathrm{Re}(p) > \alpha$ then

$$\int_0^\infty e^{-pt} e(t)\,\mathrm{d}t = 0;$$

(b) if $z \in \overline{\mathbb{C}}_-$ is a zero for R_L with $\mathrm{Re}(z) > \alpha$ then

$$\int_0^\infty e^{-zt} e(t)\,\mathrm{d}t = \frac{1}{T_L(0)z}.$$

Proof Let us first determine that all stated integrals exist. Since the closed-loop system is IBIBO stable, the normalised step response $y(t)$ will be bounded, as will be the error. Therefore, if $\mathrm{Re}(s) > 0$, the integrals

$$\int_0^\infty e^{-st} e(t)\,\mathrm{d}t \quad \text{and} \quad \int_0^\infty e^{-st} y(t)\,\mathrm{d}t$$

will exist. Now we claim that if $\lim_{s\to 0} R_L(s) = \infty$ then $\lim_{t\to 0} e(t) = 0$. Indeed, by the Final Value Theorem, Proposition E.9(ii), we have

$$\begin{aligned}\lim_{t\to\infty} e(t) &= \lim_{s\to 0} sS_L(s)\hat{r}(s)\\ &= \lim_{s\to 0} \frac{1}{1+R_L(s)}\frac{1}{T_L(0)}\\ &= 0,\end{aligned}$$

where we have noted that for the normalised step response the reference is $1(t)\frac{1}{T_L(0)}$. Now, since $\lim_{t\to\infty} e(t) = 0$ when $\lim_{s\to 0} R_L(s) = \infty$, in this case the integral

$$\int_0^\infty e^{-st}e(t)\,\mathrm{d}t$$

will exist provided s is greater than the abscissa of absolute convergence for $e(t)$. Since the abscissa of absolute convergence is α, it follows that the integrals of part (iii) exist.

(i) We have

$$\int_0^\infty e^{-pt}e(t)\,\mathrm{d}t = \hat{e}(p) \quad \text{and} \quad \int_0^\infty e^{-pt}y(t)\,\mathrm{d}t = \hat{y}(p).$$

Therefore

$$\begin{aligned}\int_0^\infty e^{-pt}e(t)\,\mathrm{d}t &= S_L(p)\hat{r}(p)\\ &= \lim_{s\to p} \frac{1}{1+R_L(s)}\frac{1}{T_L(0)s}\\ &= 0,\end{aligned}$$

if $\mathrm{Re}(p) > 0$. For the integral involving $y(t)$, when $\mathrm{Re}(p) > 0$ we have

$$\begin{aligned}\int_0^\infty e^{-pt}y(t)\,\mathrm{d}t &= T_L(p)\hat{r}(p)\\ &= \lim_{s\to p} \frac{R_L(s)}{1+R_L(s)}\frac{1}{T_L(0)s}\\ &= \frac{1}{T_L(0)p}.\end{aligned}$$

(ii) We compute

$$\begin{aligned}\int_0^\infty e^{-zt}e(t)\,\mathrm{d}t &= S_L(z)\hat{r}(z)\\ &= \lim_{s\to z} \frac{1}{1+R_L(s)}\frac{1}{T_L(0)s}\\ &= \frac{1}{T_L(0)z},\end{aligned}$$

if $\mathrm{Re}(z) > 0$. In like manner we have

$$\begin{aligned}\int_0^\infty e^{-zt}y(t) &= T_L(z)\hat{r}(z)\\ &= \lim_{s\to z} \frac{R_L(s)}{1+R_L(s)}\frac{1}{T_L(0)s}\\ &= 0,\end{aligned}$$

again provided $\mathrm{Re}(z) > 0$.

(iii) In this case, we can conclude, since the integrals have been shown to converge, that the analysis of parts (i) and (ii) still holds. ■

It is not immediately obvious why these conclusions are useful or interesting. We shall see shortly that they do lead to some results that are more obviously useful and interesting, but let's make a few comments before we move on to further discussion.

9.2 Remarks

1. The primary importance of the result is that it will apply to *any* case where a plant has unstable poles or nonminimum phase zeros. This immediately asserts that in such cases there will be some restrictions on how the step response can behave.
2. Parts (i) and (ii) can be thought of as placing limits on how good the closed-loop response can be in the presence of unstable poles or nonminimum phase zeros for the loop gain.
3. Part (iii a) ensures that even if the plant has no poles in $\mathbb{C}_+$, provided that it has a pole at $s = 0$ along with any other pole to the right of the largest closed-loop pole, there will be overshoot.
4. Along similar lines, from part (iii b), if R_L has a pole at the origin along with a zero, even a minimum phase zero, that lies to the right of the largest pole of the closed-loop system, then there will be undershoot.
5. We saw in Section 8.3 the ramifications of the assumption that $\lim_{s\to 0} R_L(s) = \infty$. What this means is that there is an integrator in the loop gain, and integrators give certain properties with respect to rejection of disturbances, and the ability to track certain reference signals. •

Let us give a few examples that illustrate the remarks 3 and 4 above.

9.3 Examples

1. Let us illustrate part (iii a) of Theorem 9.1 with the loop gain $R_L(s) = \frac{2(s+1)}{s(s-1)}$. The closed-loop transfer function is

$$T_L(s) = \frac{2(s+1)}{s^2 + s + 2},$$

which is BIBO stable. The normalised step response is shown in Figure 9.2. The thing to note, of course, is that the step response exhibits overshoot.

2. Let us illustrate part (iii b) of Theorem 9.1 with the loop gain $R_L(s) = \frac{1-s}{s(s+2)}$. The closed-loop transfer function is

$$T_L(s) = \frac{1-s}{s^2 + s + 1},$$

which is BIBO stable. The normalised step response is shown in Figure 9.2, and as expected there is undershoot in the response. •

As we saw in the previous remarks 3 and 4, an unstable pole in the loop gain immediately implies overshoot in the step response and a nonminimum phase zero implies undershoot. It turns out that we can be even more explicit about what is happening, and the following result spells this out.

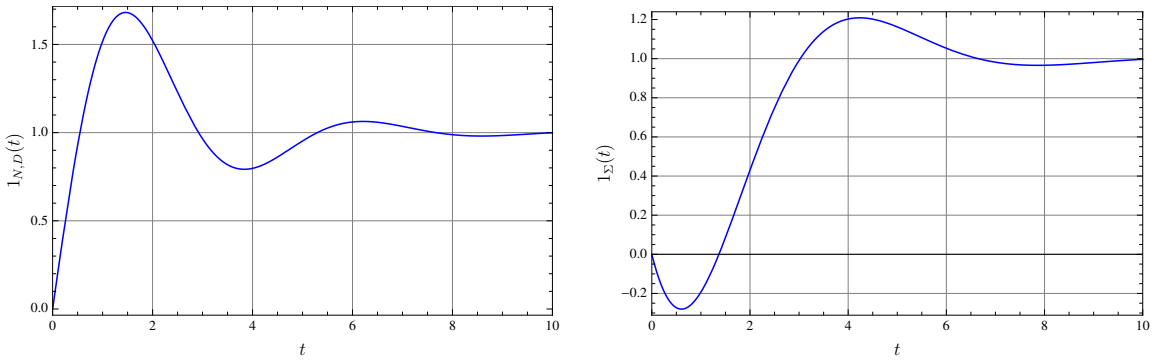


Figure 9.2 Normalised step responses for two loop gains with a pole at zero: (1) on the right an unstable pole gives overshoot and (2) on the left a nonminimum phase zero gives undershoot

9.4 Proposition *Consider the closed-loop system in Figure 9.1. Assuming the closed-loop system is steppable and IBIBO stable, the following statements hold.*

(i) If R_L has a real pole at $p \in \mathbb{C}_+$ then there is an overshoot in the normalised step response, and if t_r is the rise time, then the maximum overshoot satisfies

$$y_{\text{os}} \geq \frac{(pt_r - T_L(0))e^{pt_r} + T_L(0)}{pt_r}.$$

If $T_L(0) = 1$ then we can simplify this to

$$y_{\text{os}} \geq \frac{pt_r}{2}.$$

(ii) If R_L has a real zero at $z \in \mathbb{C}_+$ then there is an undershoot in the normalised step response, and if $t_{s,\epsilon}$ is the ϵ-settling time, then the undershoot satisfies

$$y_{\text{us}} \geq \frac{1-\epsilon}{e^{zt_{s,\epsilon}} - 1}.$$

Proof (i) That there must be overshoot follows as stated in Remark 9.2–3. Our definition of rise time implies that for $t < t_r$ we have $y(t) < \frac{t}{t_r}$. This means that

$$e(t) \geq \frac{1}{T_L(0)} - \frac{t}{t_r}$$

when $t \leq t_r$. Therefore, using Theorem 9.1(i),

$$\begin{aligned}
&\int_0^\infty e^{-pt} e(t)\,\mathrm{d}t = 0 \\
\implies \quad &-\int_{t_r}^\infty e^{-pt} e(t)\,\mathrm{d}t = \int_0^{t_r} e^{-pt} e(t)\,\mathrm{d}t \\
\implies \quad &-\int_{t_r}^\infty e^{-pt} e(t)\,\mathrm{d}t \geq \int_0^{t_r} e^{-pt}\big(\tfrac{1}{T_L(0)} - \tfrac{t}{t_r}\big)\,\mathrm{d}t.
\end{aligned}$$

Now we use this final inequality, along with the fact that $y_{\text{os}} \geq y(t) - 1 = (\frac{1}{T_L(0)} - 1) - e(t)$ to derive

$$\begin{aligned}
\frac{y_{\text{os}} e^{-pt}}{p} &= y_{\text{os}} \int_{t_r}^{\infty} e^{-pt} \, \mathrm{d}t \\
&\geq -\int_{t_r}^{\infty} e^{-pt} e(t) \, \mathrm{d}t + \int_{t_r}^{\infty} e^{-pt} \big(\tfrac{1}{T_L(0)} - 1\big) \, \mathrm{d}t \\
&\geq \int_0^{t_r} e^{-pt} \big(\tfrac{1}{T_L(0)} - \tfrac{t}{t_r}\big) \, \mathrm{d}t + \int_{t_r}^{\infty} e^{-pt} \big(\tfrac{1}{T_L(0)} - 1\big) \, \mathrm{d}t \\
&= \frac{(e^{pt_r} - 1)pt_r + T_L(0)(1 - e^{pt_r} + pt_r)}{p^2 t_r T_L(0) e^{pt_r}} + \Big(\frac{1}{T_L(0)} - 1\Big) \frac{e^{-pt_r}}{p} \\
&= \frac{pt_r + (e^{-pt_r} - 1) T_L(0)}{p^2 t_r T_L(0)}.
\end{aligned}$$

Thus we have the inequality

$$\frac{y_{\text{os}} e^{-pt}}{p} \geq \frac{pt_r + (e^{-pt_r} - 1) T_L(0)}{p^2 t_r T_L(0)}$$

which, with simple manipulation, gives the first inequality of this part of the proposition. For the second inequality one performs the Taylor expansion

$$e^{pt_t} = 1 + pt_r + \tfrac{1}{2} p^2 t_r^2 + \cdots$$

and a simple manipulation using $T_L(0) = 1$ gives the second inequality.

(ii) For $t \geq t_{s,\epsilon}$ we have $y(t) \geq 1 - \epsilon$. Therefore, using Theorem 9.1(ii), we have

$$\begin{aligned}
&\int_0^{\infty} e^{-zt} y(t) \, \mathrm{d}t = 0 \\
\implies \quad &-\int_0^{t_{s,\epsilon}} e^{-zt} y(t) \, \mathrm{d}t = \int_{t_{s,\epsilon}}^{\infty} e^{-zt} y(t) \, \mathrm{d}t \\
\implies \quad &-\int_0^{t_{s,\epsilon}} e^{-zt} y(t) \, \mathrm{d}t \geq \int_{t_{s,\epsilon}}^{\infty} e^{-zt} (1 - \epsilon) \, \mathrm{d}t \\
\implies \quad &-\int_0^{t_{s,\epsilon}} e^{-zt} y(t) \, \mathrm{d}t \geq \frac{(1 - \epsilon) e^{-z t_{s,\epsilon}}}{z}.
\end{aligned}$$

Now we use the definition of the undershoot and the previous inequality to compute

$$\begin{aligned}
\frac{y_{\text{us}}(1 - e^{-z t_{s,\epsilon}})}{z} &= y_{\text{us}} \int_0^{t_{s,\epsilon}} e^{-zt} \, \mathrm{d}t \\
&\geq -\int_0^{t_{s,\epsilon}} e^{-zt} y(t) \, \mathrm{d}t \\
&\geq \frac{(1 - \epsilon) e^{-z t_{s,\epsilon}}}{z}.
\end{aligned}$$

Thus we have demonstrated the inequality

$$\frac{y_{\text{us}}(1 - e^{-z t_{s,\epsilon}})}{z} \geq \frac{(1 - \epsilon) e^{-z t_{s,\epsilon}}}{z},$$

and from this the desired result follows easily. ■

Let us check the predictions of the proposition on the previous examples where overshoot and undershoot were exhibited.

9.5 Examples (**Example 9.3 cont'd**)

1. We resume looking at the loop gain $R_L(s) = \frac{2(s+1)}{s(s-1)}$ where the normalised step response for the closed-loop system is shown on the left in Figure 9.2. The closed-loop transfer function evaluated at $s = 0$ is $T_L(0) = 1$. Therefore, since we have a real pole at $s = 1$, the rise time and overshoot should together satisfy $y_{\text{os}} \geq \frac{t_r}{2}$. Numerically we determine that $y_{\text{os}} \approx 0.68$. For this example, it turns out that $t_r = 0$ (the order of the transfer function is not high enough to generate an interesting rise time with our definition. In any event, the inequality of part (i) of Proposition 9.4 is certainly satisfied.
2. Here we use the loop gain $R_L(s) = \frac{1-s}{s(s+2)}$ for which the closed-loop normalised step response is shown on the right in Figure 9.2. Taking $\epsilon = 0.05$, the undershoot and the ϵ-settling time are computed as $y_{\text{us}} \approx 0.28$ and $t_{s,\epsilon} \approx 6.03$. One computes

$$\frac{1-\epsilon}{e^{zt_{s,\epsilon}} - 1} \approx 0.002,$$

using $z = 1$. In this case, the inequality of part (ii) of Proposition 9.4 is clearly satisfied. •

Note that for the examples, the estimates are very conservative. The reason that this is not surprising is that the estimates hold for *all* systems, so one can expect that for a given example they will not be that sharp.

When we have *both* an unstable pole and a nonminimum phase zero for the loop gain R_L, it is possible again to provide estimates for the overshoot and undershoot.

9.6 Proposition *Consider the unity gain feedback loop of Figure 9.1, and suppose the closed-loop system is IBIBO stable and steppable. Also suppose that the loop gain R_L has a real pole at $p \in \mathbb{C}_+$ and a real zero at $z \in \mathbb{C}_+$. The following statements hold:*

(i) if $p < q$ then the overshoot satisfies

$$y_{\text{os}} \geq \frac{p}{T_L(0)(q-p)};$$

(ii) if $q < p$ then the undershoot satisfies

$$y_{\text{us}} \geq \frac{q}{T_L(0)(p-q)}.$$

Proof (i) From the formulas for concerning $e(t)$ from parts (i) and (ii) of Theorem 9.1 we have

$$\int_0^\infty (e^{-pt} - e^{-zt})(-e(t))\,\mathrm{d}t = \frac{1}{T_L(0)z}.$$

Since $y_{\text{os}} \geq -e(t)$ for all $t > 0$ this gives

$$\frac{1}{T_L(0)z} \leq y_{\text{os}} \int_0^\infty (e^{-pt} - e^{-zt})(-e(t))\,\mathrm{d}t = \frac{z-p}{pz}.$$

From this the result follows.

(ii) Here we again use parts (i) and (ii) of Theorem 9.1, but now we use the formulas concerning $y(t)$. This gives

$$\int_0^\infty (e^{-zt} - e^{-pt})(-y(t))\,\mathrm{d}t = \frac{1}{T_L(0)p}.$$

Since $y_{\text{us}} \geq -y(t)$ for all $t > 0$ we have

$$\frac{1}{T_L(0)p} \leq y_{\text{us}} \int_0^\infty (e^{-zt} - e^{-pt})(-y(t))\,\mathrm{d}t = \frac{p-z}{pz},$$

giving the result. ■

An opportunity to employ this result is given in Exercise E9.1. The implications are quite transparent, however. When one has an unstable real pole and a nonminimum phase real zero, the undershoot or overshoot can be expected to be large if the zero and the pole are widely separated.

9.2 Performance restrictions in the frequency domain for general systems

In the previous section we gave some general results indicating the effects on the time response of unstable poles and nonminimum phase zeros of the plant. In this section we carry this investigation into the frequency domain, following the excellent treatment of Seron, Braslavsky, and Goodwin [1997]. The setup is the unity gain loop depicted in Figure 9.3. Associated with this, of course, we have the closed-loop transfer function T_L and the sensitivity function S_L defined, as usual, by

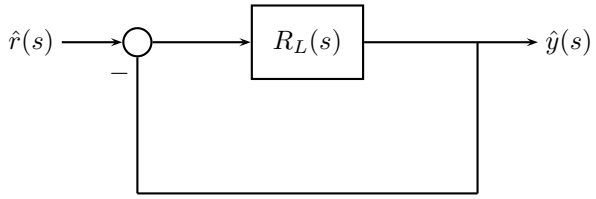


Figure 9.3 The unity gain feedback loop for investigation of performance in the frequency domain

$$T_L(s) = \frac{R_L(s)}{1 + R_L(s)}, \qquad S_L(s) = \frac{1}{1 + R_L(s)}.$$

These are obviously related, and in this section the aim is to explore fully the consequences of this relationship, as well as explore the behaviour of these two rational functions as the loop gain R_L has poles and zeros in $\overline{\mathbb{C}}_+$. If one of the goals of system design is to reduce error in the closed-loop system, then since the transfer function from input to error is the sensitivity function, a goal might be to reduce the sensitivity function. However, it is simply not possible to do this in any possible manner, and the zeros and poles of the loop gain R_L have a great deal to say about what is and is not possible.

9.2.1 Bode integral formulae

Throughout this section, we are dealing with the unity gain feedback loop of Figure 9.3. The following result gives what are called the ***Bode integral formulae***. These results,

for stable and minimum phase loop gains, are due to Horowitz [1963]. The extension to unstable loop gains for the sensitivity function are due to Freudenberg and Looze [1985], and for nonminimum phase loop gains for the transfer function results are due to Middleton and Goodwin [1990].

9.7 Theorem *Consider the feedback interconnection of Figure 9.3 and suppose that $R_L \in \mathbb{R}(s)$ is proper and has poles $p_1, \dots, p_k$ in $\mathbb{C}_+$ and zeros $z_1, \dots, z_\ell$ in $\mathbb{C}_+$. If the interconnection is IBIBO stable and if R_L has no poles on the imaginary axis then the following statements hold:*

(i) if the closed-loop system is well-posed then

$$\int_0^\infty \ln|\frac{S_L(i\omega)}{S_L(i\infty)}| \, \mathrm{d}\omega = \frac{\pi}{2} \lim_{s\to\infty} \frac{s(S_L(s) - S_L(\infty))}{S_L(\infty)} + \pi \sum_{j=1}^{k} p_j;$$

(ii) if $L(0) \neq 0$ then

$$\int_0^\infty \ln|\frac{T_L(i\omega)}{T_L(0)}| \, \frac{\mathrm{d}\omega}{\omega^2} = \frac{\pi}{2} \frac{1}{T_L(0)} \lim_{s\to 0} \frac{\mathrm{d}T_L(s)}{\mathrm{d}s} + \pi \sum_{j=1}^{\ell} \frac{1}{z_j}.$$

Proof (i) For $R > 0$ sufficiently large and $\epsilon > 0$ sufficiently small, we construct a contour $\Gamma_{R,\epsilon} \subset \overline{\mathbb{C}}_+$ comprised of $3k + 2$ arcs as follows. We define $\Gamma_\epsilon \subset i\mathbb{R}$ by

$$\Gamma_\epsilon = i\mathbb{R} \setminus \{[\mathrm{Im}(p) - \epsilon, \mathrm{Im}(p) + \epsilon] \mid p \in \{p_1, \dots, p_k\}\}.$$

Now for $j \in \{1, \dots, k\}$ define 3 contours $\Gamma^1_{\epsilon,j}$, $\Gamma^2_{\epsilon,j}$, and $\Gamma^3_{\epsilon,j}$ by

$$\begin{aligned} \Gamma^1_{\epsilon,j} &= \{x + i(\mathrm{Im}(p_j) - \epsilon) \in \mathbb{C} \mid x \in [0, \mathrm{Re}(p_j)]\} \\ \Gamma^2_{\epsilon,j} &= \{p_j + re^{i\theta} \in \mathbb{C} \mid \theta \in [\tfrac{-\pi}{2}, \tfrac{\pi}{2}]\} \\ \Gamma^3_{\epsilon,j} &= \{x + i(\mathrm{Im}(p_j) + \epsilon) \in \mathbb{C} \mid x \in [0, \mathrm{Re}(p_j)]\} \end{aligned}$$

Now define

$$\Gamma_R = \{Re^{i\theta} \in \mathbb{C} \mid \theta \in [\tfrac{-\pi}{2}, \tfrac{\pi}{2}]\}$$

The contour $\Gamma_{R,\epsilon}$ is the union of these contours, and a depiction of it is shown in Figure 9.4. Since $\ln \frac{S_L(s)}{S_L(\infty)}$ is analytic on and inside the contour $\Gamma_{R,\epsilon}$ for R sufficiently large and ϵ sufficiently small, we have

$$\lim_{\substack{\epsilon\to 0 \\ R\to\infty}} \int_{\Gamma_{R,\epsilon}} \ln \frac{S_L(s)}{S_L(\infty)} \, \mathrm{d}s = 0.$$

Let us analyse this integral bit by bit. Using the expression

$$\ln \frac{S_L(s)}{S_L(\infty)} = \ln|\frac{S_L(s)}{S_L(\infty)}| + i\angle S_L(s)/S_L(\infty),$$

we have

$$\lim_{\substack{\epsilon\to 0 \\ R\to\infty}} \int_{\Gamma_\epsilon} \ln \frac{S_L(s)}{S_L(\infty)} \, \mathrm{d}s = 2i \int_0^\infty \ln \frac{S_L(i\omega)}{S_L(i\infty)} \, \mathrm{d}\omega.$$

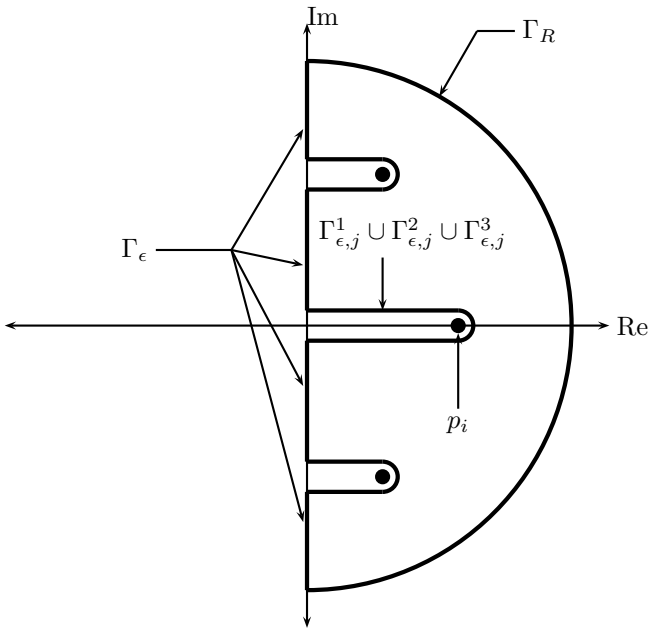


Figure 9.4 The contour for the Bode integral formulae

For $j \in \{1, \ldots, k\}$, define $f_j(s)$ by requiring that

$$\ln \frac{S_L(s)}{S_L(\infty)} = \ln(s - p_j) + \ln f_j(s).$$

Let $\Gamma_{\epsilon,j} = \Gamma^1_{\epsilon,j} \cup \Gamma^2_{\epsilon,j} \cup \Gamma^3_{\epsilon,j}$, and let $\tilde{\Gamma}_{\epsilon,j}$ be the contour obtained by closing $\Gamma_{\epsilon,j}$ along the imaginary axis. On and within $\tilde{\Gamma}_{\epsilon,j}$ the function f_j is analytic. Therefore, by Cauchy's Integral Theorem,

$$\int_{\tilde{\Gamma_{\epsilon,j}}} \ln f_j(s)\,\mathrm{d}s = \int_{\Gamma_{\epsilon,j}} \ln f_j(s)\,ds + \int_{\mathrm{Im}(p_j)+i\epsilon}^{\mathrm{Im}(p_j)-i\epsilon} \ln f_j(i\omega)\,\mathrm{d}\omega = 0$$

Combining this with

$$\int_{\Gamma_{\epsilon,j}} \ln \frac{S_L(s)}{S_L(\infty)}\,\mathrm{d}s = \int_{\Gamma_{\epsilon,j}} \ln(s - p_j)\,\mathrm{d}s + \int_{\Gamma_{\epsilon,j}} \ln f_j(s)\,\mathrm{d}s$$

gives

$$\int_{\Gamma_{\epsilon,j}} \ln \frac{S_L(s)}{S_L(\infty)}\,\mathrm{d}s = \int_{\Gamma_{\epsilon,j}} \ln(s - p_j)\,\mathrm{d}s + \int_{\mathrm{Im}(p_j)-i\epsilon}^{\mathrm{Im}(p_j)+i\epsilon} \ln f_j(i\omega)\,\mathrm{d}\omega.$$

In the limit as ϵ goes to zero, the second integral on the right will vanish. Thus we shall evaluate the first integral on the right. Since $\ln(s - p_j)$ is analytic on $\Gamma_{\epsilon,j}$ we use the fact that the antiderivative of $\ln z$ is $z \ln z - z$ to compute

$$\begin{aligned}
\int_{\Gamma_{\epsilon,j}} \ln(s - p_j)\,\mathrm{d}s &= \big((s - p_j)\ln(s - p_j) - (s - p_j)\big)\Big|_{\mathrm{Im}(p_j)-i\epsilon}^{\mathrm{Im}(p_j)+i\epsilon} \\
&= (-\mathrm{Re}(p_j) + i\epsilon)\ln(-\mathrm{Re}(p_j) + i\epsilon) - (-\mathrm{Re}(p_j) + i\epsilon) - \\
&\quad (-\mathrm{Re}(p_j) - i\epsilon)\ln(-\mathrm{Re}(p_j) - i\epsilon) + (-\mathrm{Re}(p_j) - i\epsilon) \\
&= i\mathrm{Re}(p_j)\big(\measuredangle(-\mathrm{Re}(p) - i\epsilon) - \measuredangle(-\mathrm{Re}(p) + i\epsilon)\big).
\end{aligned}$$

Taking the limit as $\epsilon \to 0$ gives

$$\lim_{\epsilon\to 0}\int_{\Gamma_{\epsilon,j}} \ln\frac{S_L(s)}{S_L(\infty)}\,\mathrm{d}s = -2\pi i \mathrm{Re}(p_j).$$

This then gives

$$\lim_{\epsilon\to 0}\sum_{j=1}^{k}\int_{\Gamma_{\epsilon,j}} = -2\pi i\sum_{j=1}^{k}\mathrm{Re}(p_i) = -2\pi i\sum_{j=1}^{k} p_j,$$

with the last equality holding since poles of R_L occur in complex conjugate pairs.

Now we look at the contour Γ_R. We have

$$\lim_{R\to\infty}\ln\frac{S_L(s)}{S_L(\infty)}\,\mathrm{d}s = i\pi \operatorname*{Res}_{s=\infty}\ln\frac{S_L(s)}{S_L(\infty)}.$$

Since R_L is proper we may write its Laurent expansion at $s=\infty$ as

$$R_L(s) = R_L(\infty) + \frac{c_{-1}}{s} + \frac{c_{-2}}{s^2} + \cdots . \tag{9.1}$$

Since $S_L(s) = (1+R_L(s))^{-1}$ we have

$$\begin{aligned}\ln S_L(s) =& -\ln(1+R_L(s))\\ =& -\ln(1+R_L(\infty)) - \ln\Big(1 + \frac{c_{-1}}{1+R_L(\infty)}\frac{1}{s} + \frac{c_{-2}}{1+R_L(\infty)}\frac{1}{s^2} + \cdots\Big).\end{aligned}$$

Therefore, writing $\ln S_L(\infty) = -\ln(1+R_L(\infty))$, we have

$$\ln\frac{S_L(s)}{S_L(\infty)} = -\ln\big(1+\tilde{R}_L(s)\big),$$

where

$$\tilde{R}_L(s) = \frac{S_L(\infty)c_{-1}}{s} + \frac{S_L(\infty)c_{-2}}{s^2} + \cdots .$$

Using the power series expansion for $\ln(1+z)$,

$$\ln(1+z) = z + \frac{z^2}{2} + \cdots ,$$

for $|z|<1$ we have

$$\begin{aligned}\ln\frac{S_L(s)}{S_L(\infty)} =& -\tilde{R}_L(s) + \frac{\tilde{R}_L^2(s)}{s} + \cdots\\ =& -\frac{S_L(\infty)c_{-1}}{s} - \frac{S_L(\infty)c_{-2}}{s^2} + \cdots ,\end{aligned}$$

which is valid for s sufficiently large. This shows that

$$\operatorname*{Res}_{s=\infty}\ln\frac{S_L(s)}{S_L(\infty)} = S_L(\infty)c_{-1}.$$

Now note that (9.1) implies that

$$\begin{aligned} c_{-1} &= \lim_{s\to\infty} s(R_L(s) - R_L(\infty)) \\ &= \lim_{s\to\infty} s\Big(\frac{1}{S_L(s)} - \frac{1}{S_L(\infty)}\Big) \\ &= \frac{1}{S_L^2(\infty)} \lim_{s\to\infty} s\big(S_L(\infty) - S_L(s)\big). \end{aligned}$$

This gives

$$\lim_{R\to\infty} \int_{\Gamma_R} \ln \frac{S_L(s)}{S_L(\infty)} = i\pi \frac{1}{S_L(\infty)} \lim_{s\to\infty} s\big(S_L(\infty) - S_L(s)\big).$$

The result now follows by summing the contours.

(ii) Here we define $G(z) = R_L(1/z)$ and

$$H(z) = \frac{1}{1 + 1/T_L(1/z)} = \frac{1}{1 + G(z)}.$$

Thus G and H now play the part of R_L and S_L in part (i). Thus we have

$$\int_0^\infty \ln|\frac{H(i\Omega)}{H(i\infty)}|\,\mathrm{d}\Omega = \frac{\pi}{2} \lim_{z\to\infty} \frac{z(H(z) - H(\infty))}{H(\infty)} + \pi \sum_{j=1}^{\ell} \frac{1}{z_j}.$$

Let us first show that

$$\lim_{z\to\infty} \frac{z(H(z) - H(\infty))}{H(\infty)} = \frac{1}{T_L(0)} \lim_{s\to 0} \frac{\mathrm{d}T_L(s)}{\mathrm{d}s}.$$

First recall that

$$\lim_{z\to\infty} \frac{z(H(z) - H(\infty))}{H(\infty)} = -\operatorname*{Res}_{z=\infty} \ln H(z),$$

following our calculations in the previous part of the proof. If the Taylor expansion of $\ln T_L(s)$ about $s = 0$ is

$$\ln T_L(s) = a_0 + a_1 s + a_2 s^2 + \cdots,$$

then the Laurent expansion of $\ln H(z)$ about $z = \infty$ is

$$\ln H(z) = a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} + \cdots.$$

Therefore $\operatorname{Res}_{z=\infty} \ln H(z) = -a_1$. However, we also have

$$a_1 = \lim_{s\to 0} \frac{\mathrm{d}\ln T_L(s)}{\mathrm{d}s} = \frac{1}{T_L(0)} \lim_{s\to 0} \frac{\mathrm{d}T_L(s)}{\mathrm{d}s}.$$

Thus we have

$$\int_0^\infty \ln|\frac{H(i\Omega)}{H(i\infty)}|\,\mathrm{d}\Omega = -\frac{\pi}{2} \frac{1}{T_L(0)} \lim_{s\to 0} \frac{\mathrm{d}T_L(s)}{\mathrm{d}s}.$$

The result now follows by making the change of variable $\omega = \frac{1}{\Omega}$. ■

Although the theorem is of some importance, its consequences do require some explication. This will be done further in the next section, but here let us make some remarks that can be easily deduced.

9.8 Remarks

1. Let us provide an interpretation for the first term on the right-hand side in part (i). Using the definition of S_L we have

$$\begin{aligned}\lim_{s\to\infty} \frac{s(S_L(s) - S_L(\infty))}{S_L(\infty)} &= \lim_{s\to\infty} \frac{s(R_L(\infty) - RL(s))}{1 + R_L(s)} \\ &= -\frac{\dot{\imath}_{N,D}(0+)}{1 + R_L(\infty)},\end{aligned}$$

where the last step follows from Proposition E.9(i). Here (N, D) is the c.f.r. of R_L. On this formula, let us make some remarks.

   (a) If the relative degree of R_L is greater than 0 then $R_L(\infty) = 0$.

   (b) If the relative degree of the plant is greater than 1 then $\dot{\imath}_{N,D}(0+) = 0$.

   (c) From the previous two remarks, if the relative degree of R_L is greater than 1 and if the plant has no poles in $\mathbb{C}_+$, then the formula

$$\int_0^\infty \ln|S_L(i\omega)| \, \mathrm{d}\omega = 0$$

   holds, since $S_L(i\infty) = 1$ in these cases. This is the formula originally due to Horowitz [1963], and is called the ***Horowitz area formula*** for the sensitivity function. It tells us that the average of the area under the magnitude part of the sensitivity Bode plot should be zero. Therefore, for any frequency intervals where S_L is less that 1, there are also frequency regions where S_L will be greater than 1.

   (d) If the relative degree of R_L is greater than 1 but R_L has poles in $\mathbb{C}_+$, then the formula from part (i) reads

$$\int_0^\infty \ln|S_L(i\omega)| \, \mathrm{d}\omega = \sum_{j=1}^{k} p_j.$$

   Since the right-hand side is positive, this tells us that the poles in $\mathbb{C}_+$ have the effect of shifting the area bias of the sensitivity function in the positive direction. That is to say, with poles for R_L in $\mathbb{C}_+$, there will be a greater frequency range for which S_L will be greater than 1.

   (e) Now let us consider the cases where the relative degree of R_L is 0 or 1, and where R_L has poles in $\mathbb{C}_+$. Here, if $\dot{\imath}_{N,D}(0+) > 0$ we have an opportunity to reduce the detrimental effect of the positive contribution to the area integral from the poles.

2. Let us try to understand part (ii) in the same manner by understanding the first term on the right-hand side. Define the scaled closed-loop transfer function $\tilde{T}_L(s) = \frac{T_L(s)}{T_L(0)}$. Let $e(t)$ be the error signal for the unit ramp input of the transfer function $\tilde{T}_L$. Thus $s^2\hat{e}(t) = \big(1 - \frac{T_L(s)}{T_L(0)}\big)$. We compute

$$\begin{aligned}\frac{1}{T_L(0)} \lim_{s\to 0} \frac{\mathrm{d}T_L(s)}{\mathrm{d}s} &= \lim_{s\to 0} \frac{\frac{T_L(s)}{T_L(0)} - 1}{s} \\ &= -\lim_{s\to 0} s\hat{e}(s) \\ &= -\lim_{t\to\infty} e(t),\end{aligned}$$

where in the final step we have used Proposition E.9(ii). Let us make some remarks, given this formula.

(a) If the scaled closed-loop system is type k for $k > 1$ then the steady-state error to the ramp input will be zero.

(b) Based upon the previous remark, if the scaled closed-loop system is type k for $k > 1$ and if the loop gain R_L has no zeros in $\mathbb{C}_+$, then the formula

$$\int_0^\infty \ln|\frac{T_L(i\omega)}{T_L(0)}| \frac{\mathrm{d}\omega}{\omega^2} = 0$$

holds. Again, this is to be seen as an area integral for the magnitude Bode plot for T_L, but now it is weighted by $\frac{1}{\omega^2}$. Thus the magnitude smaller frequencies counts for more in this integral constraint. This formula, like its sensitivity function counterpart, was first derived by Horowitz [1963], and is the ***Horowitz area formula*** for the closed-loop transfer function.

(c) If the scaled closed-loop system is type k for $k > 1$ but the loop gain does have zeros in $\mathbb{C}_+$, then the formula

$$\int_0^\infty \ln|\frac{T_L(i\omega)}{T_L(0)}| \frac{\mathrm{d}\omega}{\omega^2} = \sum_{j=1}^{\ell} \frac{1}{z_j}$$

holds. Thus we see that as with poles for the sensitivity function integral, the zeros for the loop gain shift the area up. Now we note that this effect gets worse as the zeros approach the imaginary axis.

(d) Finally, suppose that the scaled closed-loop system is type 1 (so that the steady-state error to the unit ramp is constant) and that R_L has zeros in $\mathbb{C}_+$. Then, provided that the steady-state error to the unit ramp input is positive, we can compensate for the effect of the zeros on the right hand side only by making the steady-state error to the ramp larger. •

9.2.2 Bandwidth constraints

The preceding discussion has alerted us to problems we may encounter in trying to arbitrarily shape the Bode plots for the closed-loop transfer function and the sensitivity function. Let us examine this matter further by seeing how bandwidth constraints come into play.

For the unity gain closed-loop interconnection of Figure 9.3, if ω_{b} is the bandwidth for the closed-loop transfer function, then we must have

$$\begin{aligned} & \frac{|R_L(i\omega)|}{|1 + R_L(i\omega)|} \leq \frac{1}{\sqrt{2}}, \qquad \omega > \omega_{\mathrm{b}} \\ \implies \qquad & |R_L(i\omega)| \leq \frac{1}{\sqrt{2}+1}, \qquad \omega > \omega_{\mathrm{b}} \end{aligned}$$

This implies that bandwidth constraints for the closed-loop system translate into bandwidth constraints for the loop gain, and vice versa. Typically, the bandwidth of the loop gain will be limited by the plant, and the components available for the controller. The upshot is that when performing a controller design, the designer will typically be confronted with an inequality of the form

$$|R_L(i\omega)| \leq \delta \Big(\frac{\omega_{\mathrm{b}}}{\omega}\Big)^{1+k}, \qquad \omega > \omega_{\mathrm{b}}, \tag{9.2}$$

where $\delta < \frac{1}{2}$ and where $k \in \mathbb{N}$. The objective in this section is to see how these constraints translate into constraints on the sensitivity function. This issue also came up in Section 8.6.1 in the discussion of high-frequency roll-off constraints.

The following result gives a bound on the area under the "tail" of the sensitivity function magnitude Bode plot.

9.9 Proposition *For the closed-loop interconnection of Figure 9.3, if R_L has relative degree of greater than 1 and if R_L satisfies a bandwidth constraint of the form (9.2), then*

$$|\int_{\omega_{\rm b}}^{\infty} \ln|S_L(i\omega)|\,\mathrm{d}\omega| \leq \frac{\delta\omega_{\rm b}}{2k}.$$

Proof We first note without proof that if $|s| < \frac{1}{2}$ then $|\ln(1+s)| < \frac{3|s|}{2}$. We then compute

$$\begin{aligned}
|\int_{\omega_{\rm b}}^{\infty} \ln|S_L(i\omega)|\,\mathrm{d}\omega| &\leq \int_{\omega_{\rm b}}^{\infty} |\ln|S_L(i\omega)||\,\mathrm{d}\omega \\
&\leq \int_{\omega_{\rm b}}^{\infty} |\ln S_L(i\omega)|\,\mathrm{d}\omega \\
&= \int_{\omega_{\rm b}}^{\infty} |\ln(1+R_L(i\omega)|\,\mathrm{d}\omega \\
&\leq \frac{3\delta\omega_{\rm b}^{1+k}}{2} \int \frac{1}{\omega^{1+k}}\,\mathrm{d}\omega \\
&= \frac{\delta\omega_{\rm b}}{2k},
\end{aligned}$$

as desired. ■

Let us see if we can explain the point of this.

9.10 Remark The idea is this. From Theorem 9.7(i) we know that since the relative degree of R_L is at least 2, the total area under the sensitivity function magnitude Bode plot will be nonnegative. This means that if we have a region where we have made the sensitivity function smaller than 1, there must be a region where the sensitivity function is larger than 1. What one can hope for, however, is that one can smear this necessary increase in magnitude of the sensitivity function over a large frequency range and so keep the maximum magnitude of the sensitivity function under control. Proposition 9.9 tells us that if the loop gain has bandwidth constraints, then the area contributed by the sensitivity function above the bandwidth for the loop gain is limited. Therefore, the implication is necessarily that if one wishes to significantly decrease the sensitivity function magnitude over a frequency range, there must be a significant increase in the sensitivity function magnitude at frequencies below the loop gain bandwidth. An attempt to illustrate this is given in Figure 9.5. In the figure, the sensitivity function is made small in the frequency range $[\omega_1, \omega_2]$, and the cost for this is that there is a large peak in the sensitivity function magnitude below the bandwidth since the contribution above the bandwidth is limited by Proposition 9.9. •

9.2.3 The waterbed effect

Let us first look at a phenomenon that indicates problems that can be encountered in trying to minimise the sensitivity function over a frequency range. This result is due to Francis and Zames [1984].

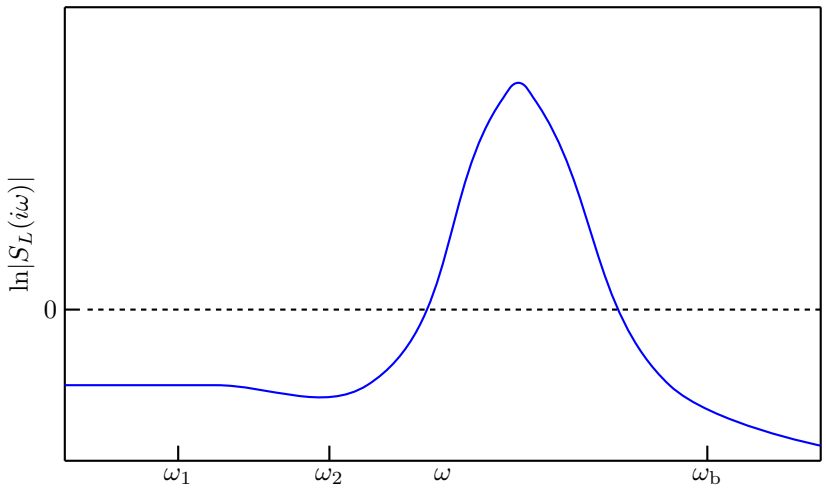


Figure 9.5 The effect of loop gain bandwidth constraints on the sensitivity function area distribiution

9.11 Theorem *Consider the feedback interconnection of Figure 9.7, and assume that the closed-loop system is IBIBO stable and well-posed. If R_L has zero in $\mathbb{C}_+$ then for $0 < \omega_1 < \omega_2$ there exists $m > 0$ so that*

$$\sup_{\omega\in[\omega_1,\omega_2]} |S_L(i\omega)| \geq \|S_L\|_\infty^{-m}.$$

Proof Suppose that $\zeta \in \mathbb{C}_+$ is a zero of R_L. Then $S_L(q) = 1$ by Proposition 8.22. Let

$$D(0,1) = \{z \in \mathbb{C} \mid |z| \leq 1\}$$

be the unit complex disk and define a mapping $\phi\colon \overline{\mathbb{C}}_+ \to D(0,1)$ by

$$\phi(s) = \frac{\zeta - s}{\bar{\zeta} + s}.$$

This map is invertible and has inverse

$$\phi^{-1}(z) = \frac{\zeta - \bar{\zeta}z}{1 + z}.$$

The interval $i[\omega_1, \omega_2] \subset i\mathbb{R}$ is mapped under ϕ to that portion of the boundary of $D(0,1)$ given by

$$\phi(i[\omega_1,\omega_2]) = \{e^{i\theta} \mid \theta \in [\theta_1,\theta_2]\},$$

where

$$e^{i\theta_j} = \frac{\zeta - i\omega_j}{\bar{\zeta} + i\omega_j}, \qquad j = 1, 2.$$

Define $\phi_* S_L\colon D(0,1) \to \mathbb{C}$ by $(\phi_* S_L)(z) = S_L(\phi^{-1}(z))$. Thus $\phi_* S_L$ is analytic in $D(0,1)$ and $\phi_* S_L(0) = 1$. Furthermore, the following equalities hold:

$$\sup_{\theta\in[\theta_1,\theta_2]} |\phi_* S_L(e^{i\theta})| = \sup_{\omega\in[\omega_1,\omega_2]} |S_L(i\omega)|, \qquad \sup_\theta |\phi_* S_L(e^{i\theta})| = \|S_L\|_\infty.$$

Define $\phi = |\theta_2 - \theta_1|$ and let $n > \frac{2\pi}{\phi}$ be an integer. Define $f\colon D(0,1) \to \mathbb{C}$ as

$$f(z) = \prod_{j=0}^{n-1} \phi_* S_L\big(ze^{2\pi ji/n}\big).$$

Being the product of analytic functions in $D(0,1)$, f itself is analytic in $D(0,1)$. For each z on the boundary of $D(0,1)$ there exists $j \in \{0,\dots,n-1\}$ so that

$$ze^{2\pi ji/n} \in \phi(i[\omega_1,\omega_2]).$$

Thus we have, using the Maximum Modulus Principle (Theorem D.9),

$$\begin{aligned} 1 &= |f(0)| \\ &\le \sup_\theta |f(e^{i\theta})| \\ &= \big(\sup_\theta |\phi_* S_L(e^{i\theta})|\big)^{n-1} \sup_{\theta\in[\theta_1,\theta_2]} |\phi_* S_L(e^{i\theta})| \\ &= \|S_L\|_\infty^{n-1} \sup_{\omega\in[\omega_1,\omega_2]} |S_L(i\omega)|. \end{aligned}$$

Taking $m = n-1$, the result follows. ■

9.12 Remark The idea here is simple. If, for a nonminimum phase plant, one wishes to reduce the sensitivity function over a certain frequency range, then one can expect that the $\|S_L\|_\infty$ will be increased in consequence. •

9.2.4 Poisson integral formulae

Now we look at other formulae that govern the behaviour of the closed-loop transfer function and the sensitivity function. In order to state the results in this section, we need to represent the closed-loop transfer function and the sensitivity function in a particular manner. Recall from (8.2) the definitions $Z(S_L)$ and $Z(T_L)$ of the sets of zeros for S_L and T_L in $\mathbb{C}_+$. Suppose that $Z(S_L) = \{p_1,\dots,p_k\}$ and $Z(T_L) = \{z_1,\dots,z_\ell\}$, the notation reflecting the fact that zeros for S_L are poles for R_L and zeros for T_L are zeros for R_L. Corresponding to these zeros are the ***Blaschke products*** for S_L and T_L defined by

$$B_{S_L}(s) = \prod_{j=1}^{k} \frac{p_j - s}{\bar{p}_j + s}, \qquad B_{T_L}(s) = \prod_{j=1}^{\ell} \frac{z_j - s}{\bar{z}_j + s}.$$

These functions all have unit magnitude. Now we define $\tilde{R}_L$ by the equality

$$R_L(s) = \frac{\tilde{R}_L(s) B_{T_L}(s)}{B_{S_L}(s)}. \tag{9.3}$$

In Section 14.2.2 we shall refer to this as an inner/outer factorisation of R_L. The key fact here is that the zeros and poles for R_L have been soaked up into the Blaschke products so that $\tilde{R}_L(s)$ has no zeros or poles in $\mathbb{C}_+$. We similarly define $\tilde{S}_L$ and $\tilde{T}_L$ by

$$S_L(s) = \tilde{S}_L(s) B_{S_L}(s), \qquad T_L(s) = \tilde{T}_L(s) B_{T_L}(s),$$

and observe that $\tilde{S}_L$ and $\tilde{T}_L$ have no zeros in $\mathbb{C}_+$.

With this notation, we state the ***Poisson integral formulae*** for the sensitivity function and the closed-loop transfer function.

9.13 Theorem *Consider the closed-loop interconnection of Figure 9.7. If the closed-loop system is IBIBO stable, the following equalities hold:*

(i) if $z = \sigma_z + i\omega_z \in \mathbb{C}_+$ *is a zero of* R_L *then*

$$\int_{-\infty}^{\infty} \ln|S_L(i\omega)| \frac{\sigma_z}{\sigma_z^2 + (\omega - \omega_z)^2} \, \mathrm{d}\omega = \pi \ln|B_{S_L}^{-1}(z)|;$$

(ii) if $p = \sigma_p + i\omega_p \in \mathbb{C}_+$ *is a pole of* R_L *then*

$$\int_{-\infty}^{\infty} \ln|T_L(i\omega)| \frac{\sigma_p}{\sigma_p^2 + (\omega - \omega_p)^2} \, \mathrm{d}\omega = \pi \ln|B_{T_L}^{-1}(p)| + \pi\sigma_p.$$

Proof (i) This follows by applying Corollary D.8 to the function $\tilde{S}_L$. To get the result, one simply observes that $\ln \tilde{S}_L$ is analytic and nonzero in $\overline{\mathbb{C}}_+$, that $|S_L(i\omega)| = |\tilde{S}_L(i\omega)|$ for all $\omega \in \mathbb{R}$, and that $S_L(z) = 1$ from Proposition 8.22.

(ii) The idea here, like part (i), follows from Corollary D.8. In this case one observes that $\ln \tilde{T}_L$ is analytic and nonzero in $\overline{\mathbb{C}}_+$, that $|T_L(i\omega)| = |\tilde{T}_L(i\omega)|$ for all $\omega \in \mathbb{R}$, and that $T_L(p) = 1$ from Proposition 8.22. ■

Let us now make some observations concerning the implications of the Poisson integral formulae as it bears on controller design. Before we say anything formal, let us make some easy observations.

9.14 Remarks

1. Like the Bode integral formulae, the Poisson integral formulae provide limits on the behaviour of the magnitude portion of the Bode plots for the sensitivity and complementary sensitivity functions.
2. Note that the weighting function

$$W_s(\omega) = \frac{\sigma_s}{\sigma_s^2 + (\omega - \omega_s)^2} \tag{9.4}$$

in each of the integrands is positive, and that the Blaschke products satisfy $|B_{S_L}^{-1}(s)| \geq 1$ and $|B_{T_L}^{-1}(s)| \geq 1$ for $s \in \mathbb{C}_+$. Thus we see that the Poisson integral formulae do indeed indicate that, for example, if there are zeros for the plant in $\mathbb{C}_+$ then the weighted integral of the sensitivity function will be positive.
3. If a plant is both unstable and nonminimum phase, then one concludes that the weighted area of sensitivity increase is greater than that for sensitivity decrease. One sees that this effect is exacerbated if there is a near cancellation of an unstable pole and a nonminimum phase zero. •

Now let us make some more structured comments about the implications of the Poisson integral formulae. Let us begin by examining carefully the weighting function (9.4) that appears in the formulae.

9.15 Lemma *For* $s_0 \in \mathbb{C}_+$ *and* $\omega_1 > 0$ *define*

$$\Theta_{s_0}(\omega_1) = \int_{-\omega_1}^{\omega_1} W_{s_0}(\omega_1) \, \mathrm{d}\omega.$$

Then the following statements hold:

(i) if s_0 is real then

$$\Theta_{s_0}(\omega_1) = -\measuredangle\Big(\frac{s_0 - i\omega_1}{s_0 + i\omega_1}\Big);$$

(ii) if s_0 is not real then

$$\Theta_{s_0}(\omega_1) = -\frac{1}{2}\Big(\measuredangle\Big(\frac{s_0 - i\omega_1}{\bar{s}_0 + i\omega_1}\Big) + \measuredangle\Big(\frac{\bar{s}_0 - i\omega_1}{s_0 + i\omega_1}\Big)\Big).$$

Proof Write $s_0 = \sigma_0 + i\omega_0$. The lemma follows from the fact that

$$\int_{-\omega_1}^{\omega_1} \frac{\sigma_0}{\sigma_0^2 + (\omega - \omega_0)^2}\,\mathrm{d}\omega = \arctan\frac{\omega_1 - \omega_0}{\sigma_0} + \arctan\frac{\omega_1 + \omega_0}{\sigma_0}.$$ ■

One should interpret this result as telling us that the length of the interval $[-\omega_1, \omega_1]$, weighted by W_{s_0}, is related to the phase lag incurred by $s_0 \in \mathbb{C}_+$. Of course, in the Poisson integral formulae, this phase lag arises from nonminimum phase zeros and/or unstable poles. This precise description of the weighting function allows us to provide some estimates pertaining to reduction of the sensitivity function, along the same lines as given by the waterbed affect, Theorem 9.11. Thus, for some $\omega_1 > 0$ and for some positive $\epsilon_1 < 1$ we wish for the sensitivity function to satisfy

$$|S_L(i\omega)| \leq \epsilon_1, \quad \omega \in [-\omega_1, \omega_1]. \tag{9.5}$$

The following result tells what are the implications of such a demand on the sensitivity function at other frequencies.

9.16 Corollary *Suppose that a proper loop gain R_L has been factored as in (9.3) and that the interconnection of Figure 9.3 is IBIBO stable. If the inequality (9.5) is satisfied and if $z \in \mathbb{C}_+$ is a zero for R_L then we have*

$$\|S_L\|_\infty \geq \Big(\frac{1}{\epsilon_1}\Big)^{\frac{\Theta_z(\omega_1)}{\pi - \Theta_z(\omega_1)}} |B_{S_L}^{-1}(z)|^{\frac{\pi}{\pi - \Theta_z(\omega_1)}}$$

Proof Using Theorem 9.13 we compute

$$\begin{aligned}\pi \ln|B_{S_L}^{-1}(z)| &= \int_{-\infty}^{\infty} \ln|S_L(i\omega)|W_z(\omega)\,\mathrm{d}\omega \\ &= \int_{-\infty}^{-\omega_1} \ln|S_L(i\omega)|W_z(\omega)\,\mathrm{d}\omega + \int_{-\omega_1}^{\omega_1} \ln|S_L(i\omega)|W_z(\omega)\,\mathrm{d}\omega + \\ &\quad \int_{\omega_1}^{\infty} \ln|S_L(i\omega)|W_z(\omega)\,\mathrm{d}\omega.\end{aligned}$$

This implies that

$$\Theta_z(\omega_1)\ln\epsilon_1 + \ln\|S_L\|_\infty(\pi - \Theta_z(\omega_1)) \geq \pi \ln|B_{S_L}^{-1}(z)|.$$

Exponentiating this inequality gives the result. ■

Since $|B_{S_L}^{-1}(z)| \geq 1$, $\epsilon_1 < 1$, and $\Theta_z(\omega_1) < \pi$, it follows that $\|S_L\|_\infty > 1$. Thus, as with the waterbed effect, the demand that the magnitude of the sensitivity function be made smaller than 1 over a certain frequency range guarantees that its H_∞-norm will be greater than 1. In some sense, the estimate of Corollary 9.16 contains more information since it contains

explicitly in the estimate the location of the zero. This enables us to say, for example, that if the phase lag contributed by the nonminimum phase zero is large over the specified frequency range (that is, $\Theta_z(\omega_1)$ is near π), then the effect is to further increase the lower bound on the H_∞-norm of S_L. To ameliorate this effect, one could design a controller so that the loop gain magnitude is small for frequencies at or above the frequency where the phase lag contributed by z is large.

A parallel process can be carried out for the complementary sensitivity function portion of Theorem 9.13, i.e., for part (ii). In this case, a design objective is not to reduce T_L over a finite frequency range, but perhaps to ensure that for large frequencies the complementary sensitivity function is not large. This is an appropriate objective where model uncertainties are a consideration, and high frequency effects can be detrimental to closed-loop stability. Thus here we ask that for some $\omega_2 > 0$ and some positive $\epsilon_2 < 1$ we have

$$|T_L(i\omega)| \leq \epsilon_2, \quad |\omega| \geq \omega_2. \tag{9.6}$$

In this case, calculations similar to those of Corollary 9.16 lead to the following result.

9.17 Corollary *Suppose that a proper loop gain R_L has been factored as in (9.3) and that the interconnection of Figure 9.3 is IBIBO stable. If the inequality (9.6) is satisfied and if $p \in \mathbb{C}_+$ is a pole for R_L then we have*

$$\|T_L\|_\infty \geq \left(\frac{1}{\epsilon_2}\right)^{\frac{\pi - \Theta_p(\omega_2)}{\Theta_p(\omega_2)}} |B_{T_L}^{-1}(p)|^{\frac{\pi}{\Theta_p(\omega_2)}}.$$

The best way to interpret this result differs somewhat from how we interpret Corollary 9.16, since reduction of the H_∞-norm of T_L is not a design objective. However, the closed-loop bandwidth is important, and Corollary 9.17 can be parlayed into an estimate as follows.

9.18 Corollary *Suppose that R_L is proper and minimum phase, that the interconnection of Figure 9.3 is IBIBO stable, and that R_L has a real pole $p \in \mathbb{C}_+$. If T_L satisfies (9.6) for $\epsilon_2 = \frac{1}{\sqrt{2}}$ then we have*

$$\|T_L\|_\infty \geq \sqrt{2}^{\frac{\pi - \Theta_p(\omega_\mathrm{b})}{\Theta_p(\omega_\mathrm{b})}},$$

where ω_b is the bandwidth for the closed-loop system. If we further ask that the lower bound on the H_∞-norm of T_L be bounded from above by $T_{\max}$, then we have

$$\omega_\mathrm{b} \geq p \tan\left(\frac{\pi}{2 + 2\frac{\ln T_{\max}}{\ln \sqrt{2}}}\right).$$

Proof The first inequality follows directly from Corollary 9.17 after adding the simplifying hypotheses. The second inequality follows from straightforward manipulation of the inequality

$$\sqrt{2}^{\frac{\pi - \Theta_p(\omega_\mathrm{b})}{\Theta_p(\omega_\mathrm{b})}} \leq T_{\max}.$$

■

The upshot of the corollary is that if one wishes to make the lower bound on the H_∞-form of T_L smaller than $\sqrt{2}$, then the closed-loop bandwidth will exceed the location of the real, unstable pole p.

Often in applications one wishes to impose the conditions (9.5) and (9.6) together, with $\omega_2 > \omega_1$, and noting that (9.6) implies that $|S_L(i\omega)| < \frac{1}{\epsilon_2}$ for $|\omega| > \omega_2$. The picture is shown in Figure 9.6: one wishes to design the sensitivity function so that it remains below

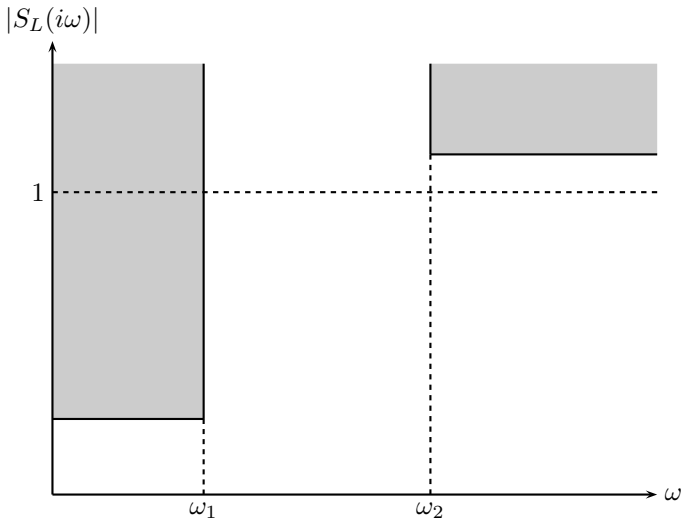


Figure 9.6 Combined sensitivity and complementary sensitivity function restrictions

the shaded area. Thus one will want to minimise the sensitivity function over a range of frequencies, as well as attenuate the complementary sensitivity function for high frequencies. Assuming that $z \in \mathbb{C}_+$ is a zero for R_L, computations like those used to prove Corollary 9.16 then give

$$\|S_L\|_\infty \geq \left(\frac{1}{\epsilon_1}\right)^{\frac{\Theta_z(\omega_1)}{\Theta_z(\omega_2)-\Theta_z(\omega_1)}} \left(\frac{1}{1+\epsilon_2}\right)^{\frac{\pi-\Theta_z(\omega_2)}{\Theta_z(\omega_2)-\Theta_z(\omega_1)}} |B_{S_L}^{-1}|^{\frac{\pi}{\Theta_z(\omega_2)-\Theta_z(\omega_1)}}. \tag{9.7}$$

The implications of this lower bound are examined in a simple case in Exercise E9.4.

9.3 The robust performance problem

So-called "H_∞ design" has received increasing attention in the recent control literature. The basic methodology has its basis in the frequency response ideas outlined in Section 8.5.2 and Section 9.2. The idea is that, motivated by Proposition 8.24, one wishes to minimise $\|S_L\|_\infty$. However, as we saw in Section 9.2, this is not an entirely straightforward task. Indeed, certain plant features—unstable poles and/or nonminimum phase zeros—can make this task attain a subtle nature. In this section we look at some ways of getting around these difficulties to specify realistic performance objectives. The approach makes contact with the topic of robust stability of Section 7.3, and enables us to state a control design problem, the so-called "robust performance problem." The resulting problem, as we shall see, takes the form of a minimisation problem, and its solution is the subject of Chapter 15.

We shall continue in this section to use the unity gain feedback loop of Figure 6.25 that we reproduce in Figure 9.7.

9.3.1 Performance objectives in terms of sensitivity and transfer functions

If you ever thought that it was possible to arbitrarily assign performance specifications to a system, it is hoped that the content of Sections 9.1 and 9.2 have made you think differently. But perhaps by now one is frightened into thinking that there is simply no way to specify performance objectives that are achievable. This is not true, of course. But what is true is

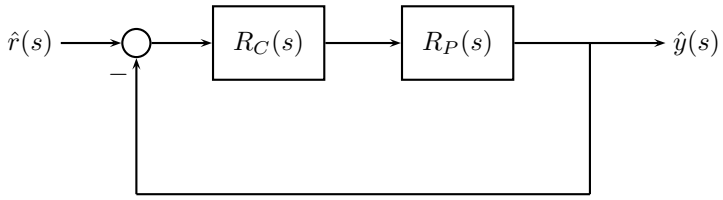


Figure 9.7 Unity gain feedback loop for robust performance problem

that one needs to be somewhat cagey about how these specifications are stated.

Let us begin by reinforcing our results about the difficulty facing us by trying something naïve. According to Proposition 8.24, if we wish to minimise the energy of the error, we should minimise the H_∞-norm of S_L. Thus, we could try to specify $\epsilon > 0$ and then set out to design a controller with the property that $\|S_L\|_\infty \leq \epsilon$. This objective may be impossible for many reasons, and let us list some of these.

9.19 Reasons why naïve sensitivity minimisation fails

1. If R_P is strictly proper, and we wish to design a proper controller R_C, then the loop gain $R_L = R_C R_P$ is strictly proper. Thus T_L is also strictly proper. Therefore, $\lim_{\omega\to\infty}|T_L(i\omega)| = 0$ from which it follows (from the fact that $S_L + T_L = 1$) that $\lim_{\omega\to\infty}|S_L(i\omega)| = 1$. Thus, in such a setting, we cannot possibly specify a H_∞-norm bound for S_L which is less than 1.
2. Based upon Proposition 8.22 we have two situations that provide limitations on possible sensitivity reduction.
   (a) If R_P and R_C have no poles in $\overline{\mathbb{C}}_+$ and if either have a zero in $\mathbb{C}_+$, then R_L is analytic in $\mathbb{C}_+$ and has a zero $z \in \mathbb{C}_+$. From Proposition 8.22 we have $S_L(z) = 1$. Thus it follows from the Maximum Modulus Principle (Theorem D.9) that $|S_L(i\omega)| \geq 1$ for some $\omega \in \mathbb{R}$.
   (b) This is much like the previous situation, except we reflect things about the imaginary axis. Thus, if R_P and R_C have no poles in $\overline{\mathbb{C}}_-$ and if either have a zero in $\mathbb{C}_-$, then R_L is analytic in $\mathbb{C}_-$ and has a zero $z \in \mathbb{C}_-$. Arguing as above, we see that $|S_L(i\omega)| \geq 1$ for some $\omega \in \mathbb{R}$.
3. If the relative degree of $R_L = R_C R_P$ exceeds 1 and if R_L is strictly proper, then, from Theorem 9.7 we know that $\|S_L\|_\infty > 1$. Furthermore, this same result tells us that this effect is exacerbated by R_L having unstable poles.
4. If R_L is nonminimum phase and if we minimise $|S_L|$ over a certain range of frequencies, then the result will be an increase in $\|S_L\|_\infty$.
5. The previous point is reiterated in the various corollaries to the Poisson integral formulae given in Theorem 9.13. Here the location of nonminimum phase zeros is explicitly seen in the estimate for the H_∞-norm for the sensitivity function. •

The above, apart from providing a neat summary of some of the material in Section 9.2, indicates that the objective of minimising $|S_L(i\omega)|$ over the *entire* frequency range is perhaps not a good objective. What's more, from a practical point of view, it is an excessively stringent condition. Indeed, remember why we want to minimise the sensitivity function: to

reduce the L_2-norm of the error to a given input signal. However, in a given application, one will often have some knowledge about the nature of the input signals one will have to deal with. This knowledge may come in the form of, or possibly be converted into the form of, frequency response information. Let us give some examples of how such a situation may arise.

9.20 Examples

1. Suppose that we are interested in tracking sinusoids with frequencies in a certain range with minimal error. Thus we take a class of nominal signals to be sinusoids of the form

$$\{r(t) = a \sin \omega t \mid \ 0 \leq a \leq 1\}.$$

Now we bias a subset of these inputs as follows:

$$\mathrm{L}_{\mathrm{ref}} = \{|W_p(i\omega)a| \sin \omega t \ \mid \ 0 \leq a \leq 1\},$$

where $W_p \in \mathbb{R}(s)$ is a function for which the magnitude of restriction to the imaginary axis captures the behaviour we wish be giving more weight to certain frequencies. With this set of inputs, the maximum error is

$$\begin{aligned}\sup_{r \in \mathrm{L}_{\mathrm{ref}}} \|e\|_\infty = & \sup_\omega |W_p(i\omega)S_L(i\omega)| \\ = & \|W_p S_L\|_\infty.\end{aligned}$$

Note here that the measure of performance we use is the L_∞-norm of the error. In the following two examples, different measures will be used.

2. One can also think of specifying the character of reference signals by providing bounds on the H_2-norm of the signal. Thus we could think of nominal signals as being those of the form

$$\{r_{\mathrm{nom}} \colon [0, \infty) \to \mathbb{R} \mid \ \|\hat{r}_{\mathrm{nom}}\|_2 \leq 1\},$$

and these are shaped to a set of possible reference signals

$$\mathrm{L}_{\mathrm{ref}} = \{r \colon [0, \infty) \to \mathbb{R} \mid \ \hat{r} = W_p r_{\mathrm{nom}}, \ \|\hat{r}_{\mathrm{nom}}\|_2 \leq 1\},$$

where W_p is a specified transfer function that shapes the energy spectrum of the signal to a desired shape. Simple manipulation then shows that

$$\mathrm{L}_{\mathrm{ref}} = \left\{ r \colon [0, \infty) \to \mathbb{R} \ \middle| \ \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \frac{\hat{r}(i\omega)}{W_p(i\omega)} \right|^2 \mathrm{d}\omega \leq 1 \right\}.$$

If we wish to minimise the L_2-norm of the error, then the maximum error is given by

$$\begin{aligned}\sup_{r \in \mathrm{L}_{\mathrm{ref}}} \|e\|_2 = & \sup\{\|W_p S_L r_{\mathrm{nom}}\|_2 \mid \ r_{\mathrm{nom}} \leq 1\} \\ = & \|W_p S_L\|_\infty,\end{aligned}$$

where we have used part (i) of Theorem 5.21. Again, we have arrived at a specification of the form of $\|W_p S_L\|_\infty < \epsilon$.

3. The above argument can be carried out for nominal reference signals

$$\{r_{\mathrm{nom}} \colon [0, \infty) \to \mathbb{R} \mid \ \mathrm{pow}(\hat{r}_{\mathrm{nom}}) \leq 1\}.$$

Similar arguments, now asking that the power spectrum of the error be minimised, lead in the same way (using part (ix) of Theorem 5.21) to a condition of the form $\|W_p S_L\|_\infty < \epsilon$.

4. Suppose that one knows from past experience that a controller will perform well if the frequency response of the sensitivity function lies below that of a given transfer function. Thus one would have a specification like

$$|S_L(i\omega)| \leq |R(i\omega)|, \quad \omega > 0, \tag{9.8}$$

where $R \in \mathbb{R}(s)$ comes from somewhere or other. In this case, if we define $W_p(s) = R(s)^{-1}$, this turns (9.8) into a condition of the form $\|W_p T_L\|_\infty < 1$. •

Note that all specifications like $\|W_p S_L\|_\infty < \epsilon$ are by simple scaling transfered to a condition of the form $\|W_p S_L\|_\infty < 1$. This is the usual form for these conditions to take, in practice.

The above examples present, in sort of general terms, possible natural ways in which one can arrive at performance criterion of the form $\|W_p S_L\|_\infty < 1$, for some $W_p \in \mathbb{R}(s)$. In making such specifications, one in only interested in the magnitude of W_p on the imaginary axis. Therefore one may as well suppose that W_p has no poles or zeros in $\mathbb{C}_+$.[1] In this book, we shall alway deal with specifications that come in the form $\|W_p S_L\|_\infty < 1$. Note also that one might use the other transfer functions

$$T_L = \frac{R_C R_P}{1 + R_C R_P}, \quad R_C S_L = \frac{R_C}{1 + R_C R_P}, \quad R_P S_L = \frac{R_P}{1 + R_C R_P}$$

for performance specifications in the case when the loop gain R_L is the product of R_C with R_P. There is nothing in principle stopping us from using these other transfer functions; our choice is motivated by a bald-faced demand for simplicity. Below we shall formulate criterion that use our performance conditions of this section, and these results are complicated if one uses the any of the other transfer functions in place of the sensitivity function.

There is a readily made graphical interpretation of the condition $\|W_p S_L\|_\infty < 1$, mirroring the pictures in Figures 7.21 and 7.25. To see how this goes, note the following:

$$\begin{aligned}
& \|W_p S_L\|_\infty < 1 \\
\iff \quad & |W_p(i\omega) S_L(i\omega)| < 1, \quad \omega \in \mathbb{R} \\
\iff \quad & \Big|\frac{W_p(i\omega)}{1 + R_L(i\omega)}\Big| < 1, \quad \omega \in \mathbb{R} \\
\iff \quad & |W_p(i\omega)| < |1 + R_L(i\omega)|, \quad \omega \in \mathbb{R} \\
\iff \quad & |W_p(i\omega)| < |-1 - R_L(i\omega)|, \quad \omega \in \mathbb{R}.
\end{aligned}$$

Now note that $|-1 - R_L(i\omega)|$ is the distance away from the point $-1 + i0$ of the point $R_L(i\omega)$ on the Nyquist contour for R_L. Thus the interpretation we make is that the Nyquist contour at frequency ω remain outside the closed disk centred at $-1 + i0$ of radius $|W_p(i\omega)|$. This is depicted in Figure 9.8. Note that we could just as well have described the condition $\|W_p S_L\|_\infty < 1$ just as we did in Figures 7.21 and 7.25 by saying that the circle of radius $|W_p(i\omega)|$ and centre $R_L(i\omega)$ does not contain the point $-1 + i0$. However, the interpretation we have given has greater utility, at least as we shall use it. Also, note that one can make similar interpretations using any of the other performance conditions $|W_p T_L|_\infty, |W_p R_C S_L|_\infty, |W_p R_P S_L|_\infty < 1$.

[1] If W_p has, say, a zero $z \in \mathbb{C}_+$, then we can write $W_p = (s - z)^k W(s)$ where $W(z) \neq 0$. One can then easily verify that the new weight $\tilde{W}_p(s) = (s + z)^k W(s)$ has the same magnitude as W_p on the imaginary axis. Doing this for all poles and zeros, we see that we can produce a function with no poles or zeros in $\mathbb{C}_+$ that has the same magnitude on $i\mathbb{R}$.

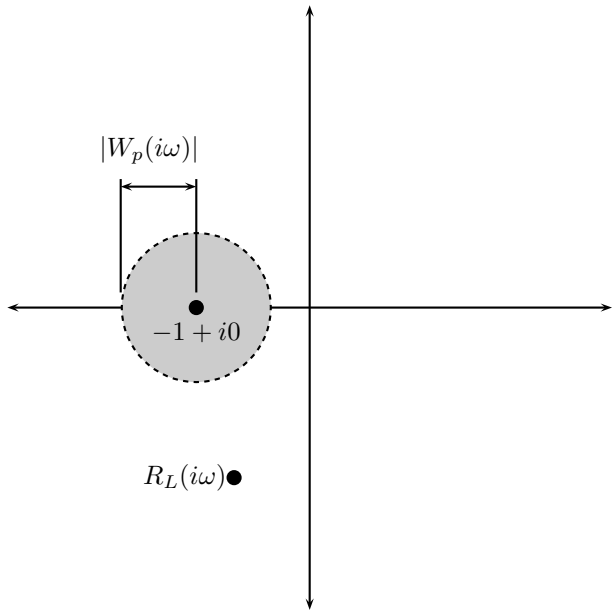


Figure 9.8 Interpretation of weighted performance condition

9.3.2 Nominal and robust performance

Now that we have a method for producing frequency domain performance specifications that we can hope are manageable, let us formulate some problems based upon combining this strategy with the uncertainty models of Section 4.5 and the notions of robust stability of Section 7.3. The situation is roughly this: we have a set of plants $\mathscr{P}$ and a performance weighting function W_p that gives a performance specification $\|W_pS_L\|_\infty < 1$. We wish to examine questions dealing with stabilising all plants in $\mathscr{P}$ while also meeting the performance criterion.

The following definition makes precise the forms of stability possible in the framework just described.

9.21 Definition Let $\bar{R}_P \in \mathbb{R}(s)$ be proper and suppose that $R_C \in \mathbb{R}(s)$ renders IBIBO stable the interconnection of Figure 9.7. Let $\bar{S}_L$ and $\bar{T}_L$ denote the corresponding sensitivity and closed-loop transfer functions. Let $W_u \in \mathrm{RH}_\infty^+$ and recall the definitions of $\mathscr{P}_\times(\bar{R}_P, W_u)$ and $\mathscr{P}_+(\bar{R}_P, W_u)$ and for $R_P \in \mathscr{P}_\times(\bar{R}_P, W_u)$ or $R_P \in \mathscr{P}_+(\bar{R}_P, W_u)$ denote by $S_L(R_P)$ and $T_L(R_P)$ the corresponding sensitivity and closed-loop transfer functions, respectively. Let $W_p \in \mathbb{R}(s)$ have no poles or zeros in $\mathbb{C}_+$.

(i) R_C provides ***nominal performance*** for $\mathscr{P}_\times(\bar{R}_P, W_u)$ (resp. $\mathscr{P}_+(\bar{R}_P, W_u)$) relative to W_p if
   (a) R_C provides robust stability for $\mathscr{P}_\times(\bar{R}_P, W_u)$ (resp. $\mathscr{P}_+(\bar{R}_P, W_u)$) and
   (b) $\|W_p\bar{S}_L\|_\infty < 1$.

(ii) R_C provides ***robust performance*** for $\mathscr{P}_\times(\bar{R}_P, W_u)$ (resp. $\mathscr{P}_+(\bar{R}_P, W_u)$) relative to W_p if
   (a) R_C provides robust stability for $\mathscr{P}_\times(\bar{R}_P, W_u)$ (resp. $\mathscr{P}_+(\bar{R}_P, W_u)$) and
   (b) $\|W_pS_L(R_P)\|_\infty < 1$ for every $R_P \in \mathscr{P}_\times(\bar{R}_L, W_u)$. •

Thus by nominal performance we mean that the controller provides robust stability, and meets the performance criterion for the nominal plant. However, robust performance requires that we have robust stability and that the performance criterion is met for *all* plants in the uncertainty set. Thus nominal performance consists of the two conditions

$$\|W_u\bar{T}_L\|_\infty < 1, \quad \|W_p\bar{S}_L\|_\infty < 1$$

for multiplicative uncertainty, and

$$\|W_uR_C\bar{S}_L\|_\infty < 1, \quad \|W_p\bar{S}_L\|_\infty < 1$$

for additive uncertainty. To interpret robust performance, note that for multiplicative perturbations we have

$$\begin{aligned} S_L(R_P) &= \frac{1}{1+R_CR_P} \\ &= \frac{1}{1+R_C(1+\Delta)\bar{R}_P} \\ &= \frac{\frac{1}{1+R_C\bar{R}_P}}{1+\Delta\frac{R_C\bar{R}_P}{1+R_C\bar{R}_P}} \\ &= \frac{\bar{S}_L}{1+\Delta W_u\bar{T}_L}, \end{aligned} \tag{9.9}$$

where Δ is the allowable perturbation giving $R_P \in \mathscr{P}_\times(\bar{R}_P, W_u)$. A similarly style calculation for additive perturbations gives

$$S_L(R_P) = \frac{\bar{S}_L}{1+\Delta W_uR_C\bar{S}_L}, \tag{9.10}$$

where Δ is the allowable perturbation giving $R_P \in \mathscr{P}_+(\bar{R}_P, W_u)$.

To state a theorem on robust performance we need some notation. For $R_1, R_2 \in \mathbb{R}(s)$ we define a $\mathbb{R}$-valued function of s by

$$s \mapsto |R_1(s)| + |R_2(s)|.$$

We denote this function by $|R_1| + |R_2|$, making a slight, but convenient, abuse of notation. Although this function is not actually in the class of functions for which we defined the H_∞-norm, we may still denote

$$\big\||R_1| + |R_2|\big\|_\infty = \sup_{\omega\in\mathbb{R}}\big(|R_1(i\omega)| + |R_2(i\omega)|\big).$$

This notation and the calculations of the preceding paragraph are useful in stating and proving the following theorem.

9.22 Theorem *Let $\bar{R}_P \in \mathbb{R}(s)$ be proper and suppose that $R_C \in \mathbb{R}(s)$ renders IBIBO stable the interconnection of Figure 9.7. Let $\bar{S}_L$ and $\bar{T}_L$ denote the corresponding sensitivity and closed-loop transfer functions. Let $W_u \in \mathrm{RH}_\infty^+$ and recall the definitions of $\mathscr{P}_\times(\bar{R}_P, W_u)$ and $\mathscr{P}_+(\bar{R}_P, W_u)$ and for $R_P \in \mathscr{P}_\times(\bar{R}_P, W_u)$ or $R_P \in \mathscr{P}_+(\bar{R}_P, W_u)$ denote by $S_L(R_P)$ and $T_L(R_P)$ the corresponding sensitivity and closed-loop transfer functions, respectively. Let $W_p \in \mathbb{R}(s)$ have no poles or zeros in $\mathbb{C}_+$. The following statements hold:*

(i) R_C *provides robust performance for* $\mathscr{P}_\times(\bar{R}_P, W_u)$ *relative to* W_p *if and only if* $\big\||W_u\bar{T}_L| + |W_p\bar{S}_L|\big\|_\infty < 1$*;*

(ii) R_C *provides robust performance for* $\mathscr{P}_+(\bar{R}_P, W_u)$ *relative to* W_p *if and only if* $\big\||W_uR_C\bar{S}_L| + |W_p\bar{S}_L|\big\|_\infty < 1$*;*

Proof (i) First we make a use equivalent formulation of the condition $\big\||W_u\bar{T}_L|+|W_p\bar{S}_L|\big\|_\infty < 1$. We compute

$$\begin{aligned}
&\big\||W_u\bar{T}_L| + |W_p\bar{S}_L|\big\|_\infty < 1 \\
\iff\quad &\big(|W_u(i\omega)\bar{T}_L(i\omega)| + |W_p(i\omega)\bar{S}_L(i\omega)|\big) < 1, \quad \omega \in \mathbb{R} \\
\iff\quad &|W_p(i\omega)\bar{S}_L(i\omega)| < 1 - |W_u(i\omega)\bar{T}_L(i\omega)|, \quad \omega \in \mathbb{R} \\
\iff\quad &|W_u(i\omega)\bar{T}_L(i\omega)| < 1, \quad \omega \in \mathbb{R}, \quad \Big|\frac{W_p(i\omega)\bar{S}_L(i\omega)}{1 - |W_u(i\omega)\bar{T}_L(i\omega)|}\Big|, \quad \omega \in \mathbb{R} \\
\iff\quad &\|W_u\bar{T}_L\|_\infty < 1, \qquad \Big\|\frac{W_p\bar{S}_L}{1 - |W_u\bar{T}_L|}\Big\|_\infty < 1. \qquad (9.11)
\end{aligned}$$

Now suppose that $\big\||W_u\bar{T}_L| + |W_p\bar{S}_L|\big\|_\infty < 1$. For any allowable Δ we then have

$$\begin{aligned}
1 &= |1 + \Delta(i\omega)W_u(i\omega)\bar{T}_L(i\omega) - \Delta(i\omega)\bar{T}_L(i\omega)| \\
&\leq |1 + \Delta(i\omega)W_u(i\omega)\bar{T}_L(i\omega)| + |W_u(i\omega)\bar{R}_L(i\omega)|,
\end{aligned}$$

for all $\omega \in \mathbb{R}$. Thus we conclude that

$$1 - |W_u(i\omega)\bar{T}_L(i\omega)| \leq |1 + \Delta(i\omega)W_u(i\omega)\bar{T}_L(i\omega)|, \quad \omega \in \mathbb{R}.$$

From this it follows that

$$\Big\|\frac{W_p\bar{S}_L}{1 - |W_u\bar{T}_L|}\Big\|_\infty \geq \Big\|\frac{W_p\bar{S}_L}{1 + \Delta W_u\bar{T}_L}\Big\|_\infty.$$

From (9.11) it now follows that

$$\Big\|\frac{W_p\bar{S}_L}{1 + \Delta W_u\bar{T}_L}\Big\|_\infty < 1,$$

and robust performance now follows from (9.9).

Now suppose that R_C provides robust performance for $\mathscr{P}_\times(\bar{R}_P, W_u)$ relative to W_p. From (9.9) and Theorem 7.18 it follows that

$$\|W_u\bar{T}_L\|_\infty < 1, \quad \Big\|\frac{W_p\bar{S}_L}{1 + \Delta W_u\bar{T}_L}\Big\|_\infty < 1$$

for all allowable Δ. Let $\bar{\omega} \geq 0$ be a frequency for which the function

$$\omega \mapsto \frac{|W_p(i\omega)\bar{S}_L(i\omega)|}{1 - |W_u(i\omega)\bar{T}_L(i\omega)|}$$

is maximum. Let $\theta = \pi - \measuredangle W_u(i\bar{\omega})\bar{T}_L(i\bar{\omega})$. Using Lemma 1 from the proof of Theorem 7.18, let $G_\theta \in \mathrm{RH}_\infty^+$ have the property that $\measuredangle G_\theta(i\bar{\omega}) = \theta$ and $|G_\theta(i\bar{\omega})| = 1$. Then $\Delta = G_\theta$ is

allowable and the quantity $\Delta(i\bar{\omega})W_u(i\bar{\omega})\bar{T}_L(i\bar{\omega})$ is real and negative. Therefore, with this allowable Δ, we have

$$1 - |W_u(i\bar{\omega})\bar{T}_L(i\bar{\omega})| = |1 + \Delta(i\bar{\omega})W_u(i\bar{\omega})\bar{T}_L(i\bar{\omega})|.$$

It now follows that for the Δ just defined,

$$\begin{aligned}\Big\|\frac{W_p\bar{S}_L}{1-|W_u\bar{T}_L|}\Big\|_\infty &= \frac{|W_p(i\bar{\omega})\bar{S}_L(i\bar{\omega})|}{1-|W_u(i\bar{\omega})\bar{T}_L(i\bar{\omega})|}\\ &= \frac{|W_p(i\bar{\omega})\bar{S}_L(i\bar{\omega})|}{|1+\Delta(i\bar{\omega})W_u(i\bar{\omega})\bar{T}_L(i\bar{\omega})|}\\ &\leq \Big\|\frac{W_p\bar{S}_L}{1+\Delta W_u\bar{T}_L}\Big\|_\infty\end{aligned}$$

Thus we have shown that

$$\Big\|\frac{W_p\bar{S}_L}{1-|W_u\bar{T}_L|}\Big\|_\infty < 1.$$

Along with our hypothesis that $\|W_u\bar{T}_L\| < 1$, from (9.11) it now follows that $\big\||W_u\bar{T}_L| + |W_p\bar{S}_L|\big\|_\infty < 1$.

(ii) Since the idea here is in spirit identical to that of part (i), we are allowed to be a little sketchy. In fact, the key is the following computation:

$$\begin{aligned}&\big\||W_uR_C\bar{S}_L| + |W_p\bar{S}_L|\big\|_\infty < 1\\ \iff\quad &\big(|W_u(i\omega)R_C(i\omega)\bar{S}_L(i\omega)| + |W_p(i\omega)\bar{S}_L(i\omega)|\big) < 1, \quad \omega\in\mathbb{R}\\ \iff\quad &|W_p(i\omega)\bar{S}_L(i\omega)| < 1 - |W_u(i\omega)R_C(i\omega)\bar{S}_L(i\omega)|, \quad \omega\in\mathbb{R}\\ \iff\quad &|W_u(i\omega)R_C(i\omega)\bar{S}_L(i\omega)| < 1, \quad \omega\in\mathbb{R}, \quad \Big|\frac{W_p(i\omega)\bar{S}_L(i\omega)}{1-|W_u(i\omega)R_C(i\omega)\bar{S}_L(i\omega)|}\Big|, \quad \omega\in\mathbb{R}\\ \iff\quad &\|W_uR_C\bar{S}_L\|_\infty < 1, \qquad \Big\|\frac{W_p\bar{S}_L}{1-|W_uR_C\bar{S}_L|}\Big\|_\infty < 1. \end{aligned} \tag{9.12}$$

First assume that $\big\||W_uR_C\bar{S}_L| + |W_p\bar{S}_L|\big\|_\infty < 1$. Now we readily compute

$$1 - |W_u(i\omega)R_C(i\omega)\bar{S}_L(i\omega)| \leq |1 + \Delta(i\omega)W_u(i\omega)R_C(i\omega)\bar{S}_L(i\omega)|$$

for all $\omega\in\mathbb{R}$ from which it follows that

$$1 > \Big\|\frac{W_p\bar{S}_L}{1-|W_uR_C\bar{S}_L|}\Big\|_\infty \geq \Big\|\frac{W_p\bar{S}_L}{1+\Delta W_uR_C\bar{S}_L}\Big\|_\infty.$$

Robust performance now follows from (9.10).

Now suppose that R_C provides robust performance for $\mathscr{P}_+(\bar{R}_P, W_u)$ relative to W_p. By (9.10) and Theorem 7.21 this means that

$$\|W_uR_C\bar{S}_L\|_\infty < 1, \quad \Big\|\frac{W_p\bar{S}_L}{1+\Delta W_uR_C\bar{S}_L}\Big\|_\infty < 1$$

for all allowable Δ. Now let $\bar{\omega}\geq 0$ be a frequency which maximises the function

$$\omega \mapsto \frac{|W_p(i\omega)\bar{S}_L(i\omega)|}{1-|W_u(i\omega)R_C(i\omega)\bar{S}_L(i\omega)|}.$$

Arguing as in part (i) we may find an allowable Δ so that

$$1 - |W_u(i\bar{\omega})R_C(i\omega)\bar{S}_L(i\bar{\omega})| = |1 + \Delta(i\bar{\omega})W_u(i\bar{\omega})R_C(i\omega)\bar{S}_L(i\bar{\omega})|.$$

Again arguing as in (i) it follows that

$$\left\| \frac{W_p\bar{S}_L}{1 - |W_uR_C\bar{S}_L|} \right\|_\infty < 1,$$

and from this it follows from (9.12) that $\big\| |W_uR_C\bar{S}_L| + |W_p\bar{S}_L| \big\|_\infty < 1$. ■

This is an important theorem in SISO robust control theory, and it forms the basis for many MIMO generalisations [see Dullerud and Paganini 1999]. It is useful because it gives a single H_∞-norm test for robust performance. For SISO systems this means that the condition can be tested by producing Bode plots, and this is something that is easily done. For MIMO systems, the matter of checking the conditions that generalise $\big\| |W_uR_C\bar{S}_L| + |W_p\bar{S}_L| \big\|_\infty < 1$ becomes a serious computational issue. In any event, the theorem allows us to state a precisely formulated design problem from that simultaneously incorporates stability, uncertainty, and performance.

9.23 Robust performance problem Given

(i) a nominal proper plant $\bar{R}_P$,

(ii) a function $W_u \in \mathrm{RH}_\infty^+$,

(iii) an uncertainty model $\mathscr{P}_\times(\bar{R}_P, W_u)$ or $\mathscr{P}_+(\bar{R}_P, W_u)$, and

(iv) a performance weight $W_p \in \mathbb{R}(s)$,

find a controller R_C that

(v) stabilises the nominal system and

(vi) satisfies either $\big\| |W_u\bar{T}_L| + |W_p\bar{S}_L| \big\|_\infty < 1$ or $\big\| |W_uR_C\bar{S}_L| + |W_p\bar{S}_L| \big\|_\infty < 1$, depending on whether one is using multiplicative or additive uncertainty. •

9.24 Remarks

1. The material leading up to the given statement in the robust performance problem has its basis in the robust stability results of Doyle and Stein [1981] and Chen and Desoer [1982], and seems to have its original statement in the book of Doyle, Francis, and Tannenbaum [1990].
2. It is possible that the robust performance problem can have no solution. Indeed, it is easy to come up with plant uncertainty models and performance weights that make the problem unsolvable. An example of how to do this is the subject of Exercise E9.8.
3. A graphical interpretation of the condition $\big\| |W_u\bar{T}_L| + |W_p\bar{S}_L| \big\|_\infty < 1$ (or $\big\| |W_uR_C\bar{S}_L| + |W_p\bar{S}_L| \big\|_\infty < 1$) is given in Figure 9.9. The picture says that for each frequency ω, the open disk of radius $|W_p(i\omega)|$ centred at $-1 + i0$ should not intersect the open disk of radius $|W_u(i\omega)\bar{R}_L(i\omega)|$ centred at $R_L(i\omega)$ (a similar statement holds, of course, for additive uncertainty). •

In Chapter 15 we shall provide a way to find a solution to a slightly modified version of the robust performance problem. In the stated form, it appears too difficult to admit a simple solution. However, for now we can content ourselves with a couple of examples that play with the problem in an *ad hoc* manner.

First we look at a case where we use multiplicative uncertainty.

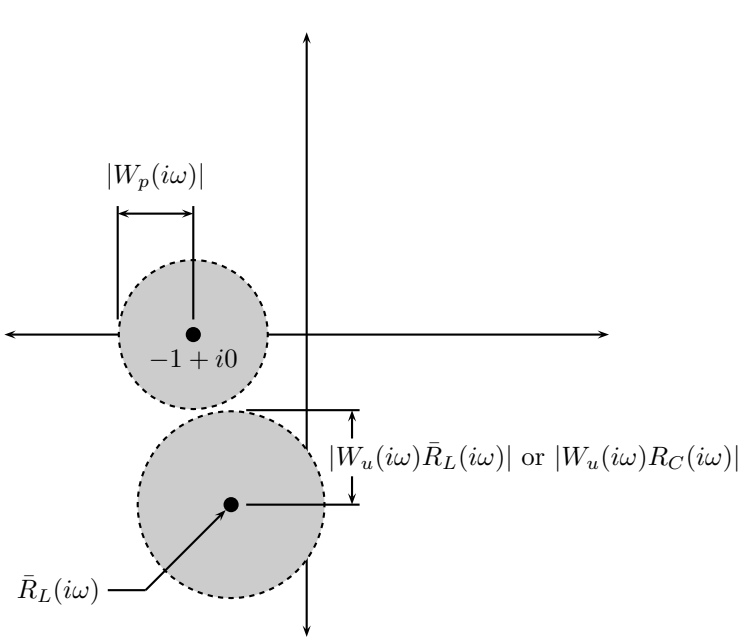


Figure 9.9 Graphical interpretation of robust performance condition

9.25 Example **(Example 7.20 cont'd)** Recall that we have taken

$$\bar{R}_P(s) = \frac{1}{s^2}$$

as our nominal plant and had used an uncertainty description of the form

$$W_u(s) = \frac{as}{s+1}, \quad a > 0.$$

In Example 7.20 we had concluded that provided that $a < a_{\max} \approx \frac{3}{4}$, the controller

$$R_C(s) = 1 + 2s + \frac{1}{s}$$

robustly stabilises $\mathscr{P}_\times(\bar{R}_P, W_u)$. To make this into a robust performance problem, we need to provide a performance weight W_p. Let us suppose that we know that the reference signals will have low energy below 10rad/ sec. Given our discussion in Example 9.20, one might say that taking

$$W_p(s) = \frac{\frac{1}{2}}{\frac{s}{10}+1}$$

is a good choice, so let us go with this. Our objective will be to decide upon the maximum value of a so that the associated robust performance problem has a solution. According to Theorem 9.22 we should choose $a > 0$ so that $\big\||W_u\bar{T}_L| + |W_p\bar{S}_L|\big\|_\infty < 1$. In Figure 9.10 we show the magnitude Bode plot for $|W_u\bar{T}_L| + |W_p\bar{S}_L|$ when $a = 1$. The peak magnitude is about 4dB. Thus we need to reduce a. However, not like the case for robust stability, the quantity $|W_u\bar{T}_L| + |W_p\bar{S}_L|$ is not linear in a so we cannot simply use a naïve scaling argument to determine a. The easiest way to proceed is by trial and error, reducing a until the magnitude Bode plot for $|W_u\bar{T}_L| + |W_p\bar{S}_L|$ dips below 0dB. Doing this trial and error gives $a_{\max} \approx \frac{1}{2}$. The magnitude Bode plot for $|W_u\bar{T}_L| + |W_p\bar{S}_L|$ when $a = \frac{2}{3}$ is also shown

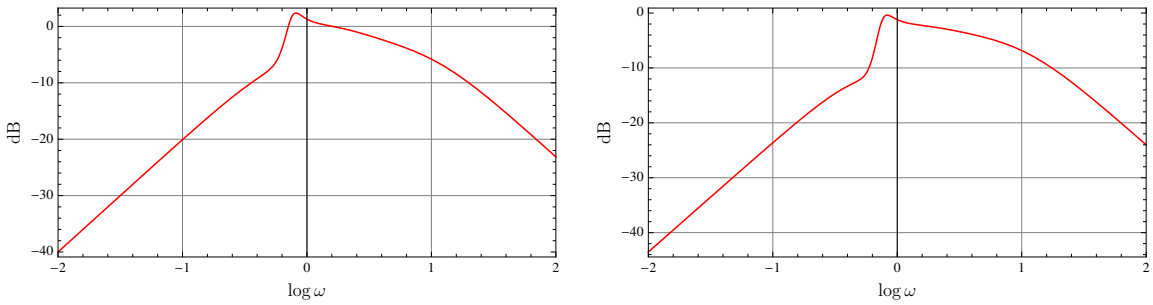


Figure 9.10 Bode plot of $|W_u\bar{T}_L| + |W_p\bar{S}_L|$ for $a = 1$ (left) and for $a = \frac{2}{3}$ (right)

in Figure 9.10, and one can see that it satisfies the robust performance constraint. Note that not surprisingly the maximum allowable value for a is smaller than was the case in Example 7.20 when we were merely looking to attain robust stability. The demand that our controller also meet the performance specifications places upon it further restrictions. In the current situation, if one wishes to allow greater variation in the set of plants contained in the uncertainty description, one might look into backing off on the performance demands. •

Next, we give an example along similar lines that uses additive uncertainty.

9.26 Example **(Example 7.23 cont'd)** We proceed along the lines of the previous example, now carrying on from Example 7.23 where we used the nominal plant and controller

$$\bar{R}_P(s) = \frac{1}{s^2}, \quad R_C(s) = 1 + 2s + \frac{1}{s}.$$

To model plant uncertainty we use

$$W_u(s) = \frac{as}{(s+1)^2},$$

and we use the same performance weight W_p as in Example 9.25. With this data we give the magnitude Bode plot for $|W_uR_C\bar{S}_L| + |W_p\bar{S}_L|$ in Figure 9.11 for the case when $a = 1$.

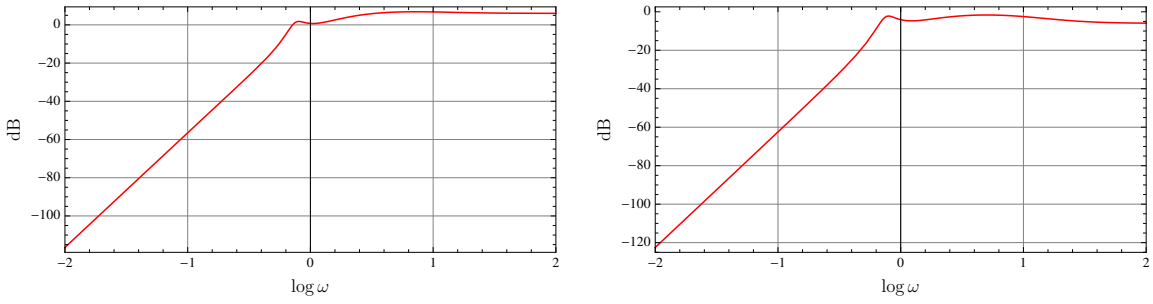


Figure 9.11 Bode plot of $|W_uR_C\bar{S}_L| + |W_p\bar{S}_L|$ for $a = 1$ (left) and for $a = \frac{1}{4}$ (right)

Again, the easiest way to determine the upper bound on a is by trial and error. Doing so gives $a_{\max} \approx \frac{1}{4}$. In Figure 9.11 is shown the Bode plot of $|W_uR_C\bar{S}_L| + |W_p\bar{S}_L|$ for this value of a, and one can see that it satisfies the bounds for robust performance. •

The preceding example illustrate that the matter of *checking* a controller for robust performance is largely a Bode plot issue. This is one thing that makes Theorem 9.22 a valuable result in this day of the easily fabricated Bode plot.

9.4 Summary

When designing a controller, the first step is typically to determine acceptable performance criterion. In this chapter, we have come up with a variety of performance measures. Let us review what we have said.

1.

Exercises

E9.1 Consider the pendulum/cart of Exercises E1.5 and E2.4. In this exercise, you will use the cart position as the output, and you will take as the equilibrium position the upright position for the pendulum.

(a) Compute the transfer function for the linearised system, and show that it has both an unstable pole and a nonminimum phase zero.

(b) Choose parameter values for the system, then design a feedback vector $\boldsymbol{f}$ that makes the closed-loop system IBIBO stable.

(c) For the closed-loop system, use Proposition 3.40 to compute the step response. Verify that the bounds of Proposition 9.6 are satisfied.

E9.2 Poisson integral formulae for the pendulum on a cart.

Vidyasagar [1986] proposes a definition of undershoot different from the one we use, his definition being as follows. Let (N, D) be a BIBO stable, strictly proper, steppable SISO linear system in input/output form and let $r > 0$ be the smallest integer for which $1^{(r)}_{N,D}(0) \neq 0$. Note that $\lim_{t\to\infty} 1_{N,D}(t) = T_{N,D}(0)$. The system (N, D) exhibits ***immediate undershoot*** if $1^{(r)}_{N,D}(0)T_{N,D}(0) < 0$. Thus, the system exhibits immediate undershoot when the initial response heads off in a different direction that is attained by the steady-state response. In the next exercise, you will explore this alternative definition of undershoot.

E9.3 Let (N, D) be a BIBO stable strictly proper, steppable SISO linear system in input/output form and let $r > 0$ be the smallest integer for which $1^{(r)}_{N,D}(0) \neq 0$.

(a) Prove the following theorem of Vidyasagar [1986].

Theorem *(N, D) exhibits immediate undershoot (in the sense of the above definition) if and only if N has an odd number of positive real roots.*

Hint: *Factor the numerator and denominator into irreducible factors, and use the fact that*

$$1^{(r)}_{N,D}(0) = \lim_{s\to\infty} s^r T_{N,D}(s).$$

Now consider the system

$$(N(s), D(s)) = (s^2 - 2s + 1, s^3 + 2s^2 + 2s + 2).$$

For this system, answer the following questions.

(b) Verify that this system is BIBO stable.

(c) According to the theorem of part (a), will the system exhibit immediate undershoot?

(d) Produce the step response for the system. Does it exhibit immediate undershoot according to the definition of Vidyasagar? Would you say it exhibits undershoot?

E9.4 Let R_L be a proper BIBO stable loop gain with the property that the standard unity gain feedback loop (e.g., Figure 9.3) is IBIBO stable. Also suppose that R_L has a single real zero $z \in \mathbb{C}_+$. We wish to have the sensitivity function fit under the shaded area in Figure 9.6, and we take $\omega_1 = \frac{3}{4}\omega_2$. We also define ϵ_2 so that ω_2 is the closed-loop bandwidth: $\epsilon_2 = \frac{1}{\sqrt{2}}$.

(a) For various values of ϵ_1, plot the lower bound for $\|S_L\|_\infty$ given by (9.7) as a function of $\frac{z}{\omega_b}$.

(b) Would you expect better performance for the closed-loop system for larger of smaller ratios $\frac{z}{\omega_b}$?

The following exercise should be done after you have designed a state feedback vector for the pendulum/cart system. An arbitrary such state feedback vector is constructed in Exercise E10.11, and an optimal state feedback vector is determined in Exercise E14.7. For the following exercise, you may use the parameter values, and the state feedback vector from either of those exercises.

E9.5 Consider the pendulum/cart system of Exercises E1.5 and E2.4, and let $\boldsymbol{f} \in \mathbb{R}^4$ be a stabilising state feedback vector, and consider the loop gain $R_{\boldsymbol{f}}$ as defined in Exercise E7.11.

(a) For this loop gain, produce the magnitude Bode plot of the corresponding sensitivity and complementary sensitivity functions.

(b) Verify that the bounds of Corollaries 9.16 and 9.17 are satisfied, as well as that of (9.7).

The next two exercises give conditions for robust performance for the uncertainty description presented in Section 4.5, but that we have not discussed in detail in the text. The conditions for robust stability for these descriptions you derived in Exercises E7.12 and E7.13. For these plant uncertainty models, it turns out to be convenient to give the performance specifications not on the sensitivity function S_L, but on the closed-loop transfer function T_L.

Thus the performance criterion for the following two exercises should take the form $\|W_p T_L\|_\infty < 1$.

With this as backdrop, you may readily adapt the proof of Theorem 9.22 to prove the conditions for robust performance in the next two exercises.

E9.6 For the plant uncertainty description of Exercise E7.12, and the performance criterion $\|W_p T_L\|_\infty < 1$ (see above), show that a controller R_C stabilising the nominal plant $\bar{R}_P$ provides robust performance if and only if $\big\||W_u \bar{R}_P \bar{S}_L| + |W_p \bar{T}_L|\big\|_\infty < 1$.

E9.7 For the plant uncertainty description of Exercise E7.13, and the performance criterion $\|W_p T_L\|_\infty < 1$ (see above), show that a controller R_C stabilising the nominal plant $\bar{R}_P$ provides robust performance if and only if $\big\||W_u \bar{S}_L| + |W_p \bar{T}_L|\big\|_\infty < 1$.

E9.8 Use the fact that $\bar{S}_L + \bar{T}_L = 1$ to show that in order for the robust performance problem to have a solution, it must be the case that

$$\min\{W_p(i\omega), W_u(i\omega)\} < 1, \quad \omega \in \mathbb{R}.$$

Part III

Controller design

Chapter 10

Stabilisation and state estimation

While Chapter 5 dealt with various types of stability, and Chapter 6 provided a general setting, with some specialisations in the later sections, for feedback, in this chapter we combine feedback and stability to get stabilisation. The idea is quite simple: one wishes to consider feedback that leaves a closed-loop system stable, or perhaps stabilises an unstable system. In this chapter we also touch upon the matter of state estimation. The need for this arises in practice where one can only measure outputs, and not the entire state. Therefore, if one wishes to design feedback laws using the state of the system, it is necessary to reconstruct the state from the output.

This is our first chapter concerned with controller design. As such, the design issue with that we are concerned is merely stability. Design for performance is dealt with in later chapters. An important outcome of this chapter is the parameterisation in Section 10.3 of *all* stabilising dynamic output feedback controllers.

Contents

10.1 Stabilisability and detectability

In Chapters 2 and 3, we saw some interconnections between controllability, observability, and pole/zero cancellations in the transfer function. At the time, we did not pay too much attention to the nature of the poles and zero that were cancelled. In fact, the illustrative ex-

amples of Section 2.3 were cooked with the goal in mind of illustrating the possible disastrous effects that lack of controllability and observability can have. This need not always be the case. Indeed, it is possible that the a system can be both uncontrollable and unobservable, yet be a system that is tolerable. In this section, we provide the language that expresses the form of this tolerability.

10.1.1 Stabilisability

In Theorem 6.49 we saw that if a system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is controllable, then it is possible by state feedback to render the closed-loop system internally asymptotically stable, and so make the system BIBO stable. The controllability hypothesis is not necessary, and this is captured by the notion of stabilisability. To wit, the system Σ is ***stabilisable*** if $\mathscr{S}_{\text{s}}(\Sigma) \neq \emptyset$. That is, Σ is stabilisable if there exists $\boldsymbol{f} \in \mathbb{R}^n$ so that $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ is Hurwitz. Note that stabilisability depends only on $(\boldsymbol{A}, \boldsymbol{b})$, so we may sometimes say that $(\boldsymbol{A}, \boldsymbol{b})$ is stabilisable rather than saying Σ is stabilisable. The following result describes stabilisability.

10.1 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system and let $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ be invertible with the property that*

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{n-\ell,\ell} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0}_{n-\ell} \end{bmatrix} \tag{10.1}$$

where $(\boldsymbol{A}_{11}, \boldsymbol{b}_1)$ is in controller canonical form (cf. Theorem 2.39). The following are equivalent:

(i) Σ is stabilisable;

(ii) $\boldsymbol{A}_{22}$ is Hurwitz;

(iii) the matrix

$$\left[\begin{array}{c|c} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \end{array}\right]$$

has rank n for all $s \in \overline{\mathbb{C}}_+$.

Proof (i) $\implies$ (ii) Let $\boldsymbol{f} \in \mathscr{S}_{\text{s}}(\Sigma)$. Then $\boldsymbol{T}(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)\boldsymbol{T}^{-1}$ is Hurwitz. Now write

$$\boldsymbol{T}^{-t}\boldsymbol{f} = (\boldsymbol{f}_1, \boldsymbol{f}_2) \in \mathbb{R}^\ell \times \mathbb{R}^{n-\ell}.$$

We then have

$$\boldsymbol{T}(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11} - \boldsymbol{b}_1\boldsymbol{f}_1^t & \boldsymbol{A}_{12} - \boldsymbol{b}_1\boldsymbol{f}_2^t \\ \boldsymbol{0}_{n-\ell,\ell} & \boldsymbol{A}_{22} \end{bmatrix}.$$

This matrix is Hurwitz if and only if $\boldsymbol{A}_{11} - \boldsymbol{b}_1\boldsymbol{f}_1^t$ and $\boldsymbol{A}_{22}$ are Hurwitz, and that $\boldsymbol{A}_{22}$ is Hurwitz is our assertion.

(ii) $\implies$ (iii) Let us define $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{b}}$ to be the expressions for $\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}$ and $\boldsymbol{T}\boldsymbol{b}$ in (10.1). The matrix $\left[\begin{array}{c|c} s\boldsymbol{I}_n - \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{b}} \end{array}\right]$ has rank n exactly when there exists no nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$ with the property that

$$\boldsymbol{x}^t \left[\begin{array}{c|c} s\boldsymbol{I}_n - \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{b}} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{x}^t(s\boldsymbol{I}_n - \tilde{\boldsymbol{A}}) & \boldsymbol{x}^t\boldsymbol{b} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{0}_n^t & 0 \end{array}\right]$$

So suppose that $\boldsymbol{x}$ has the property that this equation *does* hold for some $s_0 \in \mathbb{C}_+$. Let us write $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^\ell \times \mathbb{R}^{n-\ell}$. Thus we have

$$\begin{aligned} \boldsymbol{x}^t \left[\begin{array}{c|c} s_0\boldsymbol{I}_n - \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{b}} \end{array}\right] &= \begin{bmatrix} \boldsymbol{x}_1^t & \boldsymbol{x}_2^t \end{bmatrix} \begin{bmatrix} s_0\boldsymbol{I}_\ell - \boldsymbol{A}_{11} & -\boldsymbol{A}_{12} & \boldsymbol{b}_1 \\ \boldsymbol{0}_{n-\ell,\ell} & s_0\boldsymbol{I}_{n-k\ell} - \boldsymbol{A}_{22} & \boldsymbol{0}_{n-\ell} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{x}_1^t(s_0\boldsymbol{I}_\ell - \boldsymbol{A}_{11}) & -\boldsymbol{x}_1^t\boldsymbol{A}_{12} + s_0\boldsymbol{x}_2^t(\boldsymbol{I}_{n-k\ell} - \boldsymbol{A}_{22}) & \boldsymbol{x}_1^t\boldsymbol{b}_1 \end{bmatrix}, \end{aligned}$$

so that

$$\begin{aligned} \boldsymbol{x}_1^t(s_0\boldsymbol{I}_\ell - \boldsymbol{A}_{11}) = \boldsymbol{0}_\ell^t, \quad \boldsymbol{x}_1^t\boldsymbol{b}_1 = 0, \\ -\boldsymbol{x}_1^t\boldsymbol{A}_{12} + s_0\boldsymbol{x}_2^t(\boldsymbol{I}_{n-k\ell} - \boldsymbol{A}_{22}) = \boldsymbol{0}_{n-\ell}^t \end{aligned} \tag{10.2}$$

Since $(\boldsymbol{A}_{11}, \boldsymbol{b}_1)$ is controllable, by Exercise E2.13 the matrix $\left[\begin{array}{c|c} s_0\boldsymbol{I}_\ell - \boldsymbol{A}_{11} & \boldsymbol{b}_1 \end{array}\right]$ is full rank so that the first two of equations (10.2) implies that $\boldsymbol{x}_1 = \boldsymbol{0}_\ell$. The third of equations (10.2) then says that $\boldsymbol{x}_2$ is a vector in the eigenspace of $\boldsymbol{A}_{22}^t$ with eigenvalue $s_0 \in \overline{\mathbb{C}}_+$. However, since $\boldsymbol{A}_{22}$ is Hurwitz this implies that $\boldsymbol{x}_2 = \boldsymbol{0}_{n-\ell}$. This shows that the matrix $\left[\begin{array}{c|c} s_0\boldsymbol{I}_\ell - \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{b}} \end{array}\right]$ has rank n for $s_0 \in \overline{\mathbb{C}}_+$. Now we note that

$$\left[\begin{array}{c|c} s_0\boldsymbol{I}_\ell - \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{b}} \end{array}\right] = \boldsymbol{T}\left[\begin{array}{c|c} s_0\boldsymbol{I}_\ell - \boldsymbol{A} & \boldsymbol{b} \end{array}\right]\begin{bmatrix} \boldsymbol{T}^{-1} \\ 1 \end{bmatrix}.$$

Therefore, the ranks of $\left[\begin{array}{c|c} s_0\boldsymbol{I}_\ell - \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{b}} \end{array}\right]$ and $\left[\begin{array}{c|c} s_0\boldsymbol{I}_\ell - \boldsymbol{A} & \boldsymbol{b} \end{array}\right]$ agree and the result follows.

(iii) $\implies$ (i) This is Exercise E10.2. ■

The following corollary is obvious from the implication (i) $\implies$ (ii) of the above result.

10.2 Corollary *A SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is stabilisable if it is controllable.*

Let us explore these results with some examples.

10.3 Examples Note that if a system is controllable, then it is stabilisable. Therefore, interesting things concerning stabilisability will happen for uncontrollable systems.

1. Let us first consider a system that is not controllable and not stabilisable. We let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be defined by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

We compute

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

that has rank 1, so the system is indeed uncontrollable. To put the system into the proper form to test for stabilisability, we use the change of basis matrix $\boldsymbol{T}$ defined by

$$\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

Note that the first column of $\boldsymbol{T}^{-1}$ is the input vector $\boldsymbol{b}$, and the other column is a vector not collinear with $\boldsymbol{b}$. We then compute

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

These are in the form of (10.1). Note that $\boldsymbol{A}_{22} = [1]$ which is not Hurwitz. Thus the system is not stabilisable.

2. Now we consider an example that is not controllable but is stabilisable. We define $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

We compute

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

so the system is uncontrollable. Now we define $\boldsymbol{T}$ by

$$\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Again note that $\boldsymbol{b}$ forms the first column of $\boldsymbol{T}$. We also compute

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \boldsymbol{T}\boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

which is in the form of (10.1). Now we have $\boldsymbol{A}_{22} = [-1]$ which is Hurwitz, so the system is stabilisable. •

10.1.2 Detectablilty

The notion of detectability is dual to stabilisability in the same manner that observability is dual to controllability. But let us be precise. A SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is ***detectable*** if there exists a vector $\boldsymbol{\ell} \in \mathbb{R}^n$ with the property that the matrix $A - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz. The following result is analogous to Proposition 10.1.

10.4 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system and let $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ be invertible with the property that*

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0}_{n-k,k} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{c}_1^t & \boldsymbol{0}_{n-k}^t \end{bmatrix} \tag{10.3}$$

where $(\boldsymbol{A}_{11}, \boldsymbol{c}_1)$ are in observer canonical form (cf. Theorem 2.40). The following statements are equivalent:

(i) Σ is detectable;

(ii) $\boldsymbol{A}_{22}$ is Hurwitz;

(iii) the matrix

$$\begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} \\ \boldsymbol{c}^t \end{bmatrix}$$

has rank n for all $s \in \overline{\mathbb{C}}_+$.

Furthermore, the condition

(iv) there exists an output feedback constant $F \in \mathbb{R}$ so that the closed-loop system Σ_F is internally asymptotically stable,

implies the above three conditions.

Proof (i) $\Longrightarrow$ (ii) Let $\boldsymbol{\ell} \in \mathbb{R}^n$ have the property that $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz. Writing $\boldsymbol{T}\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \in \mathbb{R}^k \times \mathbb{R}^{n-k}$ we have

$$\boldsymbol{T}(\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t)\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11} - \boldsymbol{\ell}_1\boldsymbol{c}_1^t & \boldsymbol{0}_{k,n-k} \\ \boldsymbol{A}_{21} - \boldsymbol{\ell}_2\boldsymbol{c}_1^t & \boldsymbol{A}_{22} \end{bmatrix}.$$

This matrix is Hurwitz if and only if the matrices $\boldsymbol{A}_{11} - \boldsymbol{\ell}_1\boldsymbol{c}_1^t$ and $\boldsymbol{A}_{22}$ are Hurwitz. Therefore, if $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz then $\boldsymbol{A}_{22}$ is Hurwitz as claimed.

(ii) $\Longrightarrow$ (iii) Let $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{c}}$ be the matrix and vector in (10.3). The matrix

$$\begin{bmatrix} s\boldsymbol{I}_n - \tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{c}}^t \end{bmatrix}$$

has rank n if the only vector $\boldsymbol{x} \in \mathbb{R}^n$ for which

$$\begin{bmatrix} s\boldsymbol{I}_n - \tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{c}}^t \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \boldsymbol{0}_n \\ 0 \end{bmatrix}$$

is the zero vector. So let $\boldsymbol{x}$ be a vector for which the above equation is satisfied. Then, writing $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^k \times \mathbb{R}^{n-k}$ and letting $s_0 \in \overline{\mathbb{C}}_+$, we have

$$\begin{aligned} \begin{bmatrix} s_0\boldsymbol{I}_n - \tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{c}}^t \end{bmatrix} \boldsymbol{x} &= \begin{bmatrix} s_0\boldsymbol{I}_k - \boldsymbol{A}_{11} & \boldsymbol{0}_{k,n-k} \\ -\boldsymbol{A}_{21} & s_0\boldsymbol{I}_{n-k} - \boldsymbol{A}_{22} \\ \boldsymbol{c}_1^t & \boldsymbol{0}_{n-k}^t \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \\ &= \begin{bmatrix} (s_0\boldsymbol{I}_k - \boldsymbol{A}_{11})\boldsymbol{x}_1 \\ (s_0\boldsymbol{I}_{k-n} - \boldsymbol{A}_{22})\boldsymbol{x}_2 - \boldsymbol{A}_{21}\boldsymbol{x}_1 \\ \boldsymbol{c}_1^t\boldsymbol{x}_1 \end{bmatrix}. \end{aligned}$$

The right-hand side is zero if and only if

$$\begin{aligned} (s_0\boldsymbol{I}_{k-n} - \boldsymbol{A}_{22})\boldsymbol{x}_2 - \boldsymbol{A}_{21}\boldsymbol{x}_1 = \boldsymbol{0}_{n-k}, \\ (s_0\boldsymbol{I}_k - \boldsymbol{A}_{11})\boldsymbol{x}_1 = \boldsymbol{0}_k, \quad \boldsymbol{c}_1^t\boldsymbol{x}_1 = 0. \end{aligned} \tag{10.4}$$

Since $(\boldsymbol{A}_{11}, \boldsymbol{c}_1)$ is observable, the last two of equations (10.4) imply that $\boldsymbol{x}_1 = \boldsymbol{0}_k$ (see Exercise E2.14). Now the first of equations (10.4) imply that $\boldsymbol{x}_2$ is in the eigenspace of $\boldsymbol{A}_{22}$ for the eigenvalue $s_0 \in \overline{\mathbb{C}}_+$. However, since $\boldsymbol{A}_{22}$ is Hurwitz, this implies that $\boldsymbol{x}_2 = \boldsymbol{0}_{n-k}$. Thus we have shown that if $\boldsymbol{A}_{22}$ is Hurwitz then the matrix

$$\begin{bmatrix} s\boldsymbol{I}_n - \tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{c}}^t \end{bmatrix}$$

has full rank. Now we note that

$$\begin{bmatrix} s\boldsymbol{I}_n - \tilde{\boldsymbol{A}} \\ \tilde{\boldsymbol{c}}^t \end{bmatrix} = \begin{bmatrix} \boldsymbol{T} \\ 1 \end{bmatrix} \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} \\ \boldsymbol{c}^t \end{bmatrix} \boldsymbol{T}$$

so that if $\boldsymbol{A}_{11}$ is Hurwitz, it also follows that the matrix

$$\begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} \\ \boldsymbol{c}^t \end{bmatrix}$$

has full rank, as claimed.

(iii) $\Longrightarrow$ (i) This is Exercise E10.3.

(iv) $\Longrightarrow$ (i) This follows since $\boldsymbol{\ell} = F\boldsymbol{b}$ has the property that $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz. ■

The following corollary follows from the implication (i) $\Longrightarrow$ (ii).

10.5 Corollary *A SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is detectable if it is observable.*

10.6 Remark The content of the Propositions 10.1 and 10.4 is easily described in words. A system is stabilisable if the uncontrollable dynamics, represented by the matrix $\boldsymbol{A}_{22}$ in (10.1), are themselves asymptotically stable. Thus, even though the system may not be controllable, this does not hurt you as far as your ability to render the system stable by static state feedback. Similarly, a system is detectable if the unobservable dynamics, represented by the matrix $\boldsymbol{A}_{11}$ in (10.3), are asymptotically stable. Therefore an unobservable system may be made stable under static output feedback if it is detectable. The consequences of these observations are explored in Section 10.2. •

Let us explore these detectability results via examples.

10.7 Examples Note that if a system is observable, then it is detectable. Therefore, interesting things concerning interesting things for detectability will happen for unobservable systems.

1. Let us first consider a system that is not observable and not detectable. We let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be defined by
$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$
We compute
$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$
which has rank 1, so the system is indeed unobservable. To put the system into the proper form to test for detectability, we use the change of basis matrix $\boldsymbol{T}$ defined by
$$\boldsymbol{T}^t = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$
Note that the first column of $\boldsymbol{T}^t$ is the vector $\boldsymbol{c}$ itself, whereas the second column is a vector in $\ker(\boldsymbol{c}^t)$. We then compute
$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$
These are in the form of (10.3). Note that $\boldsymbol{A}_{22} = [1]$ which is not Hurwitz. Thus the system is not detectable.
2. Now we consider an example that is not observable but is detectable. We define $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ by
$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$
We compute
$$\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$
so the system is unobservable. Now we define $\boldsymbol{T}$ by
$$\boldsymbol{T}^t = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$
Again note that $\boldsymbol{c}$ forms the second column of $\boldsymbol{T}$ and that the first column is in $\ker(\boldsymbol{c}^t)$. We compute
$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \boldsymbol{c}^t\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & 0 \end{bmatrix},$$
which is in the form of (10.3). Now we have $\boldsymbol{A}_{22} = [-1]$ which is Hurwitz, so the system is detectable. •

10.1.3 Transfer function characterisations of stabilisability and detectability

The results of the previous two sections were concerned with state-space characterisations of stabilisability and detectability. In this section we look into how these may be manifested in the transfer function. This treatment follows closely that of Section 3.3.

First let us look at the detectability result.

10.8 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system and define polynomials*

$$P_1(s) = \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, \quad P_2(s) = \det(s\boldsymbol{I}_n - \boldsymbol{A}).$$

If $(\boldsymbol{A}, \boldsymbol{b})$ is controllable then Σ is detectable if and only if the GCD of P_1 and P_2 has no roots in $\overline{\mathbb{C}}_+$.

Proof We may as well assume that P_1 and P_2 are not coprime since the result follows from Theorem 3.5 and Corollary 10.5 otherwise. Thus we may as well suppose that $(\boldsymbol{A}, \boldsymbol{c})$ are not observable, and that

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{k,n-k} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{c}^t = \begin{bmatrix} \boldsymbol{0}_k^t & \boldsymbol{c}_2^t \end{bmatrix}$$

where $(\boldsymbol{A}_{22}, \boldsymbol{c}_2)$ is observable. Therefore, if we write $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2) \in \mathbb{R}^k \times \mathbb{R}^{n-k}$ then we compute

$$\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} = \boldsymbol{c}_2^t(s\boldsymbol{I}_{n-k} - \boldsymbol{A}_{22})\boldsymbol{b}.$$

Therefore

$$\frac{\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}}{\det(s\boldsymbol{I}_n - \boldsymbol{A})} = \frac{\boldsymbol{c}_2^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A}_{22})\boldsymbol{b}_2}{\det(s\boldsymbol{I}_{n-k} - \boldsymbol{A}_{22})}.$$

But we also have

$$\det(s\boldsymbol{I}_n - \boldsymbol{A}) = \det(s\boldsymbol{I}_k - \boldsymbol{A}_{11}) \det(s\boldsymbol{I}_{n-k} - \boldsymbol{A}_{22}),$$

from which we conclude that

$$\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b} = \det(s\boldsymbol{I}_k - \boldsymbol{A}_{11})\boldsymbol{c}_2^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A}_{22})\boldsymbol{b}_2.$$

Since $(\boldsymbol{A}_{22}, \boldsymbol{c}_2)$ is observable, the GCD of P_1 and P_2 must be exactly $\det(s\boldsymbol{I}_k - \boldsymbol{A}_{11})$. By Proposition 10.4, the roots of the GCD are in $\mathbb{C}_-$ if and only if Σ is detectable. ■

The consequences of the above result are readily observed in the detectability example we have already introduced.

10.9 Examples **(Example 10.7 cont'd)**

1. Here we had

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

   so that we compute

$$\boldsymbol{c}^t(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} = 1 - s, \quad \det(s\boldsymbol{I}_2 - \boldsymbol{A}) = s^2 - 1$$

   The GCD of these polynomials is $s - 1$ which has the root $1 \in \overline{\mathbb{C}}_+$. Thus, as we have seen, the system is not detectable.

2. Here we take
$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
and so compute
$$\boldsymbol{c}^t(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} = s + 1, \quad \det(s\boldsymbol{I}_2 - \boldsymbol{A}) = s^2 - 1.$$
The GCD of these polynomials is $s + 1$ which has the single root $-1 \in \mathbb{C}_-$, so that the system is detectable. •

Now let us give the analogous result for stabilisability.

10.10 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system and define polynomials*
$$P_1(s) = \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, \quad P_2(s) = \det(s\boldsymbol{I}_n - \boldsymbol{A}).$$
If $(\boldsymbol{A}, \boldsymbol{c})$ is observable then Σ is stabilisable if and only if the GCD of P_1 and P_2 has no roots in $\overline{\mathbb{C}}_+$.

Proof The idea is very much like the proof of Proposition 10.8. We assume that $(\boldsymbol{A}, \boldsymbol{b})$ is not controllable and that
$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{n-\ell,\ell} & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0}_{n-\ell} \end{bmatrix}$$
with $(\boldsymbol{A}_{11}, \boldsymbol{b}_1)$ controllable. By arguments like those in the proof of Proposition 10.8 we show that the GCD of P_1 and P_2 is $\det(s\boldsymbol{I}_{n-\ell} - \boldsymbol{A}_{22})$. By Proposition 10.1 the roots of the GCD are in $\mathbb{C}_-$ if and only if Σ is stabilisable. ■

Again, we may use our existing stabilisability example to illustrate the consequences of this result.

10.11 Examples **(Example 10.3 cont'd)**

1. First we take
$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$
which is observable as it is in observer canonical form. We compute
$$\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} = 1 - s, \quad \det(s\boldsymbol{I}_2 - \boldsymbol{A}) = s^2 - 1.$$
The GCD of these polynomials is $s - 1$ whose root is $1 \in \overline{\mathbb{C}}_+$, leading us to the correct conclusion that the system is not stabilisable.

2. Next we take
$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$
We compute
$$\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_2 - \boldsymbol{A})\boldsymbol{b} = s + 1, \quad \det(s\boldsymbol{I}_2 - \boldsymbol{A}) = s^2 - 1,$$
so the GCD of these polynomials is $s + 1$ whose roots are in $\mathbb{C}_-$. Thus we conclude that the system is stabilisable. •

Finally, we state a result that characterises stabilisability and detectability in terms of cancellation of poles and zeros in the transfer function. This result is rather analogous to Corollary 3.13.

10.12 Corollary *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system and define polynomials*

$$P_1(s) = \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, \quad P_2(s) = \det(s\boldsymbol{I}_n - \boldsymbol{A}).$$

The following statements are equivalent:

(i) Σ is stabilisable and detectable;

(ii) the GCD of P_1 and P_2 has no roots in $\overline{\mathbb{C}}_+$.

We comment that without additional information, one cannot decide whether a system is not stabilisable or not detectable by simply looking at the numerator and denominator of the transfer function. This is made clear in Exercise E10.6. Also, note that now we can complete the implications indicated in Figures 5.1 and Figure 5.2. The result is shown in Figure 10.1.

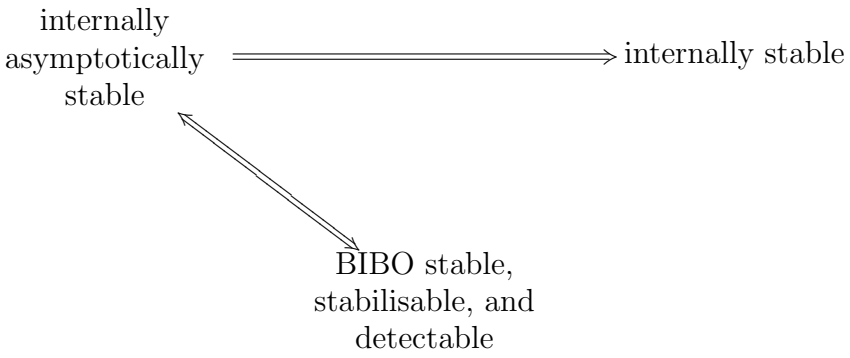


Figure 10.1 Precise relationship between internal and BIBO stability

10.2 Methods for constructing stabilising control laws

In this section, under certain assumptions, we provide explicit formulas for constructing stabilising controllers of various types. The formulas we give are most interesting in that they show that it is in principle possible to explicitly design stabilising controllers of various types. However, as a means for designing controllers, it should be emphasised that the techniques we give here may not in and of themselves be that useful as they only address one aspect of controller design; the necessity that the closed-loop system be stable. There are often present other more demanding criterion given in the form of specific performance criterion (see Chapter 8), or a demand for robustness of the controller to uncertainties in the plant model (see Chapter 15).

10.2.1 Stabilising static state feedback controllers

There is a method for systematically determining the state feedback vector $\boldsymbol{f}$ that will produce the desired poles for the closed-loop transfer function. The formula is called ***Ackermann's formula*** [Ackermann 1972].

10.13 Proposition *Let $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a controllable SISO linear system and suppose that the characteristic polynomial for $\boldsymbol{A}$ is*

$$P_{\boldsymbol{A}}(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0.$$

Let $P \in \mathbb{R}[s]$ be monic and degree n. The state feedback vector $\boldsymbol{f}$ defined by

$$\boldsymbol{f}^t = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} (\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1} P(\boldsymbol{A}).$$

has the property that the characteristic polynomial of the matrix $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ is P.

Proof For the proof of this result, it is convenient to employ a different canonical form than the controller canonical form of Theorem 2.37. From Exercise E2.32 we recall that if we define

$$\boldsymbol{T}_1 = \begin{bmatrix} 1 & -p_{n-1} & 0 & \cdots & 0 & 0 \\ 0 & 1 & -p_{n-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -p_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix},$$

then the matrix $\tilde{\boldsymbol{T}} = \boldsymbol{T}_1^{-1}(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1}$ has the property that

$$\tilde{\boldsymbol{T}}\boldsymbol{A}\tilde{\boldsymbol{T}}^{-1} = \begin{bmatrix} -p_{n-1} & -p_{n-2} & \cdots & -p_1 & -p_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \tilde{\boldsymbol{T}}\boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

One readily verifies that the vector

$$\tilde{\boldsymbol{f}} = (\alpha_{n-1} - p_{n-1}, \dots, \alpha_1 - p_1, \alpha_0 - p_0)$$

has the property that the matrix $\tilde{\boldsymbol{T}}\boldsymbol{A}\tilde{\boldsymbol{T}}^{-1} - \tilde{\boldsymbol{T}}\boldsymbol{b}\tilde{\boldsymbol{f}}^t$ has as characteristic polynomial

$$s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0.$$

The actual state feedback vector will then be defined by $\boldsymbol{f}^t = \tilde{\boldsymbol{f}}^t\tilde{\boldsymbol{T}}$.

We denote $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{T}}\boldsymbol{A}\tilde{\boldsymbol{T}}^{-1}$, and note that by Cayley-Hamilton we have

$$\tilde{\boldsymbol{A}}^n + p_{n-1}\tilde{\boldsymbol{A}}^{n-1} + \cdots + p_1\tilde{\boldsymbol{A}} + p_0\boldsymbol{I}_n = \boldsymbol{0}_n.$$

Suppose that

$$P(s) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0$$

so that

$$P(\tilde{\boldsymbol{A}}) = \tilde{\boldsymbol{A}}^n + \alpha_{n-1}\tilde{\boldsymbol{A}}^{n-1} + \cdots + \alpha_1\tilde{\boldsymbol{A}} + \alpha_0\boldsymbol{I}_n,$$

and subtracting from this the previous expression gives

$$P(\tilde{\boldsymbol{A}}) = (\alpha_{n-1} - p_{n-1})\tilde{\boldsymbol{A}}^{n-1} + \cdots + (\alpha_1 - p_1)\tilde{\boldsymbol{A}} + (\alpha_0 - p_0)\boldsymbol{I}_n. \tag{10.5}$$

One readily determines that if $\{\boldsymbol{e}_1, \dots \boldsymbol{e}_n\}$ is the standard basis for $\mathbb{R}^n$, then

$$\boldsymbol{e}_n^t\boldsymbol{A}^k = \boldsymbol{e}_{n-k}^t, \quad k = 1, \dots, n-1.$$

Therefore, using (10.5), we have

$$\boldsymbol{e}_n^t P(\tilde{\boldsymbol{A}}) = (\alpha_{n-1} - p_{n-1})\boldsymbol{e}_1^t + \cdots + (\alpha_1 - p_1)\boldsymbol{e}_{n-1}^t + (\alpha_0 - p_0)\boldsymbol{e}_n^t.$$

But this shows that $\tilde{\boldsymbol{f}}^t = \boldsymbol{e}_1^t P(\tilde{\boldsymbol{A}})$. It remains to transform $\tilde{\boldsymbol{f}}$ back into the original coordinates to get the state feedback vector $\boldsymbol{f}$. This we do by recalling that $\boldsymbol{f}^t = \tilde{\boldsymbol{f}}^t \tilde{\boldsymbol{T}}$ with $\tilde{\boldsymbol{T}} = \boldsymbol{T}_1^{-1} \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})^{-1}$. Thus we arrive at

$$\boldsymbol{f}^t = \boldsymbol{e}_n^t P(\tilde{\boldsymbol{A}}) \tilde{\boldsymbol{T}} = \boldsymbol{e}_n^t P(\tilde{\boldsymbol{T}} \boldsymbol{A} \tilde{\boldsymbol{T}}^{-1}) \tilde{\boldsymbol{T}} = \boldsymbol{e}_n^t \tilde{\boldsymbol{T}} P(\boldsymbol{A})$$

where we have used the fact that $P(\boldsymbol{T}\boldsymbol{M}\boldsymbol{T}^{-1}) = \boldsymbol{T} P(\boldsymbol{M}) \boldsymbol{T}^{-1}$ for $\boldsymbol{T}, \boldsymbol{M} \in \mathbb{R}^{n \times n}$ with $\boldsymbol{T}$ invertible. Now we see from Exercise E2.31 that $\boldsymbol{e}_n^t \boldsymbol{T}_1^{-1} = \boldsymbol{e}_n^t$, and so this proves that

$$\boldsymbol{f}^t = \boldsymbol{e}_n^t \boldsymbol{T}_1^{-1} (\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1} P(\boldsymbol{A}) = \boldsymbol{e}_n^t (\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1} P(\boldsymbol{A}),$$

as desired. ■

10.14 Remarks

1. The static state feedback of Proposition 10.13 gives an explicit controller that stabilises the closed-loop system, provided that the system is controllable. There are, of course, other stabilising state feedbacks. In fact, there are algorithms for computing static state feedback vectors that place the poles in desirable places. Often an *ad hoc* guess for pole locations will not work well in practice, as, for example, model inaccuracies may have adverse effects on the behaviour of the actual system with a controller designed as in Proposition 10.13. A popular method for designing stabilising static state feedback vectors is the ***linear quadratic regulator*** (***LQR***), where the state feedback is defined to have an optimal property. This is discussed in many books, a recent example of which is [Morris 2000]. In Chapter 14 we look at this for SISO systems using polynomial machinery.
2. In the proof of Proposition 10.13 we have made use of the canonical form developed in Exercise E2.32. The matter of choosing which canonical form to employ is often a matter of convenience. •

Let us illustrate this technology with an example.

10.15 Example (**Example 6.50 cont'd**)

We recall that in this example we had

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Suppose we wish the closed-loop characteristic polynomial to be $s^2 + 4s + 4$ which has the repeated root -2. Thus we compute

$$P(\boldsymbol{A}) = \boldsymbol{A}^2 + 4\boldsymbol{A} + 4\boldsymbol{I}_2 = \begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix}.$$

To apply Proposition 10.13 we compute

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

In this case we compute $\boldsymbol{f}$ to be

$$\boldsymbol{f}^t = \begin{bmatrix} 0 & 1 \end{bmatrix} \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})^{-1} P(\boldsymbol{A}) = \begin{bmatrix} 3 & 4 \end{bmatrix}.$$

This was the state feedback vector presented out of the blue in Example 6.50, and now we see where it came from. •

In Proposition 10.13 the assumption of controllability is made. However, only the assumption of stabilisability *need* be made for stabilisation under static state feedback (pretty much by the very definition of stabilisability). The details of how to construct a stabilising static state feedback for a system that is stabilisable but not controllable is the subject of Exercise E10.7. Let us here state the result.

10.16 Proposition *If $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is stabilisable, it is possible to explicitly construct a feedback vector $\boldsymbol{f}$ with the property that the closed-loop system $\Sigma_{\boldsymbol{f}}$ is internally asymptotically stable, i.e., so that $\boldsymbol{f} \in \mathscr{S}_{\mathrm{s}}(\Sigma)$.*

For this result, we also have the following corollary, that is the analogous result for detectability.

10.17 Corollary *If $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is detectable, it is possible to explicitly construct a vector $\boldsymbol{\ell} \in \mathbb{R}^n$ with the property that the matrix $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz.*

Proof Since Σ is detectable, it follows from part (iii) of Proposition 10.1 and part (i) of Proposition 10.4 that the system $\tilde{\Sigma} = (\boldsymbol{A}^t, \boldsymbol{c}, \boldsymbol{b}^t, \boldsymbol{D})$ is stabilisable. Therefore, by Proposition 10.16, there exists a vector $\boldsymbol{f} \in \mathbb{R}^n$ so that $\boldsymbol{A}^t - \boldsymbol{c}\boldsymbol{f}^t$ is Hurwitz. Therefore, $\boldsymbol{A} - \boldsymbol{f}\boldsymbol{c}^t$ is also Hurwitz, and the result follows by taking $\boldsymbol{\ell} = \boldsymbol{f}$. ■

Note that the result merely tells us to construct a feedback vector as in Proposition 10.13 (or Exercise E10.7 if $(\boldsymbol{A}, \boldsymbol{c})$ is not observable) using $\boldsymbol{A}^t$ and $\boldsymbol{c}$ in place of $\boldsymbol{A}$ and $\boldsymbol{b}$.

10.2.2 Stabilising static output feedback controllers

Interestingly, the problem of stabilisation by static output feedback is significantly more difficult than the problem of stabilisation by static state feedback (see [Syrmos, Abdallah, Dorato, and Grigoriadis 1987] for a fairly recent survey, and [Geromel, Souza, and Skelton 1998] for convexity results). For example, a system can be stabilised by static state feedback if and only if it is stabilisable, and stabilisability is a condition one can computationally get a handle on. However, Blondel and Tsitsiklis [1997] essentially show that the problem of determining whether a system is stabilisable by static *output* feedback is **NP**-hard.[1] Thus there is no easily computable check to see when it is even possible for a system to be stabilisable under static output feedback, never mind an algorithm to compute a static output feedback that actually stabilises.

We shall therefore have to be satisfied with a discussion of static output feedback that is not as complete as the corresponding discussion surrounding static state feedback. To motivate what we do say, recall from Theorem 6.54 that one may express the form of the closed-loop transfer functions available via static output feedback. We wish to determine conditions to test whether the poles of the closed-loop transfer function are those of a given polynomial.

[1] Let us recall, for our own amusement, roughly the idea behind this **NP**-business. The class of ***P*** problems are those for which there is a solution algorithm whose computational complexity satisfies a polynomial bound. The class of problems ***NP*** are those with the property that every solution can be verified as actually being a solution with an algorithm whose computational complexity satisfies a polynomial bound. A famous open problem is, "Does **P**=**NP**?" This problem was posed by Cook [1970] and listed by Smale [1998] as one of the most important problems in mathematics. A problem is ***NP-hard*** if every problem in **NP** can be reduced to it. An **NP**-hard problem may not be in **NP**, and all known **NP**-hard problems are not solvable by an algorithm whose complexity satisfies a polynomial bound. A problem is ***NP-complete*** if it is **NP**-hard, and further is in **NP**. These are, in essence, the "hardest" of the problems in **NP**. All known **NP**-complete problems are not solvable by an algorithm whose complexity satisfies a polynomial bound.

Our first result gives a state space test, and is extracted from [van der Woude 1988]. For a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{D})$, for $k \in \{1, \dots, n\}$ define a $n \times k$ matrix by

$$\boldsymbol{C}_k(\boldsymbol{A}, \boldsymbol{b}) = \left[\begin{array}{c|c|c|c} \boldsymbol{b} & \boldsymbol{A}\boldsymbol{b} & \cdots & \boldsymbol{A}^{k-1}\boldsymbol{b} \end{array} \right]$$

and define a $k \times n$ matrix by

$$\boldsymbol{O}_k(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} \boldsymbol{c}^t \\ \hline \boldsymbol{c}^t\boldsymbol{A} \\ \hline \vdots \\ \hline \boldsymbol{c}^t\boldsymbol{A}^{k-1} \end{bmatrix}.$$

Thus $\boldsymbol{C}_k(\boldsymbol{A}, \boldsymbol{b})$ is comprised of the first k columns of $\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})$ and $\boldsymbol{O}_k(\boldsymbol{A}, \boldsymbol{c})$ is comprised of the first k rows of $\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c})$. With this notation, we have the following result.

10.18 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{D})$ be a complete SISO linear system, and let $P \in \mathbb{R}[s]$ be a degree n monic polynomial. The following statements are equivalent:*

(i) it is possible to choose an output feedback constant F so that the poles of the closed-loop system Σ_F are the roots of P;

(ii) for some $k \in \{1, \dots, n-1\}$ we have $P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_k(\boldsymbol{A}, \boldsymbol{c}))\big) \subset \operatorname{image}(\boldsymbol{C}_{n-k}(\boldsymbol{A}, \boldsymbol{b}))$.

(iii) for all $k \in \{1, \dots, n-1\}$ we have $P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_k(\boldsymbol{A}, \boldsymbol{c}))\big) \subset \operatorname{image}(\boldsymbol{C}_{n-k}(\boldsymbol{A}, \boldsymbol{b}))$.

Proof First recall from Theorem 6.54 that the poles of the closed-loop transfer function Σ_F are the roots of

$$\det(s\boldsymbol{I}_n - (\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)).$$

First let us show that (i) is equivalent to (ii) with $k = 1$. First suppose that (i) holds. Note that if $\boldsymbol{x} \in \ker(\boldsymbol{c}^t)$ then

$$(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}.$$

Therefore,

$$(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)^2\boldsymbol{x} = \boldsymbol{A}^2\boldsymbol{x} - F\boldsymbol{b}\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{x}.$$

Thus $(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)^2\boldsymbol{x} = \boldsymbol{A}^2\boldsymbol{x} + \boldsymbol{y}_2$ where $\boldsymbol{y}_2 \in \operatorname{span}(\boldsymbol{b})$. An easy induction now shows that for $j \in \{2, \dots, n\}$ we have

$$(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)^j\boldsymbol{x} = \boldsymbol{A}^j\boldsymbol{x} + \boldsymbol{y}_j,$$

where $\boldsymbol{y}_j \in \operatorname{span}(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \dots, \boldsymbol{A}^{j-2}\boldsymbol{b})$. By the Cayley-Hamilton theorem we have

$$P(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t) = \boldsymbol{0}_n.$$

Combined with our above calculations this shows that for $\boldsymbol{x} \in \ker(\boldsymbol{c}^t)$ we have

$$\boldsymbol{0} = P(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)\boldsymbol{x} = P(\boldsymbol{A})\boldsymbol{x} + \boldsymbol{y}$$

where $\boldsymbol{y} \in \operatorname{span}(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \dots, \boldsymbol{A}^{n-2}\boldsymbol{b})$. Thus we conclude that

$$P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_1(\boldsymbol{A}, \boldsymbol{c}))\big) \subset \operatorname{image}(\boldsymbol{C}_{n-1}(\boldsymbol{A}, \boldsymbol{b})),$$

and, therefore, that (ii) holds with $k = 1$.

Now suppose that (ii) holds with $k = 1$. We first claim that

$$\operatorname{image}(\boldsymbol{C}_{n-1}(\boldsymbol{A}, \boldsymbol{b})) = \ker\big(\boldsymbol{e}_n^t(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1}\big), \tag{10.6}$$

if $\boldsymbol{e}_n$ is the nth standard basis vector for $\mathbb{R}^n$. Indeed, if $\boldsymbol{x} \in \text{image}(\boldsymbol{C}_{n-1}(\boldsymbol{A}, \boldsymbol{b}))$ if and only if the components of $\boldsymbol{x}$ in the basis $\{\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \dots, \boldsymbol{A}^{n-1}\boldsymbol{b}\}$ are of the form $(x_1, \dots, x_{n-1}, 0)$. However, these components are exactly $(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1}\boldsymbol{x}$:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{bmatrix} = (\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1}\boldsymbol{x}.$$

The relation (10.6) now follows immediately, and by our current hypothesis, so does the relation

$$\begin{aligned} & P(\boldsymbol{A})\big(\ker(\boldsymbol{c}^t)\big) \subset \ker\big(\boldsymbol{e}_n^t(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1}\big) \\ \iff \quad & \ker(\boldsymbol{c}^t) \subset \ker\big(\boldsymbol{e}_n^t(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1}P(\boldsymbol{A})\big) \\ \iff \quad & F\boldsymbol{c}^t = \boldsymbol{e}_n^t(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1}P(\boldsymbol{A}), \end{aligned}$$

for some $F \in \mathbb{R}$. Now, since $(\boldsymbol{A}, \boldsymbol{b})$ is controllable, by Proposition 10.13, if we define $\boldsymbol{f} \in \mathbb{R}^n$ by

$$\boldsymbol{f}^t = \boldsymbol{e}^t(\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1}P(\boldsymbol{A}),$$

then $\det(s\boldsymbol{I}_n - (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)) = P(s)$. This completes the proof of our assertion that (i) is equivalent to (ii) with $k = 1$.

Next let us show that (ii) holds with $k = 1$ if and only if it holds with $k = n - 1$. The equivalence of (i) and (ii) with $k = 1$ tells us that there exists $F \in \mathbb{R}$ so that

$$P(s) = \det(s\boldsymbol{I}_n - (\boldsymbol{A}^t\boldsymbol{c}\boldsymbol{b}^t)) \tag{10.7}$$

if and only if

$$P(\boldsymbol{A}^t)\big(\ker(\boldsymbol{O}_1(\boldsymbol{A}^t, \boldsymbol{b}))\big) \subset \text{image}(\boldsymbol{C}_{n-1}(\boldsymbol{A}^t, \boldsymbol{c})).$$

Now we note that $\boldsymbol{O}_1(\boldsymbol{A}^t, \boldsymbol{b}) = \boldsymbol{C}_1(\boldsymbol{A}, \boldsymbol{b})^t$ and $\boldsymbol{C}_{n-1}(\boldsymbol{A}^t, \boldsymbol{c}) = \boldsymbol{O}_{n-1}(\boldsymbol{A}, \boldsymbol{c})^t$. Thus there exists an F so that (10.7) is satisfied if and only if

$$P(\boldsymbol{A}^t)\big(\ker(\boldsymbol{C}_1(\boldsymbol{A}, \boldsymbol{b})^t)\big) \subset \text{image}(\boldsymbol{O}_{n-1}(\boldsymbol{A}, \boldsymbol{c})^t)$$

Let us make use of a lemma.

1 Lemma *Let $\boldsymbol{T} \in \mathbb{R}^{n\times n}$, $\boldsymbol{M} \in \mathbb{R}^{n\times k}$, and $\boldsymbol{L} \in \mathbb{R}^{(n-k)\times n}$. Then the following statements are equivalent:*

(i) $\boldsymbol{T}^t(\ker(\boldsymbol{M}^t)) \subset \text{image}(\boldsymbol{L}^t)$;

(ii) $\boldsymbol{T}(\ker(\boldsymbol{L})) \subset \text{image}(\boldsymbol{M})$.

Proof Assume that (i) holds. Statement (i) asserts that $\boldsymbol{T}^t$ maps $\ker(\boldsymbol{M}^t)$ to $\text{image}(\boldsymbol{L}^t)$. We claim that this implies that $\boldsymbol{T}$ maps the orthogonal complement of $\text{image}(\boldsymbol{L}^t)$ to the orthogonal complement of $\ker(\boldsymbol{M}^t)$. Indeed, if $\boldsymbol{x}$ is in the orthogonal complement to $\text{image}(\boldsymbol{L}^t)$ and if $\boldsymbol{y} \in \ker(\boldsymbol{M}^t)$ then we have

$$\langle \boldsymbol{T}\boldsymbol{x}, \boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{T}\boldsymbol{y}\rangle.$$

Since (i) holds, it follows that $\boldsymbol{T}\boldsymbol{y} \in \text{image}(\boldsymbol{L}^t)$, showing that $\langle \boldsymbol{T}\boldsymbol{x}, y\rangle = 0$ if $\boldsymbol{x}$ is in the orthogonal complement to $\text{image}(\boldsymbol{L}^t)$ and if $\boldsymbol{y} \in \ker(\boldsymbol{M}^t)$. That is, $\boldsymbol{T}$ maps the orthogonal

complement of image($\boldsymbol{L}^t$) to the orthogonal complement of ker($\boldsymbol{M}^t$). However, the orthogonal complement of image($\boldsymbol{L}^t$) is exactly ker($\boldsymbol{L}$) and the orthogonal complement of ker($\boldsymbol{M}^t$) is exactly image($\boldsymbol{M}$). Thus we have shown that

$$\boldsymbol{T}(\ker(\boldsymbol{L})) \subset \operatorname{image}(\boldsymbol{M}),$$

thus (ii) holds. Clearly this proof is symmetric, so the converse holds trivially. ▼

Applying the lemma to the case when $\boldsymbol{T} = P(\boldsymbol{A})$, $\boldsymbol{u} = \boldsymbol{C}_1(\boldsymbol{A}, \boldsymbol{b})$, and $\boldsymbol{L} = \boldsymbol{O}_{n-1}(\boldsymbol{A}, \boldsymbol{c})$ we complete the proof of the fact that (ii) with $k = 1$ implies (ii) with $k = n - 1$. The converse implication is a matter of reversing the above computations.

To this point we have shown that the following three statements are equivalent:

1. it is possible to choose an output feedback constant F so that the poles of the closed-loop system Σ_F are the roots of P;
2. $P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_1(\boldsymbol{A}, \boldsymbol{c}))\big) \subset \operatorname{image}(\boldsymbol{C}_{n-1}(\boldsymbol{A}, \boldsymbol{b}))$;
3. $P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_{n-1}(\boldsymbol{A}, \boldsymbol{c}))\big) \subset \operatorname{image}(\boldsymbol{C}_1(\boldsymbol{A}, \boldsymbol{b}))$.

We complete the proof by showing that if (ii) holds with $k = \ell$ then it also holds with $k = \ell - 1$.

Without loss of generality, suppose that $(\boldsymbol{A}, \boldsymbol{c})$ are in observer canonical form so that

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -p_0 \\ 1 & 0 & 0 & \cdots & 0 & -p_1 \\ 0 & 1 & 0 & \cdots & 0 & -p_2 \\ 0 & 0 & 1 & \cdots & 0 & -p_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -p_{n-2} \\ 0 & 0 & 0 & \cdots & 1 & -p_{n-1} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{n-2} \\ b_{n-1} \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

A direct computation than shows that

$$\ker(\boldsymbol{O}_\ell(\boldsymbol{A}, \boldsymbol{c})) = \operatorname{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{n-\ell})$$

and $\boldsymbol{A}^{\ell-1}\boldsymbol{e}_1 = \boldsymbol{e}_\ell$ for $\ell \in \{1, \ldots, n-1\}$. If we let $\boldsymbol{y} = P(\boldsymbol{A})\boldsymbol{e}_1$ it therefore follows that

$$P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_\ell(\boldsymbol{A}, \boldsymbol{c}))\big) = \operatorname{span}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}, \ldots, \boldsymbol{A}^{n-\ell-1}\boldsymbol{y}).$$

Now assume that

$$P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_\ell(\boldsymbol{A}, \boldsymbol{c}))\big) \subset \operatorname{image}(\boldsymbol{C}_{n-\ell}(\boldsymbol{A}, \boldsymbol{b})).$$

We then compute

$$\begin{aligned} P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_{\ell-1}(\boldsymbol{A}, \boldsymbol{c}))\big) &= \operatorname{span}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}, \ldots, \boldsymbol{A}^{n-\ell}\boldsymbol{y}) \\ &= \operatorname{span}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}, \ldots, \boldsymbol{A}^{n-\ell-1}\boldsymbol{y}) + \boldsymbol{A}\operatorname{span}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}, \ldots, \boldsymbol{A}^{n-\ell-1}\boldsymbol{y}) \\ &= P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_\ell(\boldsymbol{A}, \boldsymbol{c}))\big) + \boldsymbol{A}P(\boldsymbol{A})\big(\ker(\boldsymbol{O}_\ell(\boldsymbol{A}, \boldsymbol{c}))\big) \\ &\subset \operatorname{image}(\boldsymbol{C}_{n-\ell}(\boldsymbol{A}, \boldsymbol{b})) + \boldsymbol{A}\big(\operatorname{image}(\boldsymbol{C}_{n-\ell}(\boldsymbol{A}, \boldsymbol{b}))\big) \\ &= \operatorname{image}(\boldsymbol{C}_{n-\ell+1}(\boldsymbol{A}, \boldsymbol{b})). \end{aligned}$$

This completes the proof. ■

The theorem gives an insight into the difficultly in determining which closed-loop poles are available to us. Unlike its static state feedback counterpart Proposition 10.13, the conditions given by Theorem 10.18 for determining whether the closed-loop poles are the roots of a polynomial P *involve the polynomial* P *itself*. This is what makes the problem a computationally difficult one. Let us illustrate Theorem 10.18 on an example.

10.19 Example **(Example 6.55 cont'd)** We take

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1,$$

noting that $\boldsymbol{A}$ and $\boldsymbol{b}$ are as in Example 10.15. For $F \in \mathbb{R}$ we determine that

$$\det(s\boldsymbol{I}_2 - (\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)) = s^2 + 4Fs + 1 + 3F.$$

Thus, if we let $P_F(s) = s^2 + 4Fs + 1 + 3F$, we see that the only places we may assign poles under static output feedback are at the roots of P_F for some F. Let us see how this checks in with Theorem 10.18. A simple computation gives

$$P_F(\boldsymbol{A}) = \begin{bmatrix} 3F & 4F \\ -4F & 3F \end{bmatrix}.$$

Because $n = 2$, the only cases that are applicable for statements (ii) and (iii) of Theorem 10.18 are when $k = 1$. We then have $\boldsymbol{O}_1(\boldsymbol{A}, \boldsymbol{c}) = \boldsymbol{c}^t$ and $\boldsymbol{C}_1(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{b}$. Therefore

$$\ker(\boldsymbol{O}_1(\boldsymbol{A}, \boldsymbol{c})) = \operatorname{span}((4, -3)), \quad \operatorname{image}(\boldsymbol{C}_1(\boldsymbol{A}, \boldsymbol{b})) = \operatorname{span}((0, 1)).$$

We compute

$$P_F(\boldsymbol{A}) \begin{bmatrix} 4 \\ -3 \end{bmatrix} = \begin{bmatrix} 0 \\ -25F \end{bmatrix}.$$

Therefore $P_F(\boldsymbol{A})\big(\ker(\boldsymbol{O}_1(\boldsymbol{A}, \boldsymbol{c}))\big) \subset \operatorname{image}(\boldsymbol{C}_1(\boldsymbol{A}, \boldsymbol{b}))$, just as stated in Theorem 10.18.

Conversely, let $P(s) = s^2 + as + b$ and compute

$$P(\boldsymbol{A}) = \begin{bmatrix} b - 1 & a \\ -a & b - 1 \end{bmatrix},$$

and

$$P(\boldsymbol{A}) \begin{bmatrix} 4 \\ -3 \end{bmatrix} = \begin{bmatrix} 4b - 3a - 4 \\ 3 - 4a - 3b \end{bmatrix}.$$

Thus we have

$$\begin{bmatrix} 4b - 3a - 4 \\ 3 - 4a - 3b \end{bmatrix} \in \operatorname{image}(\boldsymbol{C}_1(\boldsymbol{A}, \boldsymbol{b}))$$

if and only if $4b - 3a - 4 = 0$. The linear equation

$$\begin{bmatrix} -3 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 4$$

does not have a unique solution, of course. One may check that $(a, b) = (0, 1)$ is a solution. All solutions are then of the form

$$(0, 1) + \boldsymbol{x}, \quad \boldsymbol{x} \in \ker((-3, 4)^t) = \operatorname{span}((4, 3)).$$

That is to say, if the poles of the closed-loop system are to roots of $P(s) = s^2 + as + b$, then we must have $(a, b) = (0, 1) + F(4, 3)$. Thus $P(s) = s^2 + 4Fs + (1 + 3F)$, just as we noticed initially. •

Note that Theorem 10.18 does not tell us when we may choose an output feedback constant F with the property that the closed-loop system Σ_F is stable. The following result of Kučera and Souza [1995] gives a characterisation of stabilisability in terms of the Liapunov ideas of Section 5.4 and the Riccati equations ideas that will come up in Section 14.3.2.

10.20 Theorem *For a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, 0_1)$, the following two statements are equivalent:*

(i) $\mathscr{S}_{\mathrm{o}}(\Sigma) \neq \emptyset$;

(ii) the following two conditions hold:

(a) Σ *is stabilisable and detectable;*

(b) there exists $F \in \mathbb{R}$ *and* $\boldsymbol{g} \in \mathbb{R}^n$ *so that*

$$F\boldsymbol{c}^t + \boldsymbol{b}^t\boldsymbol{P} = \boldsymbol{g}^t, \tag{10.8}$$

where $\boldsymbol{P}$ *is the unique positive-semidefinite solution of the equation*

$$\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} - \boldsymbol{P}\boldsymbol{b}\boldsymbol{b}^t\boldsymbol{P} = -\boldsymbol{c}\boldsymbol{c}^t - \boldsymbol{g}\boldsymbol{g}^t. \tag{10.9}$$

Proof (i) $\Longrightarrow$ (ii) Let $F \in \mathbb{R}$ have the property that $\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t$ is Hurwitz. Then clearly $(\boldsymbol{A}, \boldsymbol{b})$ is stabilisable and $(\boldsymbol{A}, \boldsymbol{c})$ is detectable. Since $\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t$ is Hurwitz, by part (ii) of Theorem 5.32 there exists a unique positive-semidefinite matrix $\boldsymbol{P}$ that satisfies

$$(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)^t\boldsymbol{P} + \boldsymbol{P}(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t) = -(1 + F^2)\boldsymbol{c}\boldsymbol{c}^t.$$

Straightforward manipulation of this expression gives

$$\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} - \boldsymbol{P}\boldsymbol{b}\boldsymbol{b}^t\boldsymbol{P} = -\boldsymbol{c}\boldsymbol{c}^t - (\boldsymbol{b}^t\boldsymbol{P} - F\boldsymbol{c}^t)^t(\boldsymbol{b}^t\boldsymbol{P} - F\boldsymbol{c}^t).$$

This part of the proof now follows by taking $\boldsymbol{g} = \boldsymbol{P}\boldsymbol{b}\boldsymbol{P} - F\boldsymbol{c}$.

(ii) $\Longrightarrow$ (i) Let F and $\boldsymbol{g}$ be as in the statement of the theorem. One then directly verifies using the properties of F and $\boldsymbol{g}$ that

$$(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t)^t\boldsymbol{P} + \boldsymbol{P}(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t) = -(1 + F^2)\boldsymbol{c}\boldsymbol{c}^t.$$

Let $\boldsymbol{\ell} \in \mathbb{R}$ have the property that $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz. We then note that

$$\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t = (\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t) + \begin{bmatrix}\boldsymbol{\ell} & -\boldsymbol{b}\end{bmatrix}\begin{bmatrix}\boldsymbol{c}^t \\ F\boldsymbol{c}^t\end{bmatrix}.$$

This means that the two output system $\tilde{\Sigma} = (\tilde{\boldsymbol{A}} = \boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t, \boldsymbol{b}, \boldsymbol{C}, 0_1)$ where

$$\boldsymbol{C} = \begin{bmatrix}\boldsymbol{c}^t \\ F\boldsymbol{c}^t\end{bmatrix}$$

is detectable since $(\boldsymbol{A} - F\boldsymbol{b}\boldsymbol{c}^t) - \boldsymbol{L}\boldsymbol{C}$ is Hurwitz if

$$\boldsymbol{L} = \begin{bmatrix}\boldsymbol{\ell} & -\boldsymbol{b}\end{bmatrix}.$$

From the MIMO version of Exercise E10.4, the result now follows. ■

10.21 Remark In Section 14.3.2 we will see that the equation (10.9) comes up naturally in an optimisation problem. This equation is called an ***algebraic Riccati equation***, and there are well-developed numerical schemes for obtaining solutions; it is a "nice" equation numerically. However, in the statement of Theorem 10.20 we see that not only must the algebraic Riccati equation be satisfied, but the subsidiary condition (10.8) must also be met. This, it turns out, takes the numerical problem out of the "nice" category in which the algebraic Riccati equations sits. Again, with static output feedback, things are never easy. •

Now let us adopt a different, more constructive approach. While the approach *is* more constructive, it is not a sharp construction, as we shall see. First, the following result gives a crude necessary condition for an output feedback constant to stabilise.

10.22 Proposition *Let Σ, P, and Q be as in Theorem 10.23 and write*

$$P(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0$$
$$Q(s) = q_m s^m + \cdots + q_1 s + q_0.$$

Define

$$F_{\text{lower}} = -\min\left\{\tfrac{p_j}{q_j} \;\middle|\; q_j > 0,\ j = 0, \ldots, m\right\}$$
$$F_{\text{upper}} = -\max\left\{\tfrac{p_j}{q_j} \;\middle|\; q_j < 0,\ j = 0, \ldots, m\right\}.$$

If $q_j \leq 0$ for all $j \in \{0, 1, \ldots, m\}$ then take $F_{\text{lower}} = -\infty$ and if $q_j \geq 0$ for all $j \in \{0, 1, \ldots, m\}$ then take $F_{\text{upper}} = \infty$. If the closed-loop system Σ_F is internally asymptotically stable for some $F \in \mathbb{R}$, then it must be the case that $F \in (F_{\text{lower}}, F_{\text{upper}})$.

Proof By Theorem 6.54, the poles of the closed-loop transfer function Σ_F are the roots of the polynomial

$$P_F(s) = P(s) + FQ(s).$$

If P_F is Hurwitz then all coefficients of P_F be positive (see Exercise E5.18). In particular, we should have

$$p_j + Fq_j > 0, \quad j = 1, \ldots, m.$$

This relation will be satisfied in $F \in (F_{\text{lower}}, F_{\text{upper}})$. ■

Thus we can now restrict our search for feasible output feedback constants to those in the interval $(F_{\text{lower}}, F_{\text{upper}})$. However, we still do not know when an output feedback constant F *does* stabilise. Let us address this by giving a result of Chen [1993]. To state the result, we need the notion of a "generalised eigenvalue." Given matrices $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n\times n}$, a ***generalised eigenvalue*** is a number $\lambda \in \mathbb{C}$ that satisfies

$$\det(\boldsymbol{M}_1 - \lambda \boldsymbol{M}_2) = 0.$$

We denote by $\sigma(\boldsymbol{M}_1, \boldsymbol{M}_2)$ the set of generalised eigenvalues. We also recall from Section 5.5.2 the $n \times n$ Hurwitz matrix $\boldsymbol{H}(P)$ that we may associate to a polynomial $P \in \mathbb{R}[s]$ of degree n. Note that if $Q \in \mathbb{R}[s]$ has degree less than n, then we may still define an $n \times n$ matrix $\boldsymbol{H}(Q)$ by thinking of Q as being degree n with the coefficients of the higher order terms being zero.

10.23 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a controllable SISO linear system. Denote*

$$P(s) = \det(s\boldsymbol{I}_n - \boldsymbol{A}), \quad Q(s) = \boldsymbol{c}^t \operatorname{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$$

with $\boldsymbol{H}(P)$ and $\boldsymbol{H}(Q)$ the corresponding Hurwitz matrices, supposing both to be $n \times n$. Let

$$\sigma(\boldsymbol{H}(P), -\boldsymbol{H}(Q)) \cap \mathbb{R} = \{\lambda_1, \ldots, \lambda_k\}$$

with $\lambda_1 \leq \cdots \leq \lambda_k$. The following statements hold:

(i) for $i \in \{1, \ldots, k\}$, the closed-loop system Σ_{λ_i} is not internally asymptotically stable;

(ii) for $j \in \{1, \ldots, k-1\}$, the closed-loop system Σ_F is internally asymptotically stable for all *$F \in (\lambda_j, \lambda_{j+1})$ if and only if $\Sigma_{\bar{F}}$ is internally asymptotically stable for* some *$\bar{F}(\lambda_j, \lambda_{j+1})$.*

We do not present the proof of this theorem, as the elements needed to get the proof underway would take us too far afield. However, let us indicate how one might use the theorem.

10.24 Method for generating stabilising output feedback constants Suppose you are given $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$.

1. Compute
$$P(s) = \det(s\boldsymbol{I}_n - \boldsymbol{A}), \quad Q(s) = \boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}.$$
2. Compute F_{lower} and F_{upper} as in Proposition 10.22.
3. Compute $\boldsymbol{H}(P)$ and $\boldsymbol{H}(Q)$ as in Theorem 10.23.
4. Compute and order $\{\lambda_1, \dots, \lambda_k\} = \sigma(\boldsymbol{H}(P), -\boldsymbol{H}(Q)) \cap \mathbb{R}$.
5. If $(\lambda_j, \lambda_{j+1}) \not\subset (F_{\text{lower}}, F_{\text{upper}})$, then any $F \in (\lambda_j, \lambda_{j+1})$ is not stabilising.
6. If $(\lambda_j, \lambda_{j+1}) \subset (F_{\text{lower}}, F_{\text{upper}})$, then choose $\bar{F} \in (\lambda_j, \lambda_{j+1})$ and check whether $P_{\bar{F}}(s) = P(s) + FQ(s)$ is Hurwitz (use, for example, the Routh/Hurwitz criterion).
7. If $P_{\bar{F}}$ is Hurwitz, then P_F is Hurwitz for any $F \in (\lambda_j, \lambda_{j+1})$. •

Let us try this out on a simple example.

10.25 Example ((10.19) cont'd) We resume our example where
$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \boldsymbol{D} = \boldsymbol{0}_1.$$
We compute
$$P(s) = s^2 + 1, \quad Q(s) = 4s + 3,$$
which gives
$$F_{\text{lower}} = 0, \quad F_{\text{upper}} = \infty.$$
One may also determine that
$$\boldsymbol{H}(P) = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{H}(Q) = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}.$$
We then compute
$$\det(\boldsymbol{H}(P) + \lambda \boldsymbol{H}(Q)) = 12s^2 + 4s.$$
Thus
$$\sigma(\boldsymbol{H}(P), -\boldsymbol{H}(Q)) \cap \mathbb{R} = \{-\tfrac{1}{3}, 0\},$$
so $\lambda_1 = -\frac{1}{3}$ and $\lambda_2 = 0$. We note that $(\lambda_1, \lambda_2) \not\subset (F_{\text{lower}}, F_{\text{upper}})$, so from Theorem 10.23 we may only conclude that there are no stabilising output feedback constants in (λ_1, λ_2). However, note that in this example, any $F \in (F_{\text{lower}}, F_{\text{upper}})$ is actually stabilising. Thus while Theorem 10.23 provides a way to perhaps obtain *some* stabilising output feedback constants, it does not provide all of them. This further points out the genuine difficulty of developing a satisfactory theory for stabilisation using static output feedback, even in the SISO context. •

10.26 Remark While the above discussion suggests that obtaining a fully-developed theory for stabilisation by static output feedback may be troublesome, in practice, things are not as grim as they have been painted out to be, at least for SISO systems. Indeed, the matter of finding stabilising output feedback constants is exactly the problem of finding constants F so that the polynomial

$$P_F(s) = P(s) + FQ(s)$$

is Hurwitz. A standard way to do this is using root-locus methods developed by Evans (1948, 1950), and presented here in Chapter 11. It is also possible to use the Nyquist criterion to obtain suitable values for F. Note, however, that both of these solution methods, while certainly usable in practice, are graphical, and do not involve concrete formulas, as do the corresponding formulas for static state feedback in Section 10.2.1 and for dynamic output feedback in Section 10.2.3. Thus one's belief in such solutions methods is exactly as deep as one's trust in graphical methods. •

10.2.3 Stabilising dynamic output feedback controllers

In this section we will show that it is always possible to construct a dynamic output feedback controller that renders the resulting closed-loop system internally asymptotically stable, provided that the plant is stabilisable and detectable. This is clearly interesting. First of all, we should certainly expect that we will have to make the assumption of stabilisability and detectability. If these assumptions are not made, then it is not hard to believe that there will be no way to make the plant internally asymptotically stable under feedback since the plant has internal unstable dynamics that are neither controlled by the input nor observed by the output.

First we recall that if $\Sigma_P = (\boldsymbol{A}_P, \boldsymbol{b}_P, \boldsymbol{c}_P^t, \boldsymbol{D}_P)$ is stabilisable and detectable there exists two vectors $\boldsymbol{f}, \boldsymbol{\ell} \in \mathbb{R}^n$ with the property that the matrices $\boldsymbol{A}_P - \boldsymbol{b}_P\boldsymbol{f}^t$ and $\boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t$ are Hurwitz. With this as basis, we state the following result.

10.27 Theorem *Let $\Sigma_P = (\boldsymbol{A}_P, \boldsymbol{b}_P, \boldsymbol{c}_P^t, \boldsymbol{D}_P)$ be a SISO linear plant. Then the following statements are equivalent:*

(i) Σ_P is stabilisable and detectable;

(ii) there exists a SISO linear controller $\Sigma_C = (\boldsymbol{A}_C, \boldsymbol{b}_C, \boldsymbol{c}_C^t, \boldsymbol{D}_C)$ with the property that the closed-loop system is internally asymptotically stable.

Furthermore, if either of these equivalent conditions is satisfied and if $\boldsymbol{f}, \boldsymbol{\ell} \in \mathbb{R}^n$ have the property that the matrices $\boldsymbol{A}_P - \boldsymbol{b}_P\boldsymbol{f}^t$ and $\boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t$ are Hurwitz, then the SISO linear controller Σ_C defined by

$$\begin{aligned} \boldsymbol{A}_C &= \boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t - \boldsymbol{b}_P\boldsymbol{f}^t + \boldsymbol{D}_P\boldsymbol{\ell}\boldsymbol{f}^t, \\ \boldsymbol{b}_C &= \boldsymbol{\ell}, \quad \boldsymbol{c}_C^t = \boldsymbol{f}^t, \quad \boldsymbol{D} = \boldsymbol{0}_1 \end{aligned}$$

has the property that the closed-loop system is internally asymptotically stable.

Proof (i) $\Longrightarrow$ (ii) Using Proposition 6.56 we compute the closed-loop system matrix $\boldsymbol{A}_{\text{cl}}$ to be

$$\boldsymbol{A}_{\text{cl}} = \begin{bmatrix} \boldsymbol{A}_P & \boldsymbol{b}_P\boldsymbol{f}^t \\ -\boldsymbol{\ell}\boldsymbol{c}_P^t & \boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t - \boldsymbol{b}_P\boldsymbol{f}^t \end{bmatrix}.$$

Now define $\boldsymbol{T} \in \mathbb{R}^{2n\times 2n}$ by

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{I}_n \\ \boldsymbol{0}_n & \boldsymbol{I}_n \end{bmatrix} \Longrightarrow \quad \boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{I}_n \\ \boldsymbol{0}_n & \boldsymbol{I}_n \end{bmatrix}.$$

One readily computes that

$$\boldsymbol{T}\boldsymbol{A}_{\mathrm{cl}}\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t & \boldsymbol{0}_n \\ -\boldsymbol{\ell}\boldsymbol{c}_P^t & \boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t \end{bmatrix}.$$

In particular,

$$\operatorname{spec}(\boldsymbol{A}_{\mathrm{cl}}) = \operatorname{spec}(\boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t) \cup \operatorname{spec}(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t),$$

which says that the eigenvalues of $\boldsymbol{A}_{\mathrm{cl}}$ are in $\mathbb{C}_-$, as desired, provided we choose $\boldsymbol{f}$ and $\boldsymbol{\ell}$ appropriately. This also proves the second assertion of the theorem.

(ii) $\Longrightarrow$ (i) If Σ_P is neither stabilisable nor detectable, it is also neither controllable nor observable. Therefore, by Theorem 2.41 we may suppose that $\boldsymbol{A}_P$, $\boldsymbol{b}_P$, and $\boldsymbol{c}_P$ have the form

$$\boldsymbol{A}_P = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \boldsymbol{A}_{13} & \boldsymbol{A}_{14} \\ \boldsymbol{0}_{k\times j} & \boldsymbol{A}_{22} & \boldsymbol{0}_{j\times\ell} & \boldsymbol{A}_{24} \\ \boldsymbol{0}_{\ell\times j} & \boldsymbol{0}_{\ell\times k} & \boldsymbol{A}_{33} & \boldsymbol{A}_{34} \\ \boldsymbol{0}_{m\times j} & \boldsymbol{0}_{m\times k} & \boldsymbol{0}_{m\times\ell} & \boldsymbol{A}_{44} \end{bmatrix}, \quad \boldsymbol{b}_P = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \boldsymbol{0}_\ell \\ \boldsymbol{0}_m \end{bmatrix}, \quad \boldsymbol{c}_P = \begin{bmatrix} \boldsymbol{0}_j \\ \boldsymbol{c}_2 \\ \boldsymbol{0}_\ell \\ \boldsymbol{c}_4 \end{bmatrix},$$

for suitable j, k, ℓ, and m. The assumption that Σ_P is neither stabilisable nor detectable is equivalent to saying that the matrix $\boldsymbol{A}_{33}$ is not Hurwitz. First let us suppose that $\boldsymbol{A}_{33}$ has a real eigenvalue $\lambda \in \overline{\mathbb{C}}_+$ with $\boldsymbol{v}$ an eigenvector. Consider a controller SISO linear system $\Sigma_C = (\boldsymbol{A}_C, \boldsymbol{b}_C, \boldsymbol{c}_C^t, \boldsymbol{D}_C)$, giving rise to the closed-loop equations

$$\begin{aligned} \dot{\boldsymbol{x}}_P(t) &= \boldsymbol{A}_P\boldsymbol{x}_P(t) + \boldsymbol{b}_P u(t) \\ \dot{\boldsymbol{x}}_C(t) &= \boldsymbol{A}_C\boldsymbol{x}_C(t) - \boldsymbol{b}_C y(t) \\ y(t) &= \boldsymbol{c}_P^t\boldsymbol{x}_P(t) + \boldsymbol{D}_P u(t) \\ u(t) &= \boldsymbol{c}_C^t\boldsymbol{x}_C(t) - \boldsymbol{D}_C y(t). \end{aligned}$$

If we choose initial conditions for $\boldsymbol{x}_P$ and $\boldsymbol{x}_C$ as

$$\boldsymbol{x}_P(0) = \begin{bmatrix} \boldsymbol{0}_j \\ \boldsymbol{0}_k \\ \boldsymbol{v} \\ \boldsymbol{0}_m \end{bmatrix}, \quad \boldsymbol{x}_C(0) = \boldsymbol{0},$$

the resulting solution to the closed-loop equations will simply be

$$\boldsymbol{x}_P(t) = e^{\lambda t}\begin{bmatrix} \boldsymbol{0}_j \\ \boldsymbol{0}_k \\ \boldsymbol{v} \\ \boldsymbol{0}_m \end{bmatrix}, \quad \boldsymbol{x}_C(t) = \boldsymbol{0}.$$

In particular, the closed-loop system is not internally asymptotically stable. If the eigenvalue in $\overline{\mathbb{C}}_+$ is not real, obviously a similar argument can be constructed. ■

10.28 Remark The stabilising controller Σ_C constructed in the theorem has the same order, i.e., the same number of state variables as the plant Σ_P. It can be expected that frequently one can do much better than this and design a significantly "simpler" controller. In Section 10.3 we parameterise (almost) all stabilising controllers which includes the one of Theorem 10.27 as a special case. •

This gives the following interesting corollary that relates to the feedback interconnection of Figure 10.2. Note that this answer the question raised at the end of Section 6.3.1.

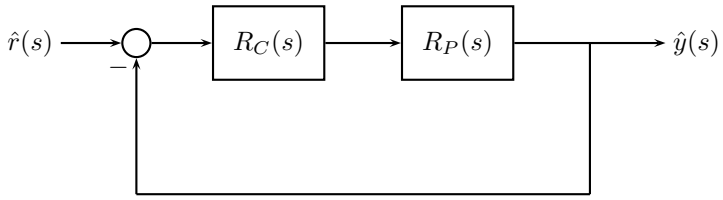


Figure 10.2 Unity gain feedback loop

10.29 Corollary *Let $R_P \in \mathbb{R}(s)$ which we assume to be represented by its c.f.r. (N_P, D_P). Then it is possible to compute a controller rational function R_C with the property that the closed-loop interconnection of Figure 10.2 is IBIBO stable.*

Proof This follows immediately from Theorem 10.27 since the canonical minimal realisation of R_P is controllable and observable, and so stabilisable and detectable. ■

Let us see how this works in an example. We return to the example of Section 6.4.3, except now we do so with a methodology in mind.

10.30 Example **(Example 6.59)** In this example we had

$$\boldsymbol{A}_P = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{b}_P = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}, \quad \boldsymbol{c}_P = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{D}_P = \boldsymbol{0}_1.$$

To apply Theorem 10.27 we need to design vectors $\boldsymbol{f}$ and $\boldsymbol{\ell}$ so that $\boldsymbol{A}_P - \boldsymbol{b}_P\boldsymbol{f}^t$ and $\boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t$ are Hurwitz. To do this, it is required that $(\boldsymbol{A}_P, \boldsymbol{b}_P)$ be stabilisable and that $(\boldsymbol{A}_P, \boldsymbol{c}_P)$ be detectable. But we compute

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus Σ_P is controllable (and hence stabilisable) and observable (and hence detectable). Thus we may use Proposition 10.13 to construct $\boldsymbol{f}$ and Corollary 10.17 to construct $\boldsymbol{\ell}$. In each case, we ask that the characteristic polynomial of the closed-loop matrix be $\alpha(s) = s^2 + 2s^2$ which has roots $-1 \pm i$. Thus we define

$$\begin{aligned} \boldsymbol{f}^t &= \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b})^{-1} \alpha(\boldsymbol{A}) = \begin{bmatrix} 2m & 2m \end{bmatrix} \\ \boldsymbol{\ell}^t &= \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \boldsymbol{O}(\boldsymbol{A}^t, \boldsymbol{c})^{-1} \alpha(\boldsymbol{A}^t) = \begin{bmatrix} 2 & 2 \end{bmatrix}. \end{aligned}$$

Here, following Proposition 10.13, we have used

$$\alpha(\boldsymbol{A}) = \boldsymbol{A}^2 + 2\boldsymbol{A} + 2\boldsymbol{I}_2, \quad \alpha(\boldsymbol{A}^t) = (\boldsymbol{A}^t)^2 + 2\boldsymbol{A}^t + 2\boldsymbol{I}_2.$$

Using Theorem 10.27 we then ascertain that

$$\begin{aligned} \boldsymbol{A}_C &= \boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t - \boldsymbol{b}_P\boldsymbol{f}^t = \begin{bmatrix} -2 & 1 \\ -4 & -2 \end{bmatrix} \\ \boldsymbol{b}_C &= \boldsymbol{\ell} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \boldsymbol{c}_C^t = \boldsymbol{f}^t = \begin{bmatrix} 2m & 2m \end{bmatrix}. \end{aligned}$$

Let us check that the closed-loop system, as defined by Proposition 6.56, is indeed internally asymptotically stable. A straightforward application of Proposition 6.56 gives

$$\boldsymbol{A}_{\mathrm{cl}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ -2 & 0 & -2 & 1 \\ -2 & 0 & -4 & -2 \end{bmatrix},$$

$$\boldsymbol{b}_{\mathrm{cl}} = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 2 \end{bmatrix}, \quad \boldsymbol{c}_{\mathrm{cl}}^t = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{D}_{\mathrm{cl}} = \boldsymbol{0}_1.$$

One may check that the eigenvalues of the $\boldsymbol{A}_{\mathrm{cl}}$ are $\{-1 \pm i\}$ where each root is has algebraic multiplicity 2. As predicted by the proof of Theorem 10.27, these are the eigenvalues of $\boldsymbol{A}_P - \boldsymbol{b}_P \boldsymbol{f}^t$ and $\boldsymbol{A}_P - \boldsymbol{\ell} \boldsymbol{c}_P^t$.

Now let us look at this from the point of view of Corollary 10.29. Instead of thinking of the plant as a SISO linear system Σ_P, let us think of it as a rational function $R_P(s) = T_{\Sigma_P}(s) = \frac{1}{ms^2}$. This is not, of course, a BIBO stable transfer function. However, if we use the controller rational function

$$R_C(s) = T_{\Sigma_C}(s) = \frac{4m(2s+1)}{s^2 + 4s + 8},$$

then we are guaranteed by Corollary 10.29 that the closed-loop configuration of Figure 10.2 is IBIBO stable. In Figure 10.3 is shown the Nyquist plot for $R_L = R_C R_P$. Note that the

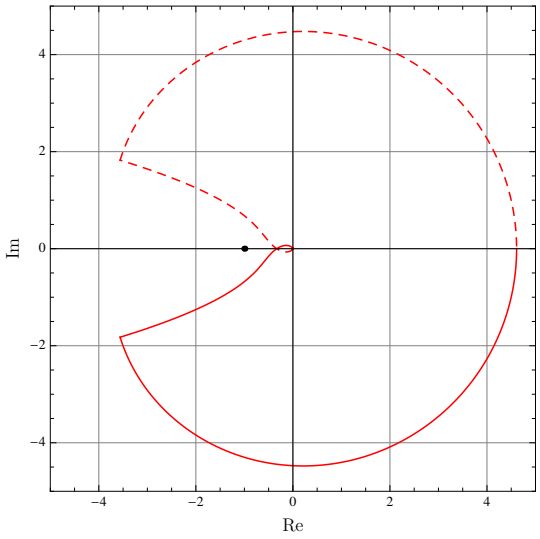


Figure 10.3 Nyquist plot for dynamic output feedback problem

Nyquist criterion is indeed satisfied. However, one could certainly make the point that the gain and phase margins could use improvement. This points out the general drawback of the purely algorithmic approach to controller design that is common to *all* of the algorithms we present. One should not rely on the algorithm to produce a satisfactory controller out of the box. The control designer will always be able to improve an initial design by employing the lessons only experience can teach. •

10.3 Parameterisation of stabilising dynamic output feedback controllers

The above discussion of construction stabilising controllers leads one to a consideration of whether it is possible to describe *all* stabilising controllers. The answer is that it is, and it is best executed in the rational function framework. We look at the block diagram configuration of Figure 10.4. We think of the plant transfer function R_P as being proper and

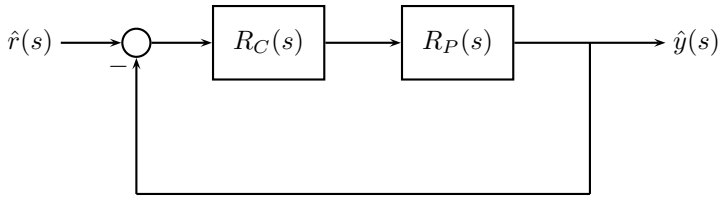


Figure 10.4 The block diagram configuration for the investigation of stabilising controller parameterisation

fixed. The objective is to find all controller transfer functions R_C so that the interconnection is IBIBO stable. This will happen when the transfer functions between all inputs and outputs are in RH_∞^+. The four relevant transfer functions are

$$\begin{aligned} T_1 &= \frac{1}{1+R_CR_P}, & T_2 &= \frac{R_C}{1+R_CR_P},\\ T_3 &= \frac{R_P}{1+R_CR_P}, & T_4 &= \frac{R_CR_P}{1+R_CR_P}. \end{aligned} \tag{10.10}$$

Let us first provide a useful properties rational functions in RH_∞^+.

10.3.1 More facts about RH_∞^+

It is useful to investigate in more detail some algebraic properties of RH_∞^+. These will be useful in this section, and again in Chapter 15. Many of our results in this section may be found in [Fuhrmann 2012].

For a rational function $R \in \mathbb{R}(s)$, the c.f.r. is a representation of R by the quotient of coprime polynomials where the denominator polynomial is monic. Let us look at another way to represent rational functions. We shall say that $R_1, R_2 \in \mathrm{RH}_\infty^+$ are ***coprime*** if they have no common zeros in $\overline{\mathbb{C}}_+$ and if at least one of them is not strictly proper.

10.31 Definition A ***coprime fractional representative*** of $R \in \mathbb{R}(s)$ is a pair (R_N, R_D) with the following properties:

(i) $R_N, R_D \in \mathrm{RH}_\infty^+$;

(ii) R_D and R_N are coprime in RH_∞^+;

(iii) $R = \frac{R_N}{R_D}$. •

The following simple result indicates that any rational function has a coprime fractional representative.

10.32 Proposition *If $R \in \mathbb{R}(s)$ then R has a coprime fractional representative.*

Proof Let (N, D) be the c.f.r. of R and let $k = \max\{\deg(D), \deg(N)\}$. Then (R_N, R_D) is a coprime fractional representative where

$$R_N(s) = \frac{N(s)}{(s+1)^k}, \quad R_D(s) = \frac{D(s)}{(s+1)^k}. \qquad \blacksquare$$

Note that unlike the c.f.r., there is no unique coprime fractional representative. However, it will be useful for us to come up with a particular coprime fractional representative. Given a polynomial $P \in \mathbb{R}[s]$ we may factor it as $P(s) = P_-(s)P_+(s)$ where all roots of P_- lie in $\mathbb{C}_-$ and all roots of P_+ lie on $\overline{\mathbb{C}}_+$. This factorisation is unique except in a trivial way; the coefficient of the highest power of s may be distributed between P_- and P_+ in an arbitrary way. Now let (N, D) be the c.f.r. for $R \in \mathbb{R}(s)$. We then have

$$R(s) = \frac{N(s)}{D(s)} = \frac{N_-(s)N_+(s)}{D(s)} = \frac{N_-(s)(s+1)^{\ell+k}}{D(s)} \frac{N_+(s)}{(s+1)^{\ell+k}} \tag{10.11}$$

where $\ell = \deg(N_+)$ and k is the relative degree of (N, D). Note that $\frac{N_-(s)(s+1)^{\ell+k}}{D(s)} \in \mathrm{RH}_\infty^+$ and that $\frac{D(s)}{N_-(s)(s+1)^{\ell+k}} \in \mathrm{RH}_\infty^+$. Generally, if $Q \in \mathrm{RH}_\infty^+$ and $Q^{-1} \in \mathrm{RH}_\infty^+$, then we say that Q is ***invertible*** in RH_∞^+. The formula (10.11) then says that any rational function $R \in \mathrm{RH}_\infty^+$ is the product of a function invertible in RH_∞^+, and a function in RH_∞^+ all of whose zeros lie in $\overline{\mathbb{C}}_+$.

The following result introduces the notion of the "coprime factorisation." This will play for us an essential rôle in determining useful representations for stabilising controllers.

10.33 Theorem *Rational functions $R_1, R_2 \in \mathrm{RH}_\infty^+$ are coprime if and only if there exists $\rho_1, \rho_2 \in \mathrm{RH}_\infty^+$ so that*

$$\rho_1 R_1 + \rho_2 R_2 = 1. \tag{10.12}$$

*We call (ρ_1, ρ_2) a **coprime factorisation** of (R_1, R_2).*

Proof First suppose that $\rho_1, \rho_2 \in \mathrm{RH}_\infty^+$ exist as stated. Clearly, if $z \in \overline{\mathbb{C}}_+$ is a zero of, say, R_1 is cannot also be a zero of R_2 as this would contradict (10.12). What's more, if both R_1 and R_2 are strictly proper then we have

$$\lim_{s\to\infty} \big(\rho_1(s)R_1(s) + \rho_2(s)R_2(s)\big) = 0,$$

again in contradiction with (10.12).

Now suppose that $R_1, R_2 \in \mathrm{RH}_\infty^+$ are coprime and suppose that R_1 is not strictly proper. Let (N_1, D_1) and (N_2, D_2) be the c.f.r.'s for R_1 and R_2. Denote $\sigma = (s+1)$, let $\ell_j = \deg(N_{j,+})$, and let k_j be the relative degree of (N_j, D_j), $j = 1, 2$. Thus $k_1 = 0$. Write

$$R_j = \tilde{R}_j \frac{N_{j,+}}{\sigma^{\ell_j+k_j}}, \quad j = 1, 2,$$

where $\tilde{R}_j$, $j = 1, 2$, is invertible in RH_∞^+. Suppose that $(\tilde{\rho}_1, \tilde{\rho}_2)$ are a coprime factorisation of $(\frac{N_{1,+}}{\sigma^{\ell_1+k_1}}, \frac{N_{2,+}}{\sigma^{\ell_2+k_2}})$. Then

$$\begin{aligned} &\tilde{\rho}_1 \frac{N_{1,+}}{\sigma^{\ell_1}} + \tilde{\rho}_2 \frac{N_{2,+}}{\sigma^{\ell_2+k_2}} = 1 \\ \implies \quad &\tilde{\rho}_1 \tilde{R}_1^{-1} \tilde{R}_1 \frac{N_{1,+}}{\sigma^{\ell_1}} + \tilde{\rho}_2 \tilde{R}_2^{-1} \tilde{R}_2 \frac{N_{2,+}}{\sigma^{\ell_2+k_2}} = 1. \end{aligned}$$

Since $\tilde{R}_1$ and $\tilde{R}_2$ are invertible in RH_∞^+ this shows that $(\tilde{\rho}_1\tilde{R}_1^{-1}, \tilde{\rho}_2\tilde{R}_2^{-1})$ is a coprime factorisation of (R_1, R_2). Thus we can assume, without loss of generality, that R_1 and R_2 are of the form

$$R_1 = \frac{P_1}{\sigma^{\ell_1}}, \quad R_2 = \frac{P_2}{\sigma^{\ell_2+k}},$$

where the coprime polynomials P_1 and P_2 have all roots in $\overline{\mathbb{C}}_+$, $\deg(P_j) = \ell_j$, $j = 1, 2$, and $\ell_2 \leq \ell_1$. By Lemma C.4 we may find polynomials $Q_1, Q_2 \in \mathbb{R}[s]$ so that

$$Q_1P_1 + Q_2P_2 = \sigma^{\ell_1+\ell_2+k},$$

and with $\deg(Q_2) < \deg(P_1)$. Thus we have

$$\frac{Q_1}{\sigma^{\ell_2+k}}\frac{P_1}{\sigma^{\ell_1}} + \frac{Q_2}{\sigma^{\ell_1}}\frac{P_2}{\sigma^{\ell_2+k}} = 1. \tag{10.13}$$

Since $R_2 \in \mathrm{RH}_\infty^+$, $\frac{P_2}{\sigma^{\ell_2+k}}$ is proper. Since $\deg(Q_2) < \deg(P_1)$, $\frac{Q_2}{\sigma^{\ell_1}}$ is strictly proper. Therefore, the second term in (10.13) is strictly proper. Since $\frac{P_1}{\sigma^{\ell_1}}$ is not strictly proper, it then follows that $\frac{Q_1}{\sigma^{\ell_2+k}}$ is also not strictly proper since we must have

$$\lim_{s\to\infty}\left(\frac{Q_1}{\sigma^{\ell_2+k}}\frac{P_1}{\sigma^{\ell_1}} + \frac{Q_2}{\sigma^{\ell_1}}\frac{P_2}{\sigma^{\ell_2+k}}\right) = 1$$

and

$$\lim_{s\to\infty}\left(\frac{Q_2}{\sigma^{\ell_1}}\frac{P_2}{\sigma^{\ell_2+k}}\right) = 0.$$

Therefore, if we take

$$\rho_1 = \frac{Q_1}{\sigma^{\ell_2+k}}, \quad \rho_2 = \frac{Q_2}{\sigma^{\ell_1}},$$

we see that the conditions of the theorem are satisfied. ■

The matter of determining rational functions ρ_1 and ρ_2 in the lemma is not necessarily straightforward. However, in the next section we shall demonstrate a way to do this, at least if one can find a single stabilising controller. The following corollary, derived directly from the computations of the proof of Theorem 10.33, declares the existence of a particular coprime factorisation that will be helpful in the course of the proof of Theorem 10.37.

10.34 Corollary *If $R_1, R_2 \in \mathrm{RH}_\infty^+$ are coprime with R_1 not strictly proper, then there exists a coprime factorisation (ρ_1, ρ_2) of (R_1, R_2) having the property that ρ_2 is strictly proper.*

Proof As in the initial part of the proof of Theorem 10.33, let us write

$$R_1 = \tilde{R}_1\frac{P_1}{\sigma^{\ell_1}}, \quad R_2 = \tilde{R}_2\frac{P_2}{\sigma^{\ell_2+k}},$$

where $\tilde{R}_1$ and $\tilde{R}_2$ are invertible in RH_∞^+ and where $\ell_j = \deg(P_j)$, $j = 1, 2$. In the proof of Theorem 10.33 were constructed $\tilde{\rho}_1$ and $\tilde{\rho}_2$ (in the proof of Theorem 10.33, these are the final ρ_1 and ρ_2) with the property that

$$\tilde{\rho}_1\frac{P_1}{\sigma^{\ell_1}} + \tilde{\rho}_2\frac{P_2}{\sigma^{\ell_2+k}} = 1$$

and ρ_2 is strictly proper. Now we note that if $\rho_j = \tilde{R}_j^{-1}\tilde{\rho}_j$, $j = 1, 2$, then (ρ_1, ρ_2) is a coprime factorisation of (R_1, R_2) and that ρ_2 is strictly proper since $\tilde{R}_2^{-1} \in \mathrm{RH}_\infty^+$, and so is proper. ■

10.3.2 The Youla parameterisation

Now we can use Theorem 10.33 to ensure a means of parameterising stabilising controllers by a single function in RH_∞^+. Before we begin, let us establish some notation that we will use to provide an important preliminary results. We let $R_P \in \mathbb{R}(s)$ be proper with coprime fractional representative (P_1, P_2), and let (ρ_1, ρ_2) be a coprime factorisation for (P_1, P_2). We call $\theta \in \mathrm{RH}_\infty^+$ ***admissible*** for the coprime fractional representative (P_1, P_2) and the coprime factorisation (ρ_1, ρ_2) if

1. $\theta \neq \dfrac{\rho_2}{P_1}$, and
2. $\lim\limits_{s\to\infty} \big(\rho_2(s) - \theta(s)P_1(s)\big) \neq 0$.

Now we define

$$\mathscr{S}_{\mathrm{pr}}(P_1, P_2, \rho_1, \rho_2) = \{\frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1} \mid \theta \text{ admissible}\}.$$

At this point, this set depends on the choice of coprime factorisation (ρ_1, ρ_2). The following lemma indicates that the set is, in fact, independent of this factorisation.

10.35 Lemma *Let $R_P \in \mathbb{R}(s)$ be proper with coprime fractional representative (P_1, P_2). If (ρ_1, ρ_2) and $(\tilde\rho_1, \tilde\rho_2)$ are coprime factorisations for (P_1, P_2), then the map*

$$\frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1} \mapsto \frac{\tilde\rho_1 + \tilde\theta(\theta) P_2}{\tilde\rho_2 - \tilde\theta(\theta) P_1}$$

from $\mathscr{S}_{\mathrm{pr}}(P_1, P_2, \rho_1, \rho_2)$ to $\mathscr{S}_{\mathrm{pr}}(P_1, P_2, \tilde\rho_1, \tilde\rho_2)$ is a bijection if

$$\tilde\theta(\theta) = \theta + \rho_1\tilde\rho_2 - \rho_2\tilde\rho_1.$$

Proof First note that $\tilde\theta(\theta) \in \mathrm{RH}_\infty^+$ so the map is well-defined. To see that the map is a bijection, it suffices to check that the map

$$\frac{\tilde\rho_1 + \tilde\theta P_2}{\tilde\rho_2 - \tilde\theta P_1} \mapsto \frac{\rho_1 + \theta(\tilde\theta) P_2}{\rho_2 - \theta(\tilde\theta) P_1}$$

is its inverse provided that

$$\theta(\tilde\theta) = \tilde\theta + \tilde\rho_1\rho_2 - \tilde\rho_2\rho_1.$$

This is a straightforward, if slightly tedious, computation. ■

The upshot of the lemma is that the set $\mathscr{S}_{\mathrm{pr}}(P_1, P_2, \rho_1, \rho_2)$ is independent of ρ_1 and ρ_2. Thus let us denote it by $\mathscr{S}_{\mathrm{pr}}(P_1, P_2)$. Now let us verify that this set is in fact only dependent on R_P, and not on the coprime fractional representative of R_P.

10.36 Lemma *If $R_P \in \mathbb{R}(s)$ is proper, and if (P_1, P_2) and $(\tilde P_1, \tilde P_2)$ are coprime fractional representatives of R_P, then $\mathscr{S}_{\mathrm{pr}}(P_1, P_2) = \mathscr{S}_{\mathrm{pr}}(\tilde P_1, \tilde P_2)$.*

Proof Since we have

$$R_P = \frac{P_1}{P_2} = \frac{\tilde P_1}{\tilde P_2}$$

it follows that $\tilde{P}_j = UP_j$ for an invertible $U \in \mathrm{RH}_\infty^+$. If (ρ_1, ρ_2) is a coprime factorisation of (P_1, P_2) it then follows that $(U^{-1}\rho_1, U^{-1}\rho_2)$ is a coprime factorisation of $(\tilde{P}_1, \tilde{P}_2)$. We then have

$$\begin{aligned}\mathscr{S}_{\mathrm{pr}}(\tilde{P}_1, \tilde{P}_2) &= \{\frac{\tilde{\rho}_1 + \tilde{\theta}\tilde{P}_2}{\tilde{\rho}_2 - \tilde{\theta}\tilde{P}_1} \mid \theta \text{ admissible}\} \\ &= \{\frac{U^{-1}\rho_1 + \tilde{\theta}UP_2}{U^{-1}\rho_2 - \tilde{\theta}UP_1} \mid \tilde{\theta} \text{ admissible}\} \\ &= \{\frac{\rho_1 + \tilde{\theta}U^2P_2}{\rho_2 - \tilde{\theta}U^2P_1} \mid \tilde{\theta} \text{ admissible}\} \\ &= \{\frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1} \mid \theta \text{ admissible}\} \\ &= \mathscr{S}_{\mathrm{pr}}(P_1, P_2),\end{aligned}$$

as desired. ■

Now we are permitted to denote $\mathscr{S}_{\mathrm{pr}}(P_1, P_2)$ simply by $\mathscr{S}_{\mathrm{pr}}(R_P)$ since it indeed only depends on the plant transfer function. With this notation we state the following famous result due to [Youla, Jabr, and Bongiorno 1976], stating that $\mathscr{S}_{\mathrm{pr}}(R_P)$ is exactly the set of proper stabilising controllers. Thus, in particular, $\mathscr{S}_{\mathrm{pr}}(R_P) \subset \mathscr{S}(R_P)$.

10.37 Theorem *Consider the block diagram configuration of Figure 10.4 and suppose that R_P is proper. For the plant rational function R_P, let (P_1, P_2) be a coprime fractional representative with (ρ_1, ρ_2) a coprime factorisation of (P_1, P_2):*

$$\rho_1 P_1 + \rho_2 P_2 = 1. \tag{10.14}$$

Then there is a one-to-one correspondence between the set of proper rational functions $R_C \in \mathbb{R}(s)$ that render the interconnection IBIBO stable and the set

$$\mathscr{S}_{\mathrm{pr}}(R_P) = \{\frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1} \mid \theta \text{ admissible}\}. \tag{10.15}$$

Furthermore, if R_P is strictly proper, every $\theta \in \mathrm{RH}_\infty^+$ is admissible.

Proof Let us first translate the conditions (10.10) into conditions on coprime fractional representatives for R_P and R_C. Let (P_1, P_2) be a coprime fractional representative of R_P as in the statement of the theorem. Also denote a coprime fractional representative of R_C as (C_1, C_2). Then the four transfer functions of (10.10) are computed to be

$$\begin{aligned}T_1 &= \frac{C_2P_2}{C_1P_1 + C_2P_2}, \quad T_2 = \frac{C_1P_2}{C_1P_1 + C_2P_2}, \\ T_3 &= \frac{C_2P_1}{C_1P_1 + C_2P_2}, \quad T_4 = \frac{C_1P_1}{C_1P_1 + C_2P_2}.\end{aligned} \tag{10.16}$$

We are thus charged with showing that these four functions are in RH_∞^+.

Now let $\theta \in \mathrm{RH}_\infty^+$ and let R_C be the corresponding rational function defined by (10.15). Let

$$C_1 = \rho_1 + \theta P_2, \quad C_2 = \rho_2 - \theta P_1.$$

We claim that (C_1, C_2) is a coprime fractional representative of R_C. This requires us to show that $C_1, C_2 \in \mathrm{RH}_\infty^+$, that the functions have no common zeros in $\overline{\mathbb{C}}_+$, and that at least one

of them is not strictly proper. That $C_1, C_2 \in \mathrm{RH}_\infty^+$ follows since $\theta, \rho_1, \rho_2, P_1, P_2 \in \mathrm{RH}_\infty^+$. A direct computation, using (10.14), shows that

$$C_1P_1 + C_2P_2 = 1. \tag{10.17}$$

From this it follows that C_1 and C_2 have no common zeros. Finally, we shall show that R_C is proper. By Lemma 10.35 we may freely choose the coprime factorisation (ρ_1, ρ_2). By Corollary 10.34 we choose (ρ_1, ρ_2) so that ρ_1 is strictly proper. Since

$$\lim_{s\to\infty} \big(\rho_1(s)P_1(s) + \rho_2(s)P_2(s)\big) = 1,$$

it follows that ρ_2 is not strictly proper. Therefore, $C_2 = \rho_2 - \theta P_1$ is also not strictly proper, provided that θ is admissible.

Now consider the case when R_P is proper (i.e., the final assertion of the theorem). Note that if R_P is strictly proper, then so is P_1. Condition 1 of the definition of admissibility then follows since if $\theta = \frac{\rho_2}{P_1}$, then θ would necessarily be improper. Similarly, if P_1 is strictly proper, it follows that $\lim_{s\to\infty} C_2(s) = \rho_2(s) \neq 0$. Thus condition 2 for admissibility holds.

Now suppose that $R_C \in \mathbb{R}(s)$ stabilises the closed-loop system so that the four transfer functions (10.10) are in RH_∞^+. Let (C_1, C_2) be a coprime fractional representative of R_C. Let $D = C_1P_1 + C_2P_2$. We claim that D and $\frac{1}{D}$ are in RH_∞^+. Clearly $D \in \mathrm{RH}_\infty^+$. Also, if $\alpha_1, \alpha_2 \in \mathrm{RH}_\infty^+$ have the property that

$$\alpha_1C_1 + \alpha_2C_2 = 1,$$

(such functions exist by Theorem 10.33), then we have

$$\begin{aligned}\frac{1}{D} &= \frac{(\alpha_1C_1 + \alpha_2C_2)(\rho_1P_1 + \rho_2P_2)}{D}\\ &= \alpha_1\rho_1T_4 + \alpha_1\rho_2T_2 + \alpha_2\rho_1T_3 + \alpha_2\rho_2T_1.\end{aligned}$$

By the assumption that the transfer functions T_1, T_2, T_3, and T_4 are all in RH_∞^+, it follows that $\frac{1}{D} \in \mathrm{RH}_\infty^+$. Thus D is proper and not strictly proper so that we may define a new coprime fractional representative for R_C by $(\frac{C_1}{D}, \frac{C_2}{D})$ so that

$$C_1P_1 + C_2P_2 = 1.$$

We therefore have

$$\begin{bmatrix}\rho_1 & \rho_2\\ C_1 & C_2\end{bmatrix}\begin{bmatrix}P_1\\ P_2\end{bmatrix} = \begin{bmatrix}1\\ 1\end{bmatrix}$$
$$\implies \quad \theta\begin{bmatrix}P_1\\ P_2\end{bmatrix} = \begin{bmatrix}-C_2 & \rho_2\\ C_1 & -\rho_1\end{bmatrix}\begin{bmatrix}1\\ 1\end{bmatrix},$$

if we take $\theta = \rho_2C_1 - \rho_1C_2 \in \mathrm{RH}_\infty^+$. It therefore follows that

$$C_1 = \rho_1 + \theta P_2, \quad C_2 = \rho_2 - \theta P_1,$$

as desired. ■

10.38 Remarks

1. This is clearly an interesting result as it allows us the opportunity to write down all proper stabilising controllers in terms of a single parameter $\theta \in \mathrm{RH}_\infty^+$. Note that there is a correspondence between proper controllers and those obtained in the dynamic output feedback setting. Thus the previous result might be thought of as capturing all the stabilising dynamics output feedback controllers.
2. Some authors say that *all* stabilising controllers are obtained by the Youla parameterisation. This is not quite correct.
3. In the case that R_P is not strictly proper, one should check that all admissible θ's give loop gains $R_L = R_C R_P$ that are well-posed in the unity gain feedback configuration of Figure 10.4. This is done in Exercise E10.15. •

Things are problematic at the moment because we are required to determine $\rho_1, \rho_2 \in \mathrm{RH}_\infty^+$ with the property that (10.14) holds. This may not be trivial. However, let us indicate a possible method for determining ρ_1 and ρ_2. First let us show that the parameter $\theta \in \mathrm{RH}_\infty^+$ uniquely determines the controller.

10.39 Proposition *Let R_P be a strictly proper rational function with (P_1, P_2) a coprime fractional representative. Let $\rho_1, \rho_2 \in \mathrm{RH}_\infty^+$ have the property that*

$$\rho_1 P_1 + \rho_2 P_2 = 1.$$

Then the map

$$\theta \mapsto \frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1}$$

from RH_∞^+ into the set of stabilising controllers for the block diagram configuration of Figure 10.4 is injective.

Proof Let Φ be the indicated map from RH_∞^+ into the set of stabilising controllers, and suppose that $\Phi(\theta_1) = \Phi(\theta_2)$. That is,

$$\frac{\rho_1 + \theta_1 P_2}{\rho_2 - \theta_1 P_1} = \frac{\rho_1 + \theta_2 P_2}{\rho_2 - \theta_2 P_1}.$$

This implies that

$$\begin{aligned} & -\theta_1\theta_2 P_1 P_2 + \theta_1 P_2 \rho_2 - \theta_2 P_1 \rho_1 + \rho_1\rho_2 = -\theta_1\theta_2 P_1 P_2 - \theta_1 P_1 \rho_1 + \theta_2 P_2 \rho_2 + \rho_1\rho_2 \\ \implies \quad & \theta_1(\rho_1 P_1 + \rho_2 P_2) = \theta_2(\rho_1 P_1 + \rho_2 P_2) \\ \implies \quad & \theta_1 = \theta_2, \end{aligned}$$

as desired. ■

Let us now see how to employ a given stabilising controller to determine ρ_1 and ρ_2.

10.40 Proposition *Let R_P be a strictly proper rational function with (P_1, P_2) a coprime fractional representative. If R_C is a stabilising controller for the block diagram configuration of Figure 10.4, define ρ_1 and ρ_2 by*

check

$$\rho_1 = \frac{R_C}{P_2 + P_1 R_C}, \quad \rho_2 = \frac{1}{P_2 + P_1 R_C}.$$

Then the following statements hold:

(i) $\rho_1, \rho_2 \in \mathrm{RH}_\infty^+$;
(ii) $\rho_1 P_1 + \rho_2 P_2 = 1$.

Proof (i) Let (C_1, C_2) be a coprime fractional representative for R_C. Then

$$\begin{aligned}\rho_1 &= \frac{\frac{C_1}{C_2}}{P_2 + P_1\frac{C_1}{C_2}}\\ &= \frac{C_1}{C_1P_1 + C_2P_2}.\end{aligned}$$

As in the proof of Theorem 10.37, it follows then that if $D = C_1P_1 + C_2P_2$ then $D, \frac{1}{D} \in \mathrm{RH}_\infty^+$. Since $C_1 \in \mathrm{RH}_\infty^+$, it follows that $\rho_1 \in \mathrm{RH}_\infty^+$. A similar computation gives $\rho_2 \in \mathrm{RH}_\infty^+$.

(ii) This is a direct computation. ■

This result allows us to compute ρ_1 and ρ_2 given a coprime fractional representative for a plant transfer function. This allows us to produce the following algorithm for parameterising the set of stabilising controllers.

10.41 Algorithm for parameterisation of stabilising controllers Given a proper plant transfer function R_P perform the following steps.

1. Determine a coprime fractional representative (P_1, P_2) for R_P using Proposition 10.32.
2. Construct the canonical minimal realisation $\Sigma_P = (\boldsymbol{A}_P, \boldsymbol{b}_P, \boldsymbol{c}_P^t, \boldsymbol{D}_P)$ for R_P.
3. Using Proposition 10.13, construct $\boldsymbol{f} \in \mathbb{R}^n$ so that $\boldsymbol{A}_P - \boldsymbol{b}_P\boldsymbol{f}^t$ is Hurwitz.
4. Using Corollary 10.17 construct $\boldsymbol{\ell} \in \mathbb{R}^n$ so that $\boldsymbol{A}_P - \boldsymbol{\ell}\boldsymbol{c}_P^t$ is Hurwitz. Note that this amounts to performing the construction of Proposition 10.13 with $\boldsymbol{A} = \boldsymbol{A}_P^t$ and $\boldsymbol{b} = \boldsymbol{c}$.
5. Using Theorem 10.27 define a stabilising controller SISO linear system $\Sigma_C = (\boldsymbol{A}_C, \boldsymbol{b}_C, \boldsymbol{c}_C^t, \boldsymbol{D}_C)$.
6. Define the stabilising controller rational function $R_C = T_{\Sigma_C}$.
7. Determine $\rho_1, \rho_2 \in \mathrm{RH}_\infty^+$ using Proposition 10.40.
8. The set of all stabilising controllers is now given by

$$\mathscr{S}_{\mathrm{pr}}(R_P) = \{\frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1} \mid \theta \text{ admissible}\}.$$ •

Let us carry this out for an example.

10.42 Example **(Example 6.59 cont'd)** We return to the example where $R_P = \frac{1}{ms^2}$, and perform the above steps.

1. A coprime fractional representative of R_P is given by (P_1, P_2) where

$$P_1(s) = \frac{1/m}{(s+1)^2}, \quad P_2(s) = \frac{s^2}{(s+1)^2}.$$

2. As in Example 6.59, we have

$$\boldsymbol{A}_P = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}, \quad \boldsymbol{b}_P = \begin{bmatrix}0\\ \frac{1}{m}\end{bmatrix}, \quad \boldsymbol{c}_P = \begin{bmatrix}1\\ 0\end{bmatrix}, \quad \boldsymbol{D}_P = \boldsymbol{0}_1.$$

3. In Example 10.30 we computed
$$\boldsymbol{f} = \begin{bmatrix} 2m \\ 2m \end{bmatrix},$$
4. In Example 10.30 we also computed
$$\boldsymbol{\ell} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$
5. Using Example 10.30 we have
$$\boldsymbol{A}_C = \begin{bmatrix} -2 & 1 \\ -4 & -2 \end{bmatrix} \boldsymbol{b}_C = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \boldsymbol{c}_C^t = \begin{bmatrix} 2m & 2m \end{bmatrix}, \quad \boldsymbol{D}_C = \boldsymbol{0}_1.$$
6. In Example 10.30 we computed
$$R_C(s) = \frac{4m(2s+1)}{s^2+4s+8}.$$
7. From Proposition 10.40 we calculate, after some simplification,
$$\rho_1 = \frac{(s+1)^2(s^2+4s+8)}{(s^2+2s+2)^2}, \quad \rho_2 = \frac{4m(s+1)^2(2s+1)}{(s^2+2s+2)^2}.$$
8. Finally, after simplification, we see that the set of stabilising controllers is given by
$$\mathscr{S}_{\text{pr}}(R_P) = \{\frac{m\big(4m(s+1)^4(2s+1) + \theta(s)s^2((s^2+2s+2)^2)\big)}{m(s+1)^4(s^2+4s+8) - \theta(s)((s^2+2s+2)^2)} \mid \theta \in \text{RH}_\infty^+\}.$$

This is a somewhat complicated expression. It can be simplified by using simpler expressions for ρ_1 and ρ_2. In computing ρ_1 and ρ_2 above, we have merely applied our rule verbatim. Indeed, simpler functions that also satisfy $\rho_1 P_1 + \rho_2 P_2 = 1$ are

finish

$$\rho_1(s) = 1, \quad \rho_2(s) = m.$$

With these functions we compute the set of stabilising controllers to be

$$\{\frac{m\big(\theta(s)s^2 - m(s^2+1)\big)}{\theta + m(s^2+1)} \mid \theta \in \text{RH}_\infty^+\},$$

which is somewhat more pleasing than our expression derived using our rules. •

10.4 Strongly stabilising controllers

In the previous section we expressed all controllers that stabilise a given plant. Of course, one will typically not want to allow *any* form for the controller. For example, one might wish for the controller transfer function to itself be stable. This is not always possible, and in this section we explore this matter.

10.5 State estimation

One of the problems with using static or dynamic state feedback is that the assumption that one knows the value of the state is typically over-optimistic. Indeed, in practice, it is often the case that one can at best only know the value of the output at any given time. Therefore, what one would like to be able to do is infer the value of the state from the knowledge of the output. A moments reflection should suggest that this is possible for observable systems. A further moments reflection should lead one to allow the possibility for detectable systems. These speculations are correct, and lead to the theory of observers that we introduce in this section.

10.5.1 Observers

Let us give a rough idea of what we mean by an observer before we get to formal definitions. Suppose that we have a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ evolving with its usual differential equations:

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t) \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t) + \boldsymbol{D}u(t). \end{aligned} \tag{10.18}$$

An ***observer*** should take as input the original input $u(t)$, along with the measured output $y(t)$. Using these inputs, the observer constructs an estimate for the state, and we denote this estimate by $\hat{\boldsymbol{x}}(t)$. In Figure 10.5 we schematically show how an observer works. Our

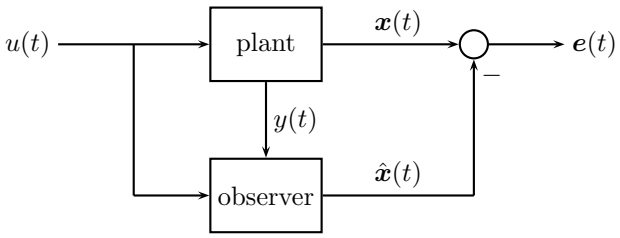


Figure 10.5 A schematic for an observer using the error

first result shows that an observer exists, although we will not use the given observer in practice.

10.43 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be an observable SISO linear system satisfying (10.18). There exists $\boldsymbol{o}_o(s), \boldsymbol{o}_i(s) \in \mathbb{R}[s]^{n\times 1}$ with the property that*

$$\boldsymbol{x}(t) = \boldsymbol{o}_{\mathrm{o}}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)y(t) + \boldsymbol{o}_{\mathrm{i}}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)u(t).$$

Proof In (10.18), differentiate $y(t)$ $n-1$ times successively with respect to t, and use the equation for $\dot{\boldsymbol{x}}(t)$ to get

$$\begin{bmatrix} y(t) \\ y^{(1)}(t) \\ y^{(2)}(t) \\ \vdots \\ y^{(n-1)}(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}^t \\ \boldsymbol{c}^t\boldsymbol{A} \\ \boldsymbol{c}^t\boldsymbol{A}^2 \\ \vdots \\ \boldsymbol{c}^t\boldsymbol{A}^{n-1} \end{bmatrix} \boldsymbol{x}(t) + \begin{bmatrix} \boldsymbol{D} & 0 & \cdots & 0 & 0 \\ \boldsymbol{c}^t\boldsymbol{b} & \boldsymbol{D} & \cdots & 0 & 0 \\ \boldsymbol{c}^t\boldsymbol{A}\boldsymbol{b} & \boldsymbol{c}^t\boldsymbol{b} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{c}^t\boldsymbol{A}^{n-2}\boldsymbol{b} & \boldsymbol{c}^t\boldsymbol{A}^{n-3}\boldsymbol{b} & \cdots & \boldsymbol{c}^t\boldsymbol{b} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} u(t) \\ u^{(1)}(t) \\ u^{(2)}(t) \\ \vdots \\ u^{(n-1)}(t). \end{bmatrix}$$

Since Σ is observable, $\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})$ is invertible, and so we have

$$\boldsymbol{x}(t) = \boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})^{-1}\begin{bmatrix} y(t) \\ y^{(1)}(t) \\ y^{(2)}(t) \\ \vdots \\ y^{(n-1)}(t) \end{bmatrix} - \boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})^{-1}\begin{bmatrix} \boldsymbol{D} & 0 & \cdots & 0 & 0 \\ \boldsymbol{c}^t\boldsymbol{b} & \boldsymbol{D} & \cdots & 0 & 0 \\ \boldsymbol{c}^t\boldsymbol{A}\boldsymbol{b} & \boldsymbol{c}^t\boldsymbol{b} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{c}^t\boldsymbol{A}^{n-2}\boldsymbol{b} & \boldsymbol{c}^t\boldsymbol{A}^{n-3}\boldsymbol{b} & \cdots & \boldsymbol{c}^t\boldsymbol{b} & \boldsymbol{D} \end{bmatrix}\begin{bmatrix} u(t) \\ u^{(1)}(t) \\ u^{(2)}(t) \\ \vdots \\ u^{(n-1)}(t), \end{bmatrix}$$

which proves the proposition if we take

$$\boldsymbol{o}_{\text{o}} = \boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})^{-1}\begin{bmatrix} 1 \\ s \\ s^2 \\ \vdots \\ s^{n-1} \end{bmatrix}, \quad \boldsymbol{o}_{\text{i}} = -\boldsymbol{O}(\boldsymbol{A},\boldsymbol{c})^{-1}\begin{bmatrix} \boldsymbol{D} & 0 & \cdots & 0 & 0 \\ \boldsymbol{c}^t\boldsymbol{b} & \boldsymbol{D} & \cdots & 0 & 0 \\ \boldsymbol{c}^t\boldsymbol{A}\boldsymbol{b} & \boldsymbol{c}^t\boldsymbol{b} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{c}^t\boldsymbol{A}^{n-2}\boldsymbol{b} & \boldsymbol{c}^t\boldsymbol{A}^{n-3}\boldsymbol{b} & \cdots & \boldsymbol{c}^t\boldsymbol{b} & \boldsymbol{D} \end{bmatrix}\begin{bmatrix} 1 \\ s \\ s^2 \\ \vdots \\ s^{n-1} \end{bmatrix}. \qquad \blacksquare$$

While the above result does indeed prove the existence of an observer which *exactly* reproduces the state given the output and the input, it suffers from repeatedly differentiating the measured output, and in practice this produces undesirable noise. To circumvent these problems, in the next section we introduce an observer that does not exactly measure the state. Indeed, it is an ***asymptotic observer***, meaning that the ***error*** $\boldsymbol{e}(t) = \boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)$ satisfies $\lim_{t\to\infty} \boldsymbol{e}(t) = \boldsymbol{0}$.

10.5.2 Luenberger observers

The error schematic in Figure 10.5 has the feature that it is driven using the error $\boldsymbol{e}(t)$. The asymptotic observer we construct in this section instead uses the so-called ***innovation*** defined as $i(t) = y(t) - \hat{y}(t)$ where $\hat{y}(t)$ is the estimated output $\hat{y}(t) = \boldsymbol{c}^t\hat{\boldsymbol{x}}(t) + \boldsymbol{D}u(t)$. The schematic for the sort of observer is shown in Figure 10.6. Note that the inputs to the

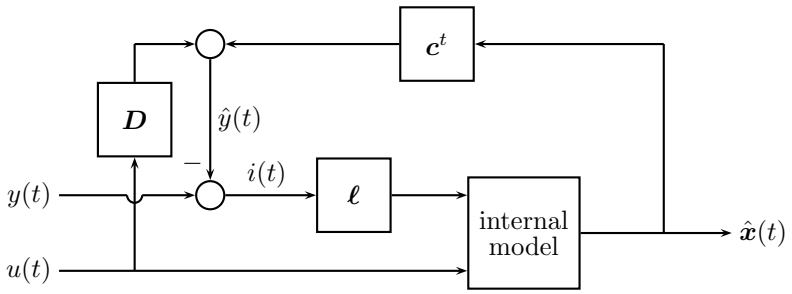


Figure 10.6 The schematic for a Luenberger observer

observer are the actual measured output $y(t)$ and the actual input $u(t)$, and that the output is the estimated state $\hat{\boldsymbol{x}}(t)$. The vector $\boldsymbol{\ell} \in \mathbb{R}^n$ we call the ***observer gain vector***. There is a gap in the schematic, and that is the "internal model." We use as the internal model the original system model, but now with the inputs as specified in the schematic. Thus we take the estimated state to satisfy the equation

internal model principle?

$$\begin{aligned} \dot{\hat{\boldsymbol{x}}}(t) &= \boldsymbol{A}\hat{\boldsymbol{x}}(t) + \boldsymbol{b}u(t) + \boldsymbol{\ell}i(t) \\ \hat{y}(t) &= \boldsymbol{c}^t\hat{\boldsymbol{x}}(t) + \boldsymbol{D}u(t) \\ i(t) &= y(t) - \hat{y}(t). \end{aligned} \tag{10.19}$$

From this equation, the following lemma gives the error $\boldsymbol{e}(t) = \boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)$.

10.44 Lemma *If $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is a SISO linear system and if $\hat{\boldsymbol{x}}(t)$, $\hat{y}(t)$, and $i(t)$ satisfy (10.19), then $\boldsymbol{e}(t) = \hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)$ satisfies the differential equation*

$$\dot{\boldsymbol{e}}(t) = (\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t)\boldsymbol{e}(t).$$

Proof Subtracting

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t), \quad \dot{\hat{\boldsymbol{x}}}(t) = \boldsymbol{A}\hat{\boldsymbol{x}}(t) + \boldsymbol{b}u(t) + \boldsymbol{\ell}i(t),$$

and using the second and third of equations (10.19) we get

$$\begin{aligned}\dot{\boldsymbol{e}}(t) &= \boldsymbol{A}\boldsymbol{e}(t) - \boldsymbol{\ell}i(t)\\ &= \boldsymbol{A}\boldsymbol{e}(t) - \boldsymbol{\ell}(y(t) - \hat{y}(t))\\ &= \boldsymbol{A}\boldsymbol{e}(t) - \boldsymbol{\ell}\boldsymbol{c}^t(\boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t))\\ &= \boldsymbol{A}\boldsymbol{e}(t) - \boldsymbol{\ell}\boldsymbol{c}^t\boldsymbol{e}(t),\end{aligned}$$

as desired. ■

The lemma now tells us that we can make our Luenberger observer an asymptotic observer provided we choose $\boldsymbol{\ell}$ so that $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz. This is very much like the Ackermann pole placement problem, and indeed can be proved along similar lines, giving the following result.

10.45 Proposition *Let $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be an observable SISO linear system and suppose that the characteristic polynomial for $\boldsymbol{A}$ is*

$$P_{\boldsymbol{A}}(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0.$$

Let $P \in \mathbb{R}[s]$ be monic and degree n. The observer gain vector $\boldsymbol{\ell}$ defined by

$$\boldsymbol{\ell} = P(\boldsymbol{A})(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))^{-1}\begin{bmatrix}0\\ \cdots\\ 0\\ 1\end{bmatrix}$$

has the property that the characteristic polynomial of the matrix $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is P.

Proof Note that observability of $(\boldsymbol{A}, \boldsymbol{c})$ is equivalent to controllability of $(\boldsymbol{A}^t, \boldsymbol{c})$. Therefore, by Proposition 10.13 we know that if

$$\boldsymbol{\ell}^t = \begin{bmatrix}0 & \cdots & 0 & 1\end{bmatrix}(\boldsymbol{C}(\boldsymbol{A}^t, \boldsymbol{c}))^{-1}P(\boldsymbol{A}^t), \tag{10.20}$$

then the matrix $\boldsymbol{A}^t - \boldsymbol{c}\boldsymbol{\ell}^t$ has characteristic polynomial P. The result now follows by taking the transpose of equation (10.20), and noting that the characteristic polynomial of $\boldsymbol{A}^t - \boldsymbol{c}\boldsymbol{\ell}^t$ is equal to that of $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$. ■

10.46 Remarks

1. As expected, the construction of the observer gain vector is accomplished along the same lines as the static state feedback vector as in Proposition 10.13. Indeed, the observer gain vector is obtained by using the formula of Proposition 10.13 with $\boldsymbol{A}^t$ in place of $\boldsymbol{A}$, and with $\boldsymbol{c}$ in place of $\boldsymbol{b}$. This is another example of the "duality" between controllability and observability.
2. The eigenvalues of $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ are called the ***observer poles*** for the given Luenberger observer.
3. The notion of an observer is lurking in the proof of Theorem 10.27. This is flushed out in Section 10.5.3.
4. We are, of course, interested in choosing the observer gain vector $\boldsymbol{\ell}$ so that $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz. This can be done if Σ is observable, or more generally, detectable. To this end, let us denote by $\mathscr{D}(\Sigma)$ those observer gain vectors for which $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz. •

Let us illustrate this with an example.

10.47 Example We consider the SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ with

$$\boldsymbol{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Suppose that we wish a closed-loop characteristic polynomial of $P(s) = s^2 + 4s + 4$. We compute

$$P(\boldsymbol{A}) = \boldsymbol{A}^2 + 2\boldsymbol{A} + 2\boldsymbol{I}_2 = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}, \boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then we have

$$\boldsymbol{\ell} = P(\boldsymbol{A}(\boldsymbol{O}(\boldsymbol{A}, \boldsymbol{c}))^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

giving

$$\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t = \begin{bmatrix} 0 & -2 \\ 1 & -2 \end{bmatrix},$$

which has the desired characteristic polynomial.

Now let us see how the observer does at observing. In Figure 10.7 we show the results of a simulation of equations (10.18) and (10.19) with

$$\boldsymbol{b} = (0, 1), \quad \boldsymbol{D} = \boldsymbol{0}_1, \quad \boldsymbol{x}(0) = (1, 1), \quad \hat{\boldsymbol{x}}(0) = \boldsymbol{0}, \quad u(t) = 1(t).$$

Note that the error is decaying to zero exponentially as expected. •

10.5.3 Static state feedback, Luenberger observers, and dynamic output feedback

In this section we bring together the ideas of static state feedback, Luenberger observers, and dynamic output feedback. It is by no means obvious that these should all be tied together, but indeed they are. To make the connection we make the obvious observation that if one does not possess accurate knowledge of the state, then static state feedback seems a somewhat optimistic means of designing a controller. However, if one uses an observer to estimate the state, one can use the estimated state for static state feedback. The schematic is depicted in Figure 10.8. If the observer is a Luenberger observer satisfying (10.19) and if

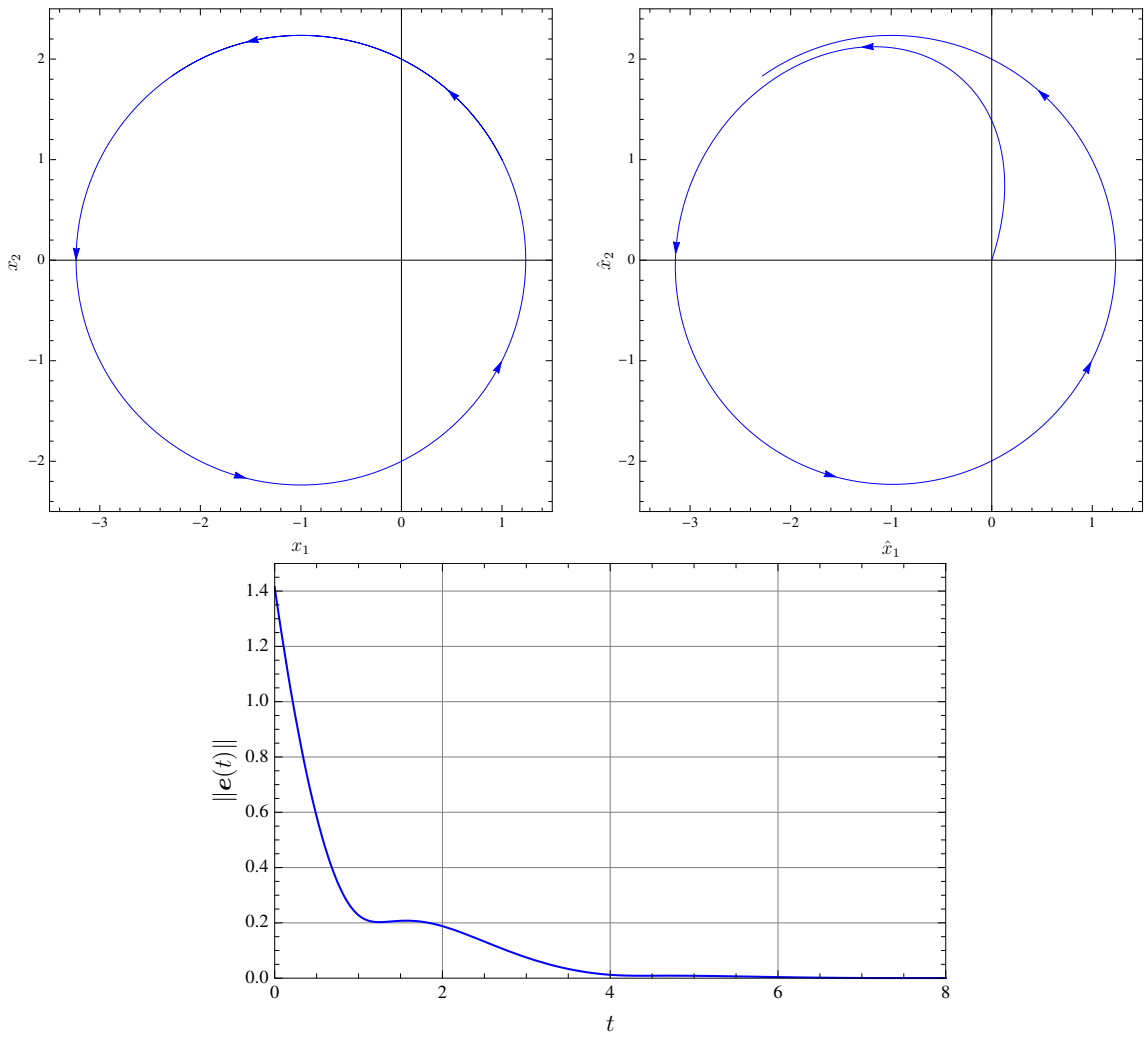


Figure 10.7 The state (top left), the estimated state (top right), and the norm of the error for a Luenberger observer

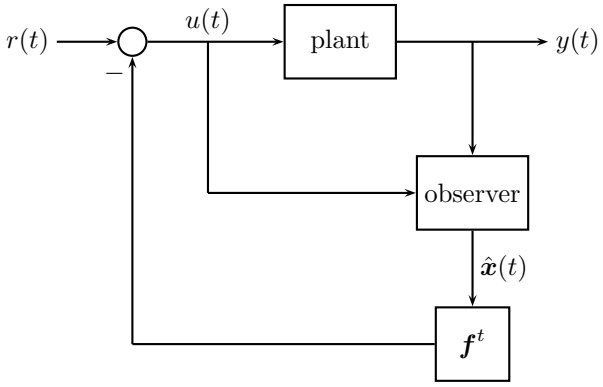


Figure 10.8 Static state feedback using the estimated state

the plant is the usual SISO state representation satisfying (10.18), then one may verify that the equations governing the interconnection of Figure 10.8 are

$$\begin{aligned}\begin{bmatrix}\dot{\boldsymbol{x}}(t)\\ \dot{\hat{\boldsymbol{x}}}(t)\end{bmatrix} &= \begin{bmatrix}\boldsymbol{A} & -\boldsymbol{b}\boldsymbol{f}^t\\ \boldsymbol{\ell}\boldsymbol{c}^t & \boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t-\boldsymbol{\ell}\boldsymbol{c}^t\end{bmatrix}\begin{bmatrix}\boldsymbol{x}(t)\\ \hat{\boldsymbol{x}}(t)\end{bmatrix} + \begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{b}\end{bmatrix} r(t)\\ y(t) &= \begin{bmatrix}\boldsymbol{c}^t & -\boldsymbol{D}\boldsymbol{f}^t\end{bmatrix}\begin{bmatrix}\boldsymbol{x}(t)\\ \hat{\boldsymbol{x}}(t)\end{bmatrix} + \boldsymbol{D}r(t).\end{aligned} \tag{10.21}$$

The next result records the characteristic polynomial and the closed-loop transfer function for these equations.

10.48 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system with $\boldsymbol{f} \in \mathbb{R}^n$ a static state feedback vector and $\boldsymbol{\ell} \in \mathbb{R}^n$ an observer gain vector for a Luenberger observer (10.19). Suppose that the observer is combined with state feedback as in Figure 10.8, giving the closed-loop equations (10.21). Also consider the interconnection of Figure 10.9 where*

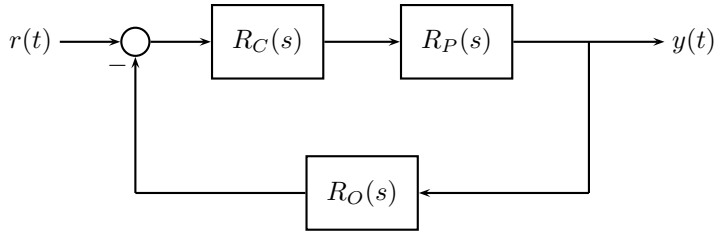


Figure 10.9 A dynamic output feedback loop giving the closed-loop characteristic polynomial of static state feedback using a Luenberger observer to estimate the state

$$\begin{aligned}R_C(s) &= \frac{\det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)}{\det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{b}\boldsymbol{f}^t + \boldsymbol{\ell}(\boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t))}\\ R_P(s) &= \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} + \boldsymbol{D}\\ R_O(s) &= \frac{\boldsymbol{f}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{\ell}}{\det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)}.\end{aligned}$$

Then the following statements hold:

(i) the closed-loop characteristic polynomial for (10.21) is the product of the characteristic polynomials of $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ and $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$.

(ii) the closed-loop system for (10.21) (i.e., the transfer function from r to y) is the same transfer function for the interconnection of Figure 10.9, and what's more, both transfer functions are exactly the transfer function for $(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}, \boldsymbol{b}, \boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t, \boldsymbol{0}_1)$.

Proof (i) The "$\boldsymbol{A}$-matrix" for the observer/feedback system (10.21) is

$$\boldsymbol{A}_{\mathrm{cl}} = \begin{bmatrix}\boldsymbol{A} & -\boldsymbol{b}\boldsymbol{f}^t\\ \boldsymbol{\ell}\boldsymbol{c}^t & \boldsymbol{A}-\boldsymbol{b}\boldsymbol{f}^t-\boldsymbol{\ell}\boldsymbol{c}^t\end{bmatrix}.$$

Let us define an invertible $2n \times 2n$ matrix

$$\begin{aligned}\boldsymbol{T} &= \begin{bmatrix}\boldsymbol{I}_n & -\boldsymbol{I}_n\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix}\\ \Longrightarrow \quad \boldsymbol{T}^{-1} &= \begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{I}_n\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix}.\end{aligned} \tag{10.22}$$

It is a simple computation to show that

$$\boldsymbol{T}\boldsymbol{A}_{\mathrm{cl}}\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t & \boldsymbol{0} \\ \boldsymbol{\ell}\boldsymbol{c}^t & \boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t \end{bmatrix}.$$

Thus we see that the characteristic polynomial of $\boldsymbol{T}\boldsymbol{A}_{\mathrm{cl}}\boldsymbol{T}^{-1}$, and therefore the characteristic polynomial of $\boldsymbol{A}_{\mathrm{cl}}$, is indeed the product of the characteristic polynomials of $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ and $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$, as claimed.

(ii) Let us use the new coordinates corresponding to the change of basis matrix $\boldsymbol{T}$ defined in (10.22). Thus we take

$$\begin{aligned} \bar{\boldsymbol{A}}_{\mathrm{cl}} &= \begin{bmatrix} \boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t & \boldsymbol{0} \\ \boldsymbol{\ell}\boldsymbol{c}^t & \boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t \end{bmatrix} \\ \bar{\boldsymbol{b}}_{\mathrm{cl}} &= \boldsymbol{T}\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{b} \end{bmatrix} \\ \bar{\boldsymbol{c}}^t_{\mathrm{cl}} &= \begin{bmatrix} \boldsymbol{c}^t & -\boldsymbol{D}\boldsymbol{f}^t \end{bmatrix}\boldsymbol{T}^{-1} = \begin{bmatrix} \boldsymbol{c}^t & \boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t \end{bmatrix} \\ \bar{\boldsymbol{D}}_{\mathrm{cl}} &= \boldsymbol{D}, \end{aligned}$$

and determine the transfer function for $\bar{\Sigma} = (\bar{\boldsymbol{A}}_{\mathrm{cl}}, \bar{\boldsymbol{b}}_{\mathrm{cl}}, \bar{\boldsymbol{c}}^t_{\mathrm{cl}}, \bar{\boldsymbol{D}}_{\mathrm{cl}})$. We have

$$(s\boldsymbol{I}_{2n} - \bar{\boldsymbol{A}}_{\mathrm{cl}})^{-1} = \begin{bmatrix} (s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1} & \boldsymbol{0} \\ * & (s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{b}\boldsymbol{f}^t)^{-1} \end{bmatrix},$$

where "$*$" denotes a term whose exact form is not relevant. One then readily computes

$$T_{\bar{\Sigma}}(s) = \bar{\boldsymbol{c}}^t_{\mathrm{cl}}(s\boldsymbol{I}_{2n} - \bar{\boldsymbol{A}}_{\mathrm{cl}})^{-1}\bar{\boldsymbol{b}}_{\mathrm{cl}} = (\boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t)(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{b}\boldsymbol{f}^t)^{-1}\boldsymbol{b}.$$

From this it follows that the transfer function of the closed-loop system (10.21) is as claimed.

Now we determine this same transfer function in a different way, proceeding directly from the closed-loop equations (10.21). First let us define

$$\begin{aligned} N_C(s) &= \det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t), & D_C(s) &= \det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{b}\boldsymbol{f}^t + \boldsymbol{\ell}(\boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t)) \\ N_P(s) &= \boldsymbol{D}\det(s\boldsymbol{I}_n - \boldsymbol{A}) + \boldsymbol{c}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}, & D_P(s) &= \det(s\boldsymbol{I}_n - \boldsymbol{A}) \\ N_O(s) &= \boldsymbol{f}^t\mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{\ell}, & D_O(s) &= \det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t). \end{aligned}$$

Now, the equation governing the observer states is

$$\dot{\hat{\boldsymbol{x}}}(t) = (\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t)\hat{\boldsymbol{x}}(t) + (\boldsymbol{b} - \boldsymbol{D}\boldsymbol{\ell})u(t) + \boldsymbol{\ell}y(t).$$

Taking left causal Laplace transforms gives

$$\hat{\boldsymbol{x}}(s) = (s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}\big((\boldsymbol{b} - \boldsymbol{D}\boldsymbol{\ell})\hat{u}(s) + \boldsymbol{\ell}\hat{y}(s)\big).$$

Let us define

$$\boldsymbol{T}_1(s) = (s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}(\boldsymbol{b} - \boldsymbol{D}\boldsymbol{\ell}), \quad \boldsymbol{T}_2(s) = (s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}\boldsymbol{\ell}$$

so that

$$\hat{\boldsymbol{x}}(s) = \boldsymbol{T}_1(s)\hat{u}(s) + \boldsymbol{T}_2(s)\hat{y}(s).$$

Given that $u(t) = r(t) - \boldsymbol{f}^t\hat{\boldsymbol{x}}(t)$, we have

$$\begin{aligned} & \hat{u}(s) = \hat{r}(s) - \boldsymbol{f}^t\boldsymbol{T}_1(s)\hat{u}(s) - \boldsymbol{f}^t\boldsymbol{T}_1(s)\hat{y}(s) \\ \Longrightarrow \quad & \hat{u}(s) = -\frac{\boldsymbol{f}^t\boldsymbol{T}_2(s)}{1+\boldsymbol{f}^t\boldsymbol{T}_1(s)}\hat{y}(s) + \frac{1}{1+\boldsymbol{f}^t\boldsymbol{T}_1(s)}\hat{r}(s). \end{aligned}$$

From the proof of Lemma A.3 we have that

$$\begin{aligned} 1 + \boldsymbol{f}^t\boldsymbol{T}_1(s) &= 1 + \boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}(\boldsymbol{b} - \boldsymbol{D}\boldsymbol{\ell}) \\ &= \det\left[1 + \boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}(\boldsymbol{b} - \boldsymbol{D}\boldsymbol{\ell})\right] \\ &= \det(\boldsymbol{I}_n + (s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}(\boldsymbol{b} - \boldsymbol{D}\boldsymbol{\ell})\boldsymbol{f}^t) \\ &= \det\big((s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}(\boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t) + \boldsymbol{b}\boldsymbol{f}^t\big) \\ &= \frac{\det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}(\boldsymbol{c}^t - \boldsymbol{D}\boldsymbol{f}^t) + \boldsymbol{b}\boldsymbol{f}^t)}{\det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)} \\ &= \frac{D_C(s)}{N_C(s)}. \end{aligned}$$

We also clearly have

$$\begin{aligned} \boldsymbol{f}^t\boldsymbol{T}_2(s) &= \boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}\boldsymbol{\ell} \\ &= \frac{\boldsymbol{f}^t\mathrm{adj}(\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)\boldsymbol{\ell}}{\det(\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)} \\ &= \frac{\boldsymbol{f}^t\mathrm{adj}(\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{\ell}}{\det(\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)} \\ &= \frac{N_O(s)}{D_O(s)}, \end{aligned}$$

where we have used Lemma A.5. Thus we have

$$\hat{u}(s) = \frac{N_O(s)}{D_O(s)}\frac{N_C(s)}{D_C(s)}\hat{y}(s) + \frac{N_C(s)}{D_C(s)}\hat{r}(s). \tag{10.23}$$

Also noting that

$$\hat{y}(s) = \frac{N_P(s)}{D_P(s)}\hat{u}(s), \tag{10.24}$$

we may eliminate $\hat{r}(s)$ from equations (10.23) and (10.24) to get

$$T_{\bar{\Sigma}}(s) = \frac{\hat{y}(s)}{\hat{r}(s)} = \frac{N_C(s)N_P(s)}{D_C(s)D_P(s) + N_P(s)N_O(s)}, \tag{10.25}$$

if we use the fact that $D_O = N_C$.

Let us now turn to the transfer function of the interconnection of Figure 10.9. The transfer function may be computed in the usual manner using Mason's Rule:

$$\frac{\hat{y}(s)}{\hat{r}(s)} = \frac{R_C(s)R_P(s)}{1 + R_C(s)R_P(s)R_O(s)} = \frac{D_O(s)N_C(s)N_P(s)}{D_C(s)D_P(s)D_O(s) + N_C(s)N_P(s)N_O(s)}.$$

Since $D_O = N_C$ this may be simplified to

$$\frac{\hat{y}(s)}{\hat{r}(s)} = \frac{N_C(s)N_P(s)}{D_C(s)D_P(s) + N_P(s)N_O(s)}.$$

Comparing this to (10.25), we see that indeed the transfer function of Figure 10.9 is the same as that of the closed-loop system (10.21). ■

10.49 Remarks

1. One of the consequences of part (i) of the theorem is that one can separately choose the observer poles and the controller poles. This is a phenomenon that goes under the name of the ***separation principle***.
2. The change of coordinates represented by the matrix defined in (10.22) can easily be seen as making a change of coordinates from $(\boldsymbol{x}, \hat{\boldsymbol{x}})$ to $(\boldsymbol{e}, \boldsymbol{x})$. It is not surprising, then, that in these coordinates we should see that the characteristic polynomial gets represented as a product of two characteristic polynomials. Indeed, we have designed the observer so that the error should be governed by the matrix $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$, and we have designed the state feedback so that the state should be governed by the matrix $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$.
3. Note that for the interconnection of Figure 10.9 there must be massive cancellation in the transfer function since the system nominally has $3n$ states, but the denominator of the transfer function has degree n. Indeed, one can see that in the loop gain there is directly a cancellation of a factor $\det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)$ from the numerator of R_C and the denominator of R_O. What is not so obvious at first glance is that there is an additional cancellation of another factor $\det(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)$ that happens when forming the closed-loop transfer function. Note that these cancellations are all stable provided one chooses $\boldsymbol{\ell}$ so that $\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t$ is Hurwitz. Thus they need not be disastrous. That there should be this cancellation in the closed-loop equations (10.21) is not surprising since the control does not affect the error. Thus the closed-loop system with the observer is not controllable (also see Exercise E10.23). One should also be careful that the characteristic polynomial for the closed-loop system (10.21) is different from that for the interconnection Figure 10.9. •

Let us illustrate Theorem 10.48 via an example.

10.50 Example **(Example 10.47 cont'd)** Recall that we had

$$\boldsymbol{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Suppose that we want both observer poles and controller poles to be roots of the polynomial $P(s) = s^2 + 4s + 4$. In Example 10.47 we found the required observer gain vector to be $\boldsymbol{\ell} = (1, 2)$. To find the static state feedback vector we employ Ackermann's formula from Proposition 10.13. The controllability matrix is

$$\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Thus we compute

$$\boldsymbol{f}^t = \begin{bmatrix} 0 & 1 \end{bmatrix} (\boldsymbol{C}(\boldsymbol{A}, \boldsymbol{b}))^{-1} P(\boldsymbol{A}) = \begin{bmatrix} -1 & 2 \end{bmatrix}.$$

This gives the closed-loop system matrix

$$\boldsymbol{A}_{\text{cl}} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & -2 \\ 0 & 1 & 0 & -2 \\ 0 & 2 & 2 & -4 \end{bmatrix}$$

for the interconnections of Figure 10.8. For the interconnection of Figure 10.8 we determine that the closed-loop input and output vectors are

$$\boldsymbol{b}_{\text{cl}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c}_{\text{cl}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

In Figure 10.10 we show the state $\boldsymbol{x}(t)$ and the estimated state $\hat{\boldsymbol{x}}(t)$ for the initial conditions

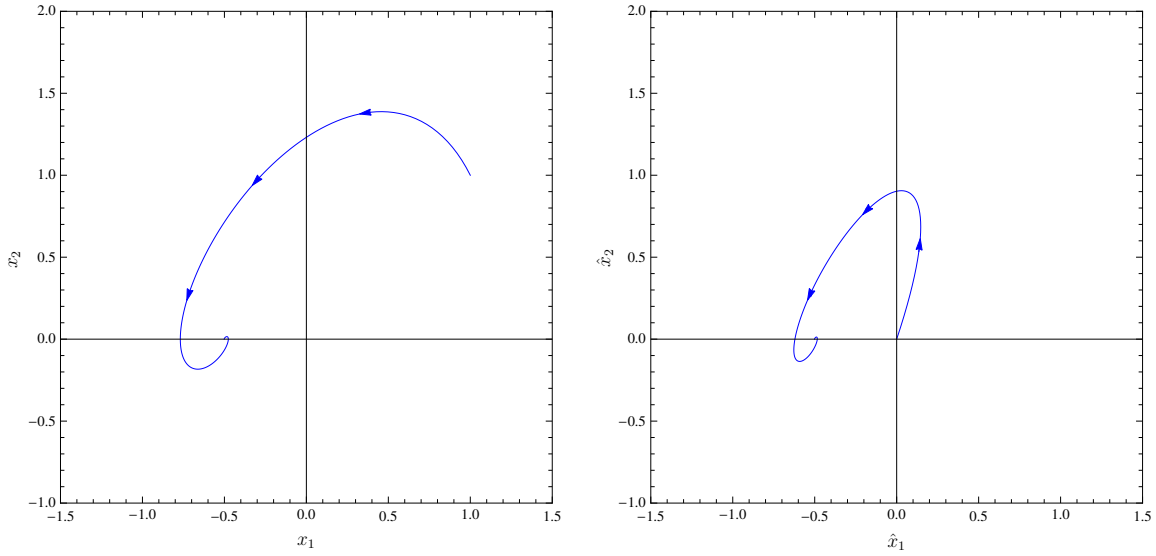


Figure 10.10 The state (left) and the estimated state (right) for the closed-loop system

$\boldsymbol{x}(0) = (1, 1)$ and $\hat{\boldsymbol{x}}(0) = (0, 0)$, and for $u(t) = 1(t)$. Note that the quantities approach the same value as $t \to \infty$, as should be the case as the observer is an asymptotic observer. In Figure 10.11 we show the output for the interconnection of Figure 10.8. •

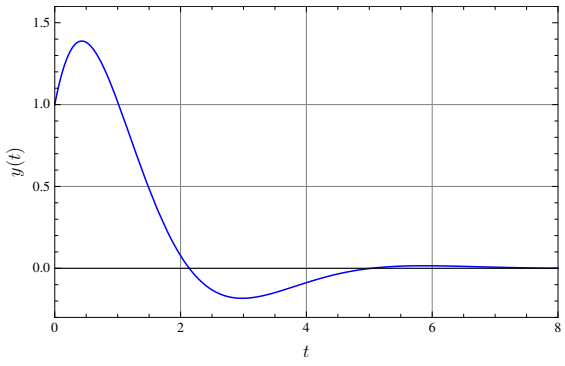


Figure 10.11 The output for the interconnection of Figure 10.8

10.6 Summary

1. Stabilisability and detectability extend the notions of controllability and observability. The extensions are along the lines of asking that the state behaviour that you cannot control or observe is stable.
2. Ackermann's formula is available as a means to place the poles for a SISO linear system in a desired location.

Exercises

E10.1 Show that $(\boldsymbol{A}, \boldsymbol{v})$ is stabilisable if and only if $(\boldsymbol{A}^t, \boldsymbol{v})$ is detectable.

E10.2 Show that $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is stabilisable if and only if the matrix

$$\left[\begin{array}{c|c} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \end{array}\right]$$

has rank n for all $s \in \overline{\mathbb{C}}_+$.

E10.3 Show that $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ is detectable if and only if the matrix

$$\begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} \\ \boldsymbol{c}^t \end{bmatrix}$$

has rank n for all $s \in \overline{\mathbb{C}}_+$.

E10.4 If $(\boldsymbol{A}, \boldsymbol{c})$ is detectable and if $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ is positive-semidefinite and satisfies

$$\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{c}\boldsymbol{c}^t, \tag{E10.1}$$

show that $\boldsymbol{A}$ is Hurwitz.
Hint: *Show that (E10.1) implies that*

$$\mathbf{P} = \mathrm{e}^{\mathbf{A}^t t}\mathbf{P}\mathrm{e}^{\mathbf{A}t} + \int_0^t \mathrm{e}^{\mathbf{A}^t\tau}\mathbf{c}\mathbf{c}^t\mathrm{e}^{\mathbf{A}\tau}\,\mathrm{d}\tau.$$

In the next exercise, we will introduce the notion of a ***linear matrix inequality*** (***LMI***). Such a relation is, in general, a matrix equation, invariant under transposition (i.e., if one takes the matrix transpose of the equation, it remains the same), for an unknown matrix. Since the equation is invariant under transposition, it makes sense to demand that the unknown matrix render the equation positive or negative-definite or semidefinite. In recent years, LMI's have become increasingly important in control theory. A survey is [El Ghaoui and Niculescu 1987]. The reason for the importance of LMI's is one can often determine their solvability using "convexity methods." These are often numerically tractable. This idea forms the backbone, for example, of the approach to robust control taken by Dullerud and Paganini [1999].

E10.5 Consider a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$.

(a) Use the Liapunov methods of Section 5.4 to show that Σ is stabilisable if and only if there exists $\boldsymbol{f} \in \mathbb{R}^n$ so that the linear matrix inequality

$$(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)\boldsymbol{P} + \boldsymbol{P}(\boldsymbol{A}^t - \boldsymbol{f}\boldsymbol{b}^t) < 0$$

has a solution $\boldsymbol{P} > 0$.

(b) Use your result from part (a) to prove the following theorem.

Theorem *For a stabilisable SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$, a state feedback vector $\boldsymbol{f} \in \mathbb{R}^n$ stabilises the closed-loop system if and only if there exists $\boldsymbol{P} > 0$ and $\boldsymbol{g} \in \mathbb{R}^n$ so that*

(i) $\boldsymbol{f} = \boldsymbol{P}^{-1}\boldsymbol{g}$ *and*

(ii) so that the LMI

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{b}\end{bmatrix}\begin{bmatrix}\boldsymbol{P} \\ \boldsymbol{g}^t\end{bmatrix} + \begin{bmatrix}\boldsymbol{P} & \boldsymbol{g}\end{bmatrix}\begin{bmatrix}\boldsymbol{A}^t \\ \boldsymbol{b}^t\end{bmatrix} < 0$$

is satisfied.

The point of the exercise is that the LMI gives a way of parameterising all stabilising state feedback vectors.

E10.6 Consider the two polynomials that have a common factor in $\overline{\mathbb{C}}_+$:

$$P_1(s) = s^2 - 1, \quad P_2(s) = s^3 + s^2 - s - 1.$$

Construct three SISO linear systems $\Sigma_i = (\boldsymbol{A}_i, \boldsymbol{b}_i, \boldsymbol{c}_i^t, \boldsymbol{0}_1)$, $i = 1, 2, 3$, with the following properties:

1. $T_{\Sigma_i} = \frac{P_1}{P_2}$, $i = 1, 2, 3$;
2. Σ_1 is stabilisable but not detectable;
3. Σ_2 is detectable but not stabilisable;
4. Σ_3 is neither stabilisable nor detectable.

E10.7 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ be a SISO linear system.

(a) If Σ is stabilisable show how, using Proposition 10.13, to construct a stabilising state feedback vector $\boldsymbol{f}$ in the case when Σ is not controllable.

(b) Show that if Σ is not stabilisable then it is not possible to construct a stabilising state feedback vector.

E10.8 Let $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ be controllable and let $P \in \mathbb{R}[s]$ be a monic polynomial of degree n. Show that there exists a *unique* static state feedback vector $\boldsymbol{f} \in \mathbb{R}^n$ with the property that $P(s) = \det(s\boldsymbol{I}_n - (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t))$.
Hint: *Refer to Exercise E2.38.*

E10.9 Suppose that $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n$ is controllable, and let $P \in \mathbb{R}[s]$ be a monic polynomial of degree $n - 1$.

(a) Show that there exists a static state feedback vector $\boldsymbol{f} \in \mathbb{R}^n$ with the following properties:

1. $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ possesses an invariant $(n-1)$-dimensional subspace $V \subset \mathbb{R}^n$ with the property that for each basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-1}\}$ for V, $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-1}, \boldsymbol{b}\}$ is a basis for $\mathbb{R}^n$;
2. The characteristic polynomial of $(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)|V$ is P.

(b) For V as in part (a), define a linear map $A_V \colon V \to V$ by $A_V(\boldsymbol{v}) = \mathrm{pr}_V \circ \boldsymbol{A}(\boldsymbol{v})$, where $\mathrm{pr}_V \colon \mathbb{R}^n \to V$ is defined by

$$\mathrm{pr}_V(a_1\boldsymbol{v}_1 + \cdots + a_{n-1}\boldsymbol{v}_{n-1} + a_n\boldsymbol{b}) = a_1\boldsymbol{v}_1 + \cdots + a_{n-1}\boldsymbol{v}_{n-1},$$

for $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-1}\}$ a basis for V.

E10.10 Consider the pendulum/cart system of Exercises E1.5. In this problem, we shall change the input for the system from a force applied to the cart, to a force applied to the mass on the end of the pendulum that is tangential to the pendulum motion; see Figure E10.1. For the following cases,

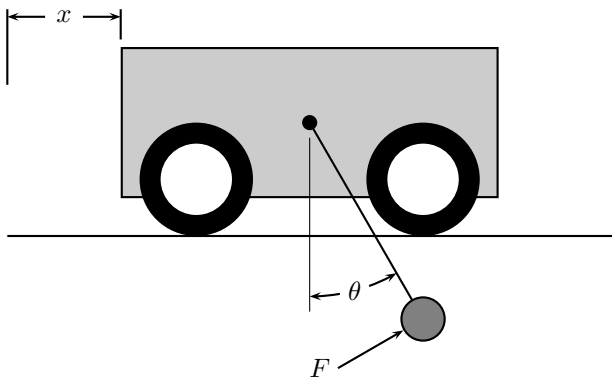


Figure E10.1 The pendulum/cart with an alternate input

(a) the equilibrium point $(0, 0)$ with cart position as output;
(b) the equilibrium point $(0, 0)$ with cart velocity as output;
(c) the equilibrium point $(0, 0)$ with pendulum angle as output;
(d) the equilibrium point $(0, 0)$ with pendulum angular velocity as output;
(e) the equilibrium point $(0, \pi)$ with cart position as output;
(f) the equilibrium point $(0, \pi)$ with cart velocity as output;
(g) the equilibrium point $(0, \pi)$ with pendulum angle as output;
(h) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output,

do the following:

1. obtain the linearisation $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ of the system;
2. obtain the transfer function for Σ;
3. determine whether the system is observable and/or controllable;
4. determine whether the system is detectable and/or stabilisable.

E10.11 Consider the pendulum/cart system of Exercises E1.5 and E2.4. Construct a state feedback vector $\boldsymbol{f}$ that makes the linearisation of pendulum/cart system stable about the equilibrium point with the pendulum pointing "up." Verify that the closed-loop system has the eigenvalues you asked for.

E10.12 Consider the double pendulum system of Exercises E1.6 and E2.5. For the following equilibrium points and input configurations, construct a state feedback vector $\boldsymbol{f}$ that makes the linearisation of double pendulum about that equilibrium point stable:

(a) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;
(b) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;
(c) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;
(d) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;
(e) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;
(f) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input.

In each case, take the output to be the angle of the second link. Verify that the closed-loop system has the eigenvalues you asked for.

E10.13 Consider the coupled tank system of Exercises E1.11, E2.6. For the following equilibrium points and input configurations, construct a state feedback vector $\boldsymbol{f}$ that makes the linearisation of double pendulum about that equilibrium point stable:

(a) the output is the level in tank 1;
(b) the output is the level in tank 2;
(c) the output is the difference in the levels.

In each case, take the output to be the angle of the second link. Verify that the closed-loop system has the eigenvalues you asked for.

E10.14 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$ with

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Answer the following questions.

(a) Show that Σ is stabilisable by static output feedback only if all components of $\boldsymbol{c}$ are nonzero and of the same sign.
(b) If $\boldsymbol{A}$ is as given, and if all components of $\boldsymbol{c}$ are nonzero and of the same sign, is it true that Σ can be stabilised using static output feedback? If so, prove it. If not, find a counterexample.

E10.15 (a) Show that the closed-loop transfer function for any one of the stabilising controllers from Theorem 10.37 is an affine function of θ. That is to say, show that the closed-loop transfer function has the form $R_1\theta + R_2$ for rational functions R_1 and R_2.
(b) Show that in the expression $R_1\theta + R_2$ from part (a) that $R_1 \in \mathrm{RH}_\infty^+$, and is strictly proper if and only if R_P is strictly proper.
(c) Conclude that if

$$R_C = \frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1}$$

for an admissible θ then the unity gain feedback loop with $R_L = R_C R_P$ is well-posed.

E10.16 Let $R_P \in \mathrm{RH}_\infty^+$ be a stable plant transfer function. Show that the set of controllers R_C for which the closed-loop system in Figure 10.4 is given by

$$\{\frac{\theta}{1 - \theta R_P} \mid \theta \in \mathrm{RH}_\infty^+\}.$$

E10.17 Consider the plant transfer function $R_P(s) = \frac{1}{s^2}$, and consider the interconnection of Figure 10.4. Show that there is a rational function $R_C(s)$, not of the form given in Theorem 10.37, for which the closed-loop system is IBIBO stable.
Hint: *Consider PD control.*

E10.18 Exercise on restricted domain for poles. Finish

In the next exercise you will demonstrate what is a useful feature of RH_∞^+. Recall that a ***ring*** is a set R with the operations of addition and multiplication, and these operations should satisfy the "natural" properties associated with addition and multiplication (see [Lang 2005]). A ring is ***commutative*** when the operation of multiplication is commutative (as for example, with the integers). A commutative ring is an ***integral domain*** when the product of nonzero elements is nonzero. An ***ideal*** in a ring R is a subset I for which

1. $0 \in I$,
2. if $x, y \in I$ then $x + y \in I$, and
3. if $x \in I$ then $ax \in I$ and $xa \in I$ for all $a \in R$.

An ideal I is ***principal*** if it has the form

$$I = \{xa \mid a \in R\}.$$

The ideal above is said to be ***generated*** by x. An integral domain R is a ***principal ideal domain*** when every ideal I is principal.

E10.19 Answer the following questions.

(a) Show that RH_∞^+ is an integral domain.
(b) Show that RH_∞^+ is a principal ideal domain.
(c) Show that H_∞^+ is not a principal ideal domain.

Normally, only condition 1 is given in the definition of admissibility for a function $\theta \in \mathrm{RH}_\infty^+$ to parameterise $\mathscr{S}_{\mathrm{pr}}(R_P)$. This is a genuine omission of hypotheses. In the next exercise, you will show that, in general, the condition 2 also must be included.

E10.20 Consider the plant transfer function

$$R_P(s) = \frac{s}{s+1}.$$

For this plant, do the following.

(a) Show that $(P_1(s), P_2(s)) = (\frac{s}{s+1}, 1)$ is a coprime fractional representative for R_P.
(b) Show that

$$(\rho_1(s), \rho_2(s)) = \Big(-\frac{s+1}{(s+2)^2}, \frac{(s+1)(s+4)}{(s+2)^2}\Big)$$

is a coprime factorisation of (P_1, P_2).
(c) Define $\theta \in \mathrm{RH}_\infty^+$ by

$$\theta(s) = \frac{s}{s+1}.$$

Is θ admissible?
(d) For the function $\theta \in \mathrm{RH}_\infty^+$ from part (c), show that the controller

$$R_C = \frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1}$$

is improper. Thus for the conclusions of Theorem 10.37 to hold, in particular, for the controller parameterisation to be one for *proper* controller transfer functions, the condition 2 for admissibility is necessary.
(e) Show that despite our conclusions of part (d), the unity gain interconnection with loop gain $R_L = R_C R_P$, using R_C from part 2, is IBIBO stable.

E10.21 Consider the plant

$$R_P(s) = \frac{s}{(s+1)(s-1)}.$$

For this plant, do the following.

(a) Find a coprime fractional representative (P_1, P_2) for R_P (choose the obvious one, if you want to make life easy).

(b) Show that

$$R_C(s) = \frac{2(s+2)}{s - \frac{1}{2}}$$

stabilises R_P.

(c) Use the stabilising controller from the previous step to construct a coprime factorisation (ρ_1, ρ_2) for (P_1, P_2).

(d) Write the expression for the set of proper stabilising controllers for R_P depending on the parameter $\theta \in \mathrm{RH}_\infty^+$.

(e) For two different values of θ, produce the Nyquist plots for the loop gain $R_C R_P$ and comment on the properties of the controller you have chosen.

E10.22 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a SISO linear system.

(a) Show that for a Luenberger observer with observer gain vector $\boldsymbol{\ell}$ the state and the estimated state together satisfy the vector differential equation

$$\begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ \dot{\hat{\boldsymbol{x}}}(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{\ell}\boldsymbol{c}^t & \boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \hat{\boldsymbol{x}}(t) \end{bmatrix} + \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{b} \end{bmatrix} u(t).$$

(b) What is the differential equation governing the behaviour of $\boldsymbol{x}(t)$ and the error $\boldsymbol{e}(t) = \boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)$.

E10.23 Verify that the system (10.21) is not controllable.

Hint: *Consider the system in the basis defined by the change of basis matrix of (10.21).*

Chapter 11

Ad hoc methods I: The root-locus method

The root-locus method we study in this section was put forward in the papers of Evans (1948, 1950). The study is of roots of polynomials when the coefficients depend linearly on a parameter. In control systems, the parameter is typically the gain of a feedback loop, and our interest is in choosing the gain so that the closed-loop system is IBIBO stable. As we saw in Section 10.2.2, with static output feedback of SISO systems, there naturally arises a control problem where one has a polynomial with coefficients linear in a parameter, and stabilisation requires choosing this parameter so that the polynomial is Hurwitz. In Section 11.1.1 below, we discuss some control problems where this scenario arises. The manner of studying such problems in this chapter is to study how the roots move in the complex plane as functions of the parameter. That is to say, we look at the locus of all roots of the polynomial as the parameter varies, hence the name "root-locus."

In many introductory texts, one will find a laying out of a "design method" using root-locus methods. We do not devote significant effort to this for the reason that, according to Horowitz [1963], "It appears, therefore, that the root locus approach to the sensitivity problem is justified only in systems where there are few dominant poles and zeros." These systems are quite well understood in any case (see Section 13.2.3).

Contents

11.1 The root-locus problem, and its rôle in control

The aim in this section is to provide some situations in control where a certain type of problem involving a certain type of polynomial arises. Once this has been nailed down, we can talk in generality about this problem, and some of its broad properties. We reserve for Section 11.2 a more or less complete discussion of the properties of such polynomials.

11.1.1 A collection of problems in control

Our first task is to indicate that there are a collection of control problems which can be reduced to a problem of a certain type.

11.1 Examples

1. To get things rolling, let us consider the unity gain feedback loop of Figure 11.1 where

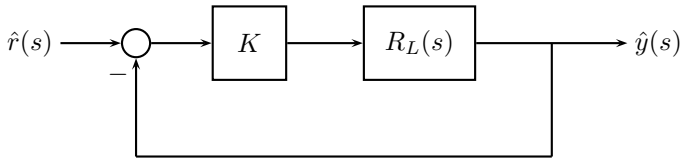


Figure 11.1 Root-locus from a negative feedback configuration

the loop gain is known up to a multiplicative constant K, and let us suppose that we ask that K be nonnegative. The transfer function from $\hat{r}(s)$ to $\hat{y}(s)$ is, as usual,

$$\frac{\hat{y}(s)}{\hat{r}(s)} = \frac{KR_L(s)}{1 + KR_L(s)}.$$

If R_L has the c.f.r. (N_L, D_L), then the characteristic polynomial for the interconnection is $D_L + KN_L$. If R_L is strictly proper, as is quite often the case, then we have $\deg(N_L) < \deg(D_L)$. The design objective is to determine whether there is a constant $K_0 \geq 0$ so that the characteristic polynomial $D_L + K_0 N_L$ is Hurwitz. One may even want to choose K so that certain performance objectives are met. This is discussed in Section 11.3.

2. As we saw in Section 6.4.2, the closed-loop characteristic polynomial for static output feedback of a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ is given by

$$P_{\boldsymbol{A}}(s) + F\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$$

where F is the output feedback constant. Since $\boldsymbol{c}^t \mathrm{adj}(s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{b}$ is a polynomial of degree at most $n-1$, this will have the form of $P_1 + FP_2$ where $\deg(P_2) < \deg(P_2)$.

3. It is possible that the variable constant K may not appear in as simple a manner as indicated in Figure 11.1. However, in some such cases, one can still reduce the characteristic polynomial to one of the desired form. Suppose that we have a plant with $R_P(s) = \frac{1}{ms^2}$ and we wish to stabilise it in a unity gain feedback loop with a PID controller $R_C(s) = K(1 + T_D s + \frac{1}{T_I}s)$. Let us suppose that, for whatever reason, we are interested only in changing the reset time T_I, and not the gain K. Furthermore, we restrict our interest to $T_I > 0$. Thus we are not immediately in the situation illustrated in Figure 11.1. Nevertheless, we proceed. The closed-loop characteristic polynomial is

$$s^3 + KT_D s^2 + Ks + \frac{K}{T_I}.$$

Now note that we may write this closed-loop characteristic polynomial as $P_1 + \alpha P_2$ if we take

$$P_1 = s^3 + KT_D s^2 + Ks, \quad P_2 = K, \quad \alpha = \frac{1}{T_I}.$$

Now, as T_I runs from 0 to ∞, α runs from ∞ to 0. Thus, even though we are not immediately in the form of Figure 11.1, we can write the characteristic polynomial is the form of a sum of two polynomials, one with a positive coefficient. Note that this may not always be possible, but sometimes it is. •

In each of the preceding three examples, we arrive at a characteristic polynomial that is of the form $P_1 + KP_2$ where $\deg(P_2) < \deg(P_1)$. It is such polynomials that are of interest to us in this chapter.

11.1.2 Definitions and general properties

Based on the discussion of the preceding section, we make the following definition.

11.2 Definition Let $N, D \in \mathbb{R}[s]$ have the following properties:

(i) D is monic;

(ii) either N or $-N$ is monic;

(iii) $\deg(N) < \deg(D)$.

A ***(N, D)-polynomial family*** is a family of polynomials of the form $\mathscr{P}(N, D) = \{D + KN \mid K \geq 0\}$. For a fixed $K \geq 0$ we shall denote $P_K = D + KN \in \mathbb{R}[s]$. The ***root-locus*** of an (N, D)-polynomial family is the subset

$$\mathrm{RL}(\mathscr{P}(N, D)) = \overline{\{z \in \mathbb{C} \mid P_K(z) = 0 \text{ for some } K \geq 0\}}$$

of the complex plane.[1] A (N, D)-polynomial family is ***Hurwitz*** if there exists $K \geq 0$ so that P_K is Hurwitz. •

11.3 Remark Note that we do ask for D to be monic and either N or $-N$ to be monic when defining a (N, D)-polynomial family. That D should be taken as monic seems natural. As for N, clearly one can make either N or $-N$ monic by simply redefining K if necessary. Thus we loose no generality with these assumptions. •

Note that the problem of determining whether a (N, D)-polynomial family is Hurwitz is exactly equivalent to the problem of determining whether a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ admits a stabilising static output feedback controller (see Exercise E11.2). Therefore, to determine whether a (N, D)-polynomial family is Hurwitz is as difficult as proving the existence of stabilising static output feedback. This problem, as was noted at the beginning of Section 10.2.2 is NP hard. Thus we cannot expect to solve this problem is an easily computable manner. However, we shall provide a rough description of the root-locus for a (N, D)-polynomial family which suggests that you might be able to numerically ascertain whether such a family is Hurwitz.

Note that the root-locus consists of collections of roots of polynomials that depend continuously, in this case linearly, on a parameter K. Let us record some properties of the roots of such parameterised polynomials.

11.4 Lemma *Let $P(s, K) = s^n + p_{n-1}(K)s^{n-1} + \cdots + p_1(K)s + p_0(K)$ be a polynomial with coefficients differentiable functions of the parameter $K \in \mathbb{R}$. For $K \in \mathbb{R}$, denote by $\{z_1(K), \ldots, z_n(K)\}$ the roots of $P(s, K)$.*

(i) The functions $\mathbb{R} \ni K \mapsto z_i(K) \in \mathbb{C}$, $i = 1, \ldots, n$, may be chosen to be continuous.

(ii) If the roots $z_i(K)$, $i = 1, \ldots, n$, are chosen to be continuous functions, then, if $z_i(K_0)$ is a root of multiplicity one for $P(s, K_0)$ then there exists $\epsilon > 0$ so that $z_i|[K_0-\epsilon, K_0+\epsilon]$ so that

(a) $f_i(K_0) = z_i(K_0)$ and

[1] Note that $\overline{S}$ denotes the closure of the set S, meaning the set and all of its limit points.

(b) $P(z_i(K), K) = 0$.

More succinctly, near nondegenerate roots of $P(s, K_0)$*, the roots are differentiable functions of the parameter* K*.*

11.5 Remark With the notation of the lemma, we may write the roots of $P(s, K)$ as $\{z_1(K), \dots, z_n(K)\}$. We shall do this frequently, and we always make the assumption that this is done in such a way that the functions $K \mapsto z_i(K)$, $i = 1, \dots, n$, are continuous. •

We do not prove this lemma, although part (ii) is easily proved using the implicit function theorem (see Exercise E11.3). Let us illustrate the lemma with a simple example.

11.6 Example Let us take $P(s, K) = s^2 + K$. For $K \leq 0$ the roots are $\{z_1(K) = \sqrt{-K}, z_2(K) = -\sqrt{-K}\}$ and for $K > 0$ the roots are $\{z_1(K) = i\sqrt{K}, z_2(K) = -i\sqrt{K}\}$. In Figure 11.2 we

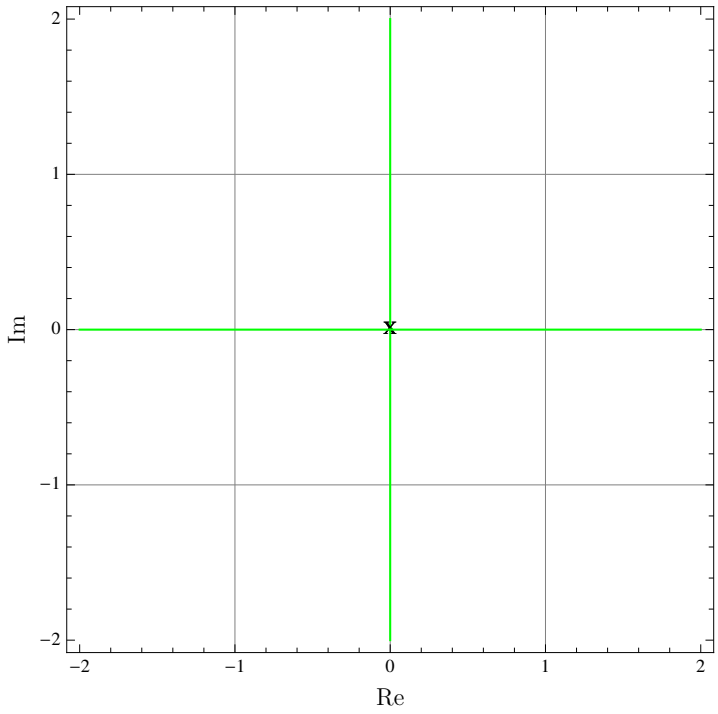


Figure 11.2 Locus of roots for $P(s, K) = s^2 + K$

plot the locus of roots for this polynomial as K runs from -4 to 4. Note that if K varies slightly, the location of the roots also change only slightly. What's more, as long as $K \neq 0$ the roots are distinct, and the locus of roots near such values of K are smooth curves in the complex plane (lines in this case). However, for the repeated root when $K = 0$, the character of the locus of roots is not smooth near this repeated root. Indeed, at the origin where this repeated root lies, the locus of roots has an intersection. This will typically be the case, and constitutes one of the more challenging aspects of making root-locus plots. •

11.2 Properties of the root-locus

In this section we do two things. First we provide a list of provable properties of the root-locus of a (N, D)-polynomial family. These are presented with proofs as these do not

seem to be part of the standard discussion of the root-locus method. The presentation in this first part of the section produces for us a good understanding of what is known and unknown about the nature of the root-locus. The next thing we do is put this understanding together to develop a methodology for graphically producing the root-locus for a (N, D)-polynomial family. It is this graphical method which forms the bulk of the presentation on the root-locus method in most standard texts.

11.2.1 A rigorous discussion of the root-locus

We follow here the paper of Krall [1961], starting with a (N, D)-polynomial family $\mathscr{P}(N, D)$. The essential idea is that when $K = 0$ the roots of P_K are obviously those of D. Suppose that $\deg(D) = n$ so that P_0 has n roots, if one counts multiplicities. Now, as K increases, Lemma 11.4 suggests that these n roots should move about continuously in the complex plane. Now suppose that $m = \deg(N)$. What will turn out to happen is that m of the n roots of P_K will start at the roots for $D = P_0$ and end up at the roots of N as $K \to \infty$. One must then account for the remaining $n - m$ roots, which, it turns out, shoot off to infinity in predictable ways.

Let's get down to it. For concreteness let us write

$$\begin{aligned} D(s) &= s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \\ N(s) &= \pm s^m + q_{m-1}s^{m-1} + \cdots + q_1 s + q_0. \end{aligned}$$

We define the ***centre of gravity*** of $\mathscr{P}(N, D)$ to be

$$\mathrm{CG}(\mathscr{P}(N, D)) = \begin{cases} -\frac{p_{n-1}-q_{m-1}}{n-m}, & N \text{ monic} \\ -\frac{p_{n-1}+q_{m-1}}{n-m}, & -N \text{ monic.} \end{cases} \tag{11.1}$$

The following result gives a simple characterisation of the centre of gravity in terms of the roots of N and D.

11.7 Lemma $\mathrm{CG}(\mathscr{P}(N, D)) = \dfrac{1}{n-m}\Big(\sum_{j=1}^{n} z_j - \sum_{k=1}^{m} \zeta_k\Big)$, *where* $z_1, \ldots, z_n$ *are the roots of* D *and* $\zeta_1, \ldots, \zeta_m$ *are the roots for* N.

Proof For a monic polynomial

$$P(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0,$$

one may easily show that the sum of the roots of P is equal to $-a_{n-1}$ (see Exercise EC.3). Thus the sum of the roots for D is p_{n-1}, and the sum of the roots for N is q_{m-1} is N is monic and $-q_{n-1}$ is $-N$ is monic. ■

Now, through the centre of gravity we construct $n - m$ rays in the complex plane. We denote these rays by $\alpha_1, \ldots, \alpha_{n-m}$ and define them by

$$\alpha_j = \begin{cases} \{\mathrm{CG}(\mathscr{P}(N, D)) + re^{(2j-1)\pi i/(n-m)} \mid r \geq 0\}, & N \text{ monic} \\ \{\mathrm{CG}(\mathscr{P}(N, D)) + re^{2j\pi i/(n-m)} \mid r \geq 0\}, & -N \text{ monic.} \end{cases} \tag{11.2}$$

We call these rays the ***asymptotes*** for $\mathscr{P}(N, D)$. In Figure 11.3 we show the asymptotes when the centre of gravity is at $1 + i0$ and when $n - m = 3$. Note that generally the asymptotes will depend on whether N or $-N$ is monic.

With the notion of asymptotes in place, we may state the following result which indicates in what manner some of the roots of P_K go to infinity.

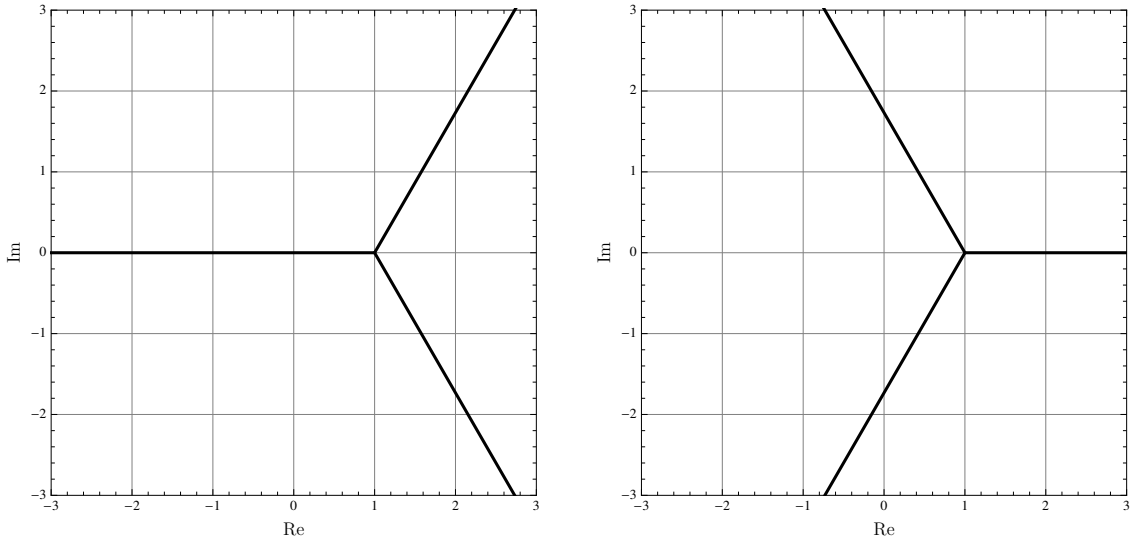


Figure 11.3 Asymptotes if $\mathrm{CG}(\mathscr{P}(N,D)) = 1 + i0$ and $n - m = 3$, and when N is monic (left) and $-N$ is monic (right)

11.8 Proposition *Let $\mathscr{P}(N,D)$ be a (N,D)-polynomial family and denote the roots of $P_K = D + KN$ by $\{z_1(K), \dots, z_n(K)\}$. Then there exists distinct $j_1, \dots, j_{n-m} \in \{1, \dots, n\}$ so that*

(i) if N is monic, $\lim_{K\to\infty} \left| z_{j_k}(K) - \mathrm{CG}(\mathscr{P}(N,D)) - K^{1/(n-m)} e^{(2k-1)\pi i/(n-m)} \right| = 0$, and

(ii) if $-N$ is monic, $\lim_{K\to\infty} \left| z_{j_k}(K) - \mathrm{CG}(\mathscr{P}(N,D)) - K^{1/(n-m)} e^{2k\pi i/(n-m)} \right| = 0$,

for each $k = 1, \dots, n - m$.

Proof The bulk of the proof is contained in the following result.

1 Lemma *The proposition holds when $\mathrm{CG}(\mathscr{P}(N,D)) = 0 + i0$.*

Proof In this case we can write

$$P_K(s) = \left(s^n + as^{n-1} + p_{n-2}s^{n-2} + \cdots + p_1 s + p_0\right) - Ke^{i\theta}\left(s^m + as^{m-1} + q_{m-2}s^{m-2} + \cdots + q_1 s + q_0\right),$$

where $K \geq 0$ and $\theta \in \{0, -\pi\}$. The main point is that since $\mathrm{CG}(\mathscr{P}(N,D)) = 0 + i0$, the next to highest coefficients of D and N must be equal if N is monic. Now fix $k \in \{1, \dots, n-m\}$ and make the substitution

$$s = we^{i\theta_k}, \quad \theta_k = \frac{2k\pi + \theta}{n-m}, w \in \mathbb{C},$$

so that

$$e^{-i\theta_k} P_K(we^{i\theta_k}) = \left(w^n + aw^{n-1}e^{-i\theta_k} + \sum_{j=0}^{n-2} a_j w^j\right) - K\left(w^m + aw^{m-1}e^{-i\theta_k} + \sum_{j=0}^{m-2} b_j w^j\right),$$

for some coefficients $a_0, \dots, a_{n-2}, b_0, \dots, b_{m-2}$ whose exact character are not of much interest to us. Now define

$$M = \sum_{j=0}^{n-2} |a_j| (3|a| + 1)^{n-2-j} \left(\frac{4}{3}\right)^j + \sum_{j=0}^{m-2} |b_j| (3|a| + 1)^{m-2-j} \left(\frac{4}{3}\right)^j$$

and

$$\rho = K^{1/(n-m)}, \quad \epsilon = \frac{3^m M}{\rho^2} = \frac{3^m M}{K^{2/(n-m)}}.$$

The key observation to make is that M is independent of K. Let

$$K_0 = \max\Big\{(3|a|+1)^{n-m}, (3^{m+1}M)^{(n-m)/2}, \Big(\frac{3^m M}{\frac{1}{2}|1-e^{\pm\pi i/(2(n-m))}|}\Big)^{(n-m)/2}\Big\}.$$

If $K > K_0$ then we have

$$\rho = K^{1/(n-m)} > \big((3|a|+1)^{n-m}\big)^{1/(n-m)} > 3|a|+1,$$

and

$$\epsilon = \frac{3^m M}{K^{2/(n-m)}} < \frac{3^m M}{\big((3^{m+1}M)^{(n-m)/2}\big)^{2/(n-m)}} = \frac{1}{3},$$

and

$$\epsilon = \frac{3^m M}{K^{2/(n-m)}} < \frac{3^m M}{\Big(\big(\frac{3^m M}{\frac{1}{2}|1-e^{\pm\pi i/(2(n-m))}|}\big)^{(n-m)/2}\Big)^{2/(n-m)}} = \frac{1}{2}|1-e^{\pm\pi i/(2(n-m))}|. \tag{11.3}$$

Now make the substitution $\xi = \frac{w}{\rho} \in \mathbb{C}$ so that, with our previous substitution for s, we have

$$\begin{aligned}\rho^{-n}e^{-i\theta_k}P_K(\rho\xi e^{i\theta_k}) &= \Big(\xi^n + \frac{a}{\rho}e^{-i\theta_k}\xi^{n-1} + \sum_{j=0}^{n-2} a_j\rho^{j-n}\xi^j\Big) - \Big(\xi^m + \frac{a}{\rho}e^{-i\theta_k}\xi^{m-1} + \sum_{j=0}^{m-2} b_j\rho^{j-m}\xi^j\Big)\\ &= \xi^{m-1}\Big(\xi\tfrac{a}{\rho}e^{-i\theta_k}\Big)(\xi^{n-m}-1) + \frac{1}{\rho^2}\Big(\sum_{j=0}^{n-2} a_j\rho^{j-n+2}\xi^j - \sum_{j=0}^{m-2} b_j\rho^{j-m+2}\xi^j\Big).\end{aligned}$$

Now define

$$\begin{aligned}f(\xi) &= \xi^{m-1}\Big(\xi + \frac{a}{\rho}e^{-i\theta_k}\Big)(\xi^{n-m}-1)\\ g(\xi) &= \frac{1}{\rho^2}\Big(\sum_{j=0}^{n-2} a_j\rho^{j-n+2}\xi^j - \sum_{j=0}^{m-2} b_j\rho^{j-m+2}\xi^j\Big).\end{aligned}$$

Now let Γ_ϵ be the circle of radius ϵ centred at $1+i0$, with ϵ as previously defined. Since $\epsilon < \frac{1}{3}$, the real part of $\xi^k \in \mathbb{C}$ is positive on Γ_ϵ for $k = 1, \dots, n-m$. Note that $\xi^{n-m}-1$ vanishes when $\xi \in \{1, e^{2\pi i/(n-m)}, e^{-2\pi i/(n-m)}, \dots\}$. Therefore, by the inequality (11.3), $\xi^{n-m}-1$ is greater than ϵ when $\xi \in \Gamma_\epsilon$. By the triangle inequality we also have check

$$\begin{aligned}\Big|\xi + \frac{a}{\rho}e^{-i\theta_k}\Big| &\le |\xi| |\frac{a}{\rho}e^{-i\theta_k}|\\ &\le\end{aligned}$$

Therefore, for $\xi \in \Gamma_\epsilon$ we have

$$|f(\xi)| > (1-\epsilon)^{m-1}\cdot\tfrac{1}{3}\cdot 1\cdot\epsilon > \big(\tfrac{1}{3}\big)^m\epsilon.$$

Thus there exists K_0 (redefine this if necessary) so that for all $K > K_0$, f has only the zero at $1+i0$ within Γ_ϵ. By definition of M, for $\xi \in \Gamma_\epsilon$ we have

$$|g(\xi)| \le \frac{M}{\rho^2} = \big(\tfrac{1}{3}\big)^m\epsilon.$$

Therefore, for $\xi \in \Gamma_\epsilon$ we have $|f(\xi)| > |g(\xi)|$. From Rouchés theorem, Theorem D.6, we may conclude that the number of zeros for $f+g$ within Γ_ϵ is the same as the number of zeros f, namely 1, provided that $K > K_0$. This shows that there is a zero ξ_0 inside Γ_ϵ so that

$$\rho^{-n} e^{-i\theta_k} P_K(\rho \xi_0 e^{i\theta_k}) = 0 \quad \Longrightarrow \quad P_K(z_0) = 0,$$

where

$$z_0 = \rho \xi_0 e^{i\theta_k} = K^{1/(n-m)} \xi_0 e^{i\theta_k}.$$

The lemma now follows since $\lim_{K\to\infty} \xi_0 = 1$, since $\lim_{K\to\infty} \epsilon = 0$. ▼

With the lemma in hand, it is now comparatively straightforward to prove the proposition. Let us proceed in the case when N is monic so that

$$\mathrm{CG}(\mathscr{P}(N,D)) = -\frac{p_{n-q} - q_{n-1}}{n-m}.$$

Making the substitution $s = w + \mathrm{CG}(\mathscr{P}(N,D))$ gives

$$P_K(w + \mathrm{CG}(\mathscr{P}(N,D))) = w^n + \frac{nq_{n-1} - mp_{n-1}}{n-m} w^{n-1} + \sum_{j=0}^{n-2} a_j w^j + K\Big(w^m + \frac{nq_{n-1} - mp_{n-1}}{n-m} w^{m-1} + \sum_{j=0}^{m-2} b_j w^j\Big),$$

for some coefficients $a_0, a_{n-2}, b_0, \ldots, b_{m-2}$ whose exact form is not of particular interest to us. Now, by Lemma 1, there is a collection of roots $\{\omega_{j_1}(K), \ldots, \omega_{j_{n-m}}(K)\}$ to this polynomial that satisfy

$$\lim_{K\to\infty} \left|\omega_{j_k}(K) - K^{1/(n-m)} e^{2k\pi i/(n-m)}\right| = 0.$$

These clearly give rise to the roots $\{z_{j_1}(K), \ldots, z_{j_{n-m}}(K)\}$ as given in the statement of the proposition. ■

Let us see how this works in a simple example.

11.9 Example Let us take $(N(s), D(s)) = (\pm 1, s^2)$. Note that $\mathrm{CG}(\mathscr{P}(N,D)) = 0 + i0$.

First let us consider the case where $N = 1$ so that $P_K(s) = s^2 + K$. Since $n - m = 2$ we expect there to be 2 roots that shoot off to infinity. According to Proposition 11.8(i), in the limit as $K \to \infty$ these roots should behave like $\sqrt{K}e^{\pi i/2} = i\sqrt{K}$ and $\sqrt{K}e^{3\pi i/2} = -i\sqrt{K}$ as K gets large. Not only do the roots behave like this as K gets large, they are exactly given by these expressions for *all* K!

When $N = -1$ we have $P_K(s) = s^2 - K$. In this case, Proposition 11.8(ii) predicts that as $K \to \infty$ the two roots of P_K behave like $\sqrt{K}e^{\pi i} = -\sqrt{K}$ and $\sqrt{K}e^{2\pi i} = +\sqrt{K}$. Again, these happen to be the exact values of the roots. Clearly we do not expect this to generally be the case. •

Now let us consider what happens to those remaining m roots of P_K as K varies.

11.10 Proposition *Let $\mathscr{P}(N,D)$ be a (N,D) polynomial family with $\{z_1(K),\dots,z_n(K)\}$ the roots of P_K for $K \geq 0$. Also denote $\{\zeta_1,\dots,\zeta_m\}$ as the roots for N. There exists distinct $j_1,\dots,j_m \in \{1,\dots,n\}$ so that for each $k=1,\dots,m$, the curve $K \mapsto z_{j_k}(K)$ is a continuous curve starting from $z_{j_k}(0)$ and satisfying $\lim_{K\to\infty} z_{j_k}(K) = \zeta_k$.*

Proof By part (i) of Lemma 11.4 we know that the curve $K \mapsto z_j(K)$ is continuous for every $j=1,\dots,n$. Now fix $k \in \{1,\dots,m\}$ and suppose that the multiplicity of the root ζ_k is ℓ. Choose $\epsilon > 0$ and let Γ_ϵ be the circle of radius ϵ centred at ζ_k. Take ϵ sufficiently small that there are no roots of N within Γ_ϵ but z_k. The number of zeros of P_K within Γ_ϵ is given by the Principle of the Argument to be

$$\frac{1}{2\pi i}\int_{\Gamma_\epsilon} \frac{P_K'(s)}{P_K(s)}\,\mathrm{d}s.$$

Using $P_K(s) = D(s) + KN(s)$ we compute

$$\frac{1}{2\pi i}\int_{\Gamma_\epsilon} \frac{P_K'(s)}{P_K(s)}\,\mathrm{d}s = \frac{1}{2\pi i}\int_{\Gamma_\epsilon} \frac{N'(s)}{N(s)}\,\mathrm{d}s + \frac{1}{2\pi i}\int_{\Gamma_\epsilon} \frac{1}{K}\frac{D'(s)N(s) - D(s)N'(s)}{(\frac{1}{K}D(s)+N(s))N(s)}\,\mathrm{d}s.$$

Since Γ_ϵ is bounded, both polynomials N and D are bounded on Γ_ϵ. Therefore, we may choose K sufficiently large that $\frac{1}{K}|D(s)| < |N(s)|$ for $s \in \Gamma_\epsilon$. Therefore we see that the number of zeros of P_K within Γ_ϵ is given by $\ell + \delta$ where δ may be made as small as we please by increasing K. Thus we conclude that for K sufficiently large, P_K has ℓ zeros in Γ_ϵ. As this holds for all $\epsilon > 0$, we conclude that ℓ zeros of P_K limit to ζ_k as $K \to \infty$. As this is true for all roots ζ_k for N, this proves the proposition. ■

Let us illustrate this proposition in another simple example.

11.11 Example We take $(N(s),D(s)) = (s+1, s^2)$ so that $P_K(s) = s^2 + Ks + K$. Note that $\mathrm{CG}(\mathscr{P}(N,D)) = 1 + i0$. Since $n - m = 1$ and N is monic, by part (i) of Proposition 11.8 we expect to see one of the roots approaching $1 + Ke^{\pi i} = 1 - K$ as $K \to \infty$. The other root should start at one of the roots for D and end up in the root $\zeta = -1$ for N. We compute the roots to be

$$-\tfrac{K}{2} \pm \sqrt{(\tfrac{K}{2})^2 - K},$$

and the root-locus is shown in Figure 11.4. Note that one of the roots does indeed start out at $0 + i0$, a root for D, and moves continuously toward $-1 + i0$, a root for N. In the root-locus of Figure 11.4, the path taken by this root is not unique. It can start at $0 + i0$ and go down along the semi-circle, then turn right toward $-1 + i0$, or it can go up along the other semi-circle, and then again turn right toward $-1 + i0$. Note that Proposition 11.10 does not preclude this ambiguity. Also, we see in Figure 11.4 that one of the roots is going off to $-\infty$ as predicted.

Let us see if we can verify the limiting behaviour analytically. For K large, both roots of P_K are real. Let us write

$$z_1(K) = -\tfrac{K}{2} - \sqrt{(\tfrac{K}{2})^2 - K}, \quad z_2(K) = -\tfrac{K}{2} + \sqrt{(\tfrac{K}{2})^2 - K}$$

Note that

$$\begin{aligned} z_1(K) - (1-K) &= -1 + \tfrac{K}{2} - \sqrt{(\tfrac{K}{2})^2 - K} \\ &= -1 + \tfrac{K}{2} - (\tfrac{K}{2})\sqrt{1 - \tfrac{4}{K}} \\ &= -1 + \tfrac{K}{2} - (\tfrac{K}{2})\big(1 - \tfrac{1}{2}\tfrac{4}{K} - \tfrac{1}{8}(\tfrac{4}{K})^2 + \cdots\big) \\ &= \tfrac{1}{K} + \cdots, \end{aligned}$$

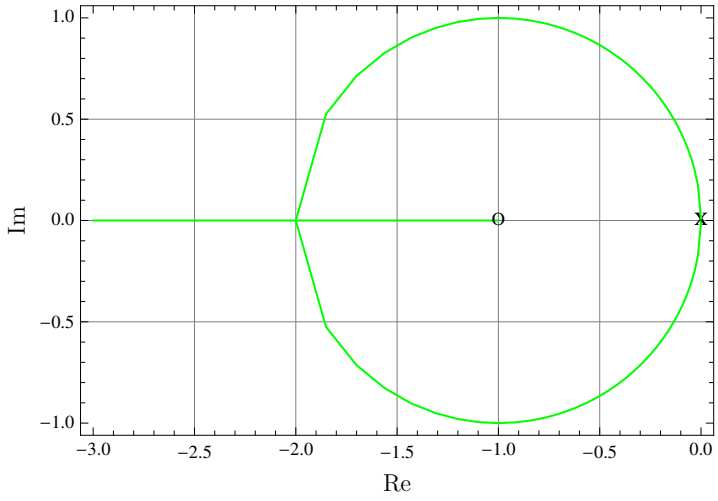


Figure 11.4 Root locus for $(N(s), D(s)) = (s+1, s^2)$

so that $\lim_{K\to\infty}(z_1(K) - (1-K)) = 0$ as desired. Similarly we have

$$\begin{aligned} z_2(K) - (-1) &= 1 - \tfrac{K}{2} + \sqrt{(\tfrac{K}{2})^2 - K} \\ &= 1 - \tfrac{K}{2} + (\tfrac{K}{2})\sqrt{1 - \tfrac{4}{K}} \\ &= 1 - \tfrac{K}{2} + (\tfrac{K}{2})\big(1 - \tfrac{1}{2}\tfrac{4}{K} - \tfrac{1}{8}(\tfrac{4}{K})^2 + \cdots\big) \\ &= -\tfrac{1}{K} + \cdots, \end{aligned}$$

so that $\lim_{K\to\infty}(z_2(K) - (-1)) = 0$, as predicted. In both of these computations, we have used the Taylor series for $\sqrt{1-x}$ about $x = 0$. •

Now we wish to examine the way that the root-locus passes through certain points on the root-locus. Let $z_0 \in \mathrm{RL}(\mathscr{P}(N, D))$ so that z_0 is a root of P_{K_0} for some $K_0 \geq 0$. If the multiplicity of z_0 is one, then the root-locus passes differentiably through z_0. Denote by $z_i(K)$ that root for P_K for which $z_i(K_0) = z_0$. We define the ***arrival angle*** of $\mathrm{RL}(\mathscr{P}(N, D))$ at z_0 to be the angle $\theta_a(z_0) \in (-\pi, \pi]$ for which

$$\lim_{K\uparrow K_0} \frac{z_i(K) - z_0}{K - K_0} = \alpha_a e^{i\theta_a(z_0)}$$

for some suitable $\alpha_a > 0$. This make precise the intuitive notion of arrival angle. Similarly, the ***departure angle*** of $\mathrm{RL}(\mathscr{P}(N, D))$ at z_0 is the angle $\theta_d(z_0) \in (-\pi, \pi]$ for which

$$\lim_{K\downarrow K_0} \frac{z_i(K) - z_0}{K - K_0} = \alpha_d e^{i\theta_d(z_0)}$$

for some suitable $\alpha_d > 0$. Since the root-locus is differentiable at z_0, one can easily see that we must have $\theta_d(z_0) \in \{\theta_a(z_0) + \pi, \theta_a(z_0) - \pi\}$. When z_0 has multiplicity $\ell > 1$, things are not so transparent since there will be more than one arrival and departure angle. In this case we have ℓ roots $z_{j_1}(K), \ldots, z_{j_\ell}(K)$ which can pass through z_0 when K_0. We may define ℓ ***arrival angles***, $\theta_{a,1}(z_0), \ldots, \theta_{a,\ell}(z_0)$, by

$$\lim_{K\uparrow K_0} \frac{z_i(K) - z_0}{K - K_0} = \alpha_{a,k} e^{i\theta_{a,k}(z_0)}, \quad k = 1, \ldots, \ell,$$

for some suitable $\alpha_{a,1},\dots,\alpha_{a,\ell} > 0$. Similarly, we may define ℓ ***departure angles***, $\theta_{d,1}(z_0),\dots,\theta_{d,\ell}(z_0)$, by

$$\lim_{K\uparrow K_0}\frac{z_i(K)-z_0}{K-K_0} = \alpha_{d,k}e^{i\theta_{d,k}(z_0)},\quad k=1,\dots,\ell,$$

for some suitable $\alpha_{d,1},\dots,\alpha_{d,\ell} > 0$. For any multiplicity of the root z_0, let us denote by $\Theta_a(z_0)$ the collection of all arrival angles, and by $\Theta_d(z_0)$ the collection of all departure angles.

We are now ready to state the character of some of the arrival and departure angles, namely the departure angles for the roots of D and the arrival angles for the roots of N. In practice, these are the most helpful to know in terms of being able to produce the root-locus for an (N,D)-polynomial family.

11.12 Proposition *Let $\mathscr{P}(N,D)$ be an (N,D)-polynomial family and let $\{z_1,\dots,z_n\}$ be the roots of D and $\{\zeta_1,\dots,\zeta_m\}$ be the roots of N.*

(i) If z_j has multiplicity ℓ, then $\Theta_d(z_j)$ is the following collection of ℓ angles:

(a) if N is monic, take

$$\theta_{d,k}(z_j) = \frac{1}{\ell}\Big(\sum_{\alpha=1}^{m}\measuredangle(z_j-\zeta_\alpha) - \sum_{\substack{\alpha=1\\ \text{excluding } z_j}}^{n}\measuredangle(z_j-z_\alpha) - (2k-1)\pi\Big),\quad k=1,\dots,\ell;$$

(b) if $-N$ is monic, take

$$\theta_{d,k}(z_j) = \frac{1}{\ell}\Big(\sum_{\alpha=1}^{m}\measuredangle(z_j-\zeta_\alpha) - \sum_{\substack{\alpha=1\\ \text{excluding } z_j}}^{n}\measuredangle(z_j-z_\alpha) - 2k\pi\Big),\quad k=1,\dots,\ell.$$

(ii) If ζ_j has multiplicity ℓ, then $\Theta_a(\zeta_j)$ is the following collection of ℓ angles:

(a) if N is monic, take

$$\theta_{a,k}(\zeta_j) = \frac{1}{\ell}\Big(\sum_{\alpha=1}^{n}\measuredangle(\zeta_j-z_\alpha) - \sum_{\substack{\alpha=1\\ \text{excluding } z_j}}^{m}\measuredangle(\zeta_j-\zeta_\alpha) + (2k-1)\pi\Big),\quad k=1,\dots,\ell;$$

(b) if $-N$ is monic, take

$$\theta_{d,k}(\zeta_j) = \frac{1}{\ell}\Big(\sum_{\alpha=1}^{n}\measuredangle(\zeta_j-z_\alpha) - \sum_{\substack{\alpha=1\\ \text{excluding } \zeta_j}}^{m}\measuredangle(\zeta_j-\zeta_\alpha) + 2k\pi\Big),\quad k=1,\dots,\ell.$$

Proof Let us for convenience write $P_K = D - Ke^{i\theta}N$ and assume that N is monic. Thus, we recover the monic cases for N and $-N$ by taking $\theta=\pi$ and $\theta=0$, respectively. We then write

$$D(s) = \prod_{j=1}^{n}(s-z_j),\quad N(s) = \prod_{j=1}^{m}(s-\zeta_j). \tag{11.4}$$

(i) For $j\in\{1,\dots,n\}$, let ℓ be the multiplicity of z_j. For $k\in\{1,\dots,\ell\}$ let $z_{j_k}(K)$ be a root of P_K which approaches z_j as $K\to 0$. Note that

$$D(z_{j_k}(K)) - Ke^{i\theta}N(z_{j_k}(K)) = 0 \quad\Longrightarrow\quad \frac{Ke^{i\theta}N(z_{j_k}(K))}{D(z_{j_k}(K))} = 1.$$

Taking complex logarithms of both sides, using (11.4), we obtain

$$\theta + \sum_{j=1}^{m} \measuredangle(s - \zeta_j) - \sum_{j=1}^{n} \measuredangle(s - z_j) = 2k\pi$$

where $k \in \mathbb{Z}$. Therefore,

$$\ell\measuredangle(s - z_j) = \sum_{\alpha=1}^{m} \measuredangle(s - \zeta_\alpha) - \sum_{\substack{\alpha=1 \\ \text{excluding } z_j}}^{n} \measuredangle(s - z_\alpha) + \theta - 2k\pi, \quad k \in \mathbb{Z}.$$

In the limit as $K \to 0$, the result follows, noting the relation between θ and N or $-N$ being monic.

(ii) The argument is exactly as in part (i), except we let $K \to \infty$. ■

Let us examine this again in an example.

11.13 Example **(Example 11.11 cont'd)** We again take $(N(s), D(s)) = (s + 1, s^2)$. In this case we have a root $z_1 = 0$ for D of multiplicity 2 and a root $\zeta_1 = -1$ for N of multiplicity 1. Since N is monic, Proposition 11.12 predicts that

$$\Theta_d(z_1) = \{-\tfrac{\pi}{2}, -\tfrac{3\pi}{2}\}, \quad \Theta_a(\zeta_1) = \{-\pi\}.$$

This is indeed consistent with Figure 11.4. •

The next result tells us that we should expect to see certain parts of the real axis within the root-locus.

11.14 Proposition *Let $\mathscr{P}(N, D)$ be an (N, D)-polynomial family, and denote the roots of D by $\{z_1, \ldots, z_n\}$ and the roots of N by $\{\zeta_1, \ldots, \zeta_m\}$.*

(i) Suppose that N is monic and let $x_0 \in \mathbb{R} \subset \mathbb{C}$. Then $x_0 \in \mathrm{RL}(\mathscr{P}(N, D))$ if and only if the set

$$\{j \mid \mathrm{Re}(z_j) > x_0\} \cup \{j \mid \mathrm{Re}(\zeta_j) > x_0\}$$

has odd cardinality.[2]

(ii) Suppose that $-N$ is monic and let $x_0 \in \mathbb{R} \subset \mathbb{C}$. Then $x_0 \in \mathrm{RL}(\mathscr{P}(N, D))$ if and only if the set

$$\{j \mid \mathrm{Re}(z_j) > x_0\} \cup \{j \mid \mathrm{Re}(\zeta_j) > x_0\}$$

has even cardinality.

Proof (i) First note that since the roots of P_K come in complex conjugate pairs, roots with nonzero imaginary part will always contribute an even number to the cardinality of the sets

$$\{j \mid \mathrm{Re}(z_j) > x_0\}, \quad \{j \mid \mathrm{Re}(\zeta_j) > x_0\}.$$

Thus it suffices to consider only the sets

$$\{j \mid z_j \text{ real and } z_j > x_0\}, \quad \{j \mid \zeta_j \text{ real and } \zeta_j > x_0\}$$

[2]The ***cardinality*** of a set S is simply the number of points in S.

for odd cardinality. Let $x_0 \in \mathbb{R} \cap \mathrm{RL}(\mathscr{P}(N,D))$. Then

$$D(x_0) + KN(x_0) = 1 \quad \Longrightarrow \quad \frac{KN(x_0)}{D(x_0)} = -1.$$

Taking complex logarithms, and using the fact that D and N can be written as in (11.4), gives

$$\sum_{j=1}^{m} \measuredangle(x_0 - \zeta_j) - \sum_{j=1}^{n} \measuredangle(x_0 - z_j) = (2k-1)\pi \tag{11.5}$$

for $k \in \mathbb{Z}$. Clearly, the sums can be taken as being over real roots since the terms corresponding to a complex root for N or D and its conjugate will cancel from (11.5). If the cardinality of the set

$$\{j \mid \mathrm{Re}(z_j) > x_0\} \cup \{j \mid \mathrm{Re}(\zeta_j) > x_0\}$$

is odd, then we will have

$$\sum_{j=1}^{m} \measuredangle(x_0 - \zeta_j) - \sum_{j=1}^{n} \measuredangle(x_0 - z_j) = (2r-1)\pi$$

for some $r \in \mathbb{Z}$. Thus this part of the proposition follows.

(ii) The proof here goes just as in part (i), except that we write $D(x_0) - KN(x_0) = 1$ and assume N monic. ■

As always, let us check the conclusions of the proposition on an example.

11.15 Example **(Example 11.11 cont'd)** We again take $(N(s), D(s)) = (s+1, s^2)$. The roots of D are $\{z_1 = 0, z_2 = 0\}$ and the roots of N are $\{\zeta_1 = -1\}$. Thus, if $x_0 \in \mathbb{R} \subset \mathbb{C}$, there are an odd number of zeros for both D and N to the right of x_0 is and only if $x_0 < -1$. Then Proposition 11.14 tells us that

$$\mathrm{RL}(\mathscr{P}(N,D)) \cap \mathbb{R} = \{x + i0 \mid x < -1\}.$$

This is indeed consistent with Figure 11.4. •

11.2.2 The graphical method of Evans

Now we may present the graphical technique typically presented in classical texts for producing the root-locus. This rather ingenious technique was that developed by Evans (1948, 1950). For us, this is simply a matter of applying the results of the preceding section, and we shall only enumerate the steps one typically takes in such a construction.

11.16 Steps for making a plot of the root-locus Given: an (N,D)-polynomial family $\mathscr{P}(N,D)$.

1. Compute the roots $\{z_1, \ldots, z_n\}$ for D and place an X at the location of each root in $\mathbb{C}$.
2. Compute the roots $\{\zeta_1, \ldots, \zeta_m\}$ for N and place an O at the location of each root in $\mathbb{C}$.
3. Compute the centre of gravity using (11.1) or Lemma 11.7.
4. Draw the $n - m$ asymptotes using (11.2).
5. Use Proposition 11.14 to determine $\mathrm{RL}(\mathscr{P}(N,D)) \cap \mathbb{R}$.
6. Use Proposition 11.12 to determine how the root-locus departs from the roots of D.
7. Use Proposition 11.12 to determine how the root-locus arrives at the roots of N.

8. If you are lucky, you can give a reasonable guess as to how the root locus behaves. For filling in the gaps in a root-locus plot, a useful property of the root-locus is that it is invariant under complex conjugation. •

The last step is in some sense the most crucial. It is possible that one can do the steps preceding it, and still get the root-locus wrong. Some experience is typically involved in knowing how a "typical" root-locus diagram looks, and then extrapolating this to a given example. Thankfully, computers typically do a good job with producing root-locus plots. These can run into problems when there are repeated roots of P_K, however. Thus one should check that a computer produced root-locus has the essential properties, mainly the correct number of branches.

Let us go through the steps outlined above in an example to see how well it works.

11.17 Example We take $(N(s), D(s)) = (s+1, s^4+6s^3+14s^2+16s+8)$.

1. The roots for D are $\{z_1 = -2, z_2 = -2, z_3 = -1-i, z_4 = -1+i\}$.
2. The roots for N are $\{\zeta_1 = -1\}$.
3. $\mathrm{CG}(\mathscr{P}(N, D)) = -\frac{5}{3}$.
4. The asymptotes are given by

$$\begin{aligned}\alpha_1 &= \left\{ -\tfrac{5}{3} + re^{\frac{\pi i}{3}} \;\middle|\; r \geq 0 \right\}\\ \alpha_2 &= \left\{ -\tfrac{5}{3} + re^{\pi i} \;\middle|\; r \geq 0 \right\}\\ \alpha_3 &= \left\{ -\tfrac{5}{3} + re^{\frac{5\pi i}{3}} \;\middle|\; r \geq 0 \right\}.\end{aligned}$$

We show these asymptotes in Figure 11.5.

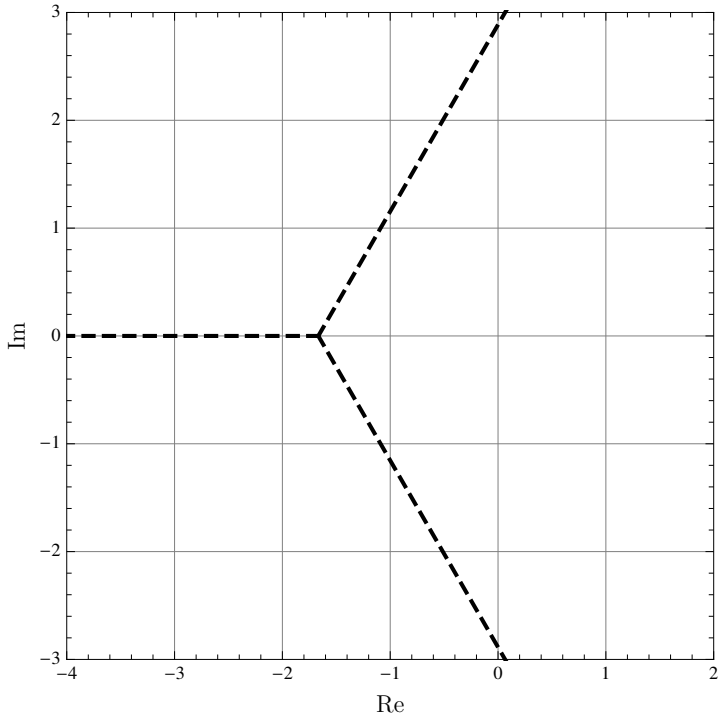


Figure 11.5 The asymptotes for $(N(s), D(s)) = (s+1, s^4+6s^3+14s^2+16s+8)$.

5. The point on the real-axis which lie on the root-locus are those points $x < -2$.
6. The departure angles from the roots of D are

$$\Theta_d(z_1) = \Theta_d(z_2) = \{0, -Pi\}, \quad \Theta_d(z_3) = \{-\tfrac{\pi}{2}\}, \quad \Theta_d(z_4) = \{-\tfrac{3\pi}{2}\}.$$

These departure angles are shown in Figure 11.6

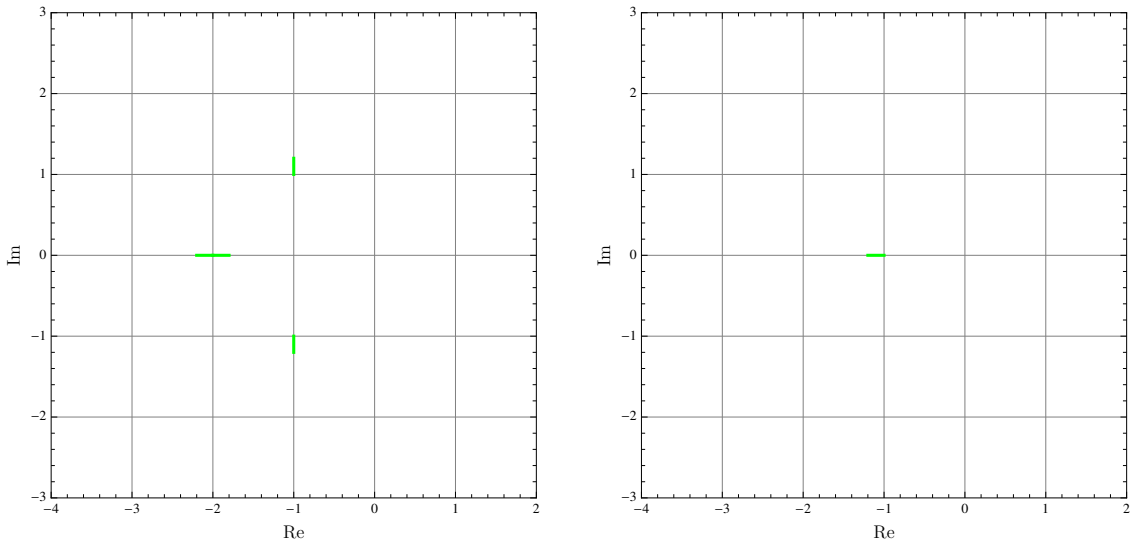


Figure 11.6 Departure angles from roots of $D(s) = s^4+6s^3+14s^2+16s+8$ (left) and arrival angle from the roots of $N(s) = s+1$ (right).

7. The arrival angles from the roots of N are

$$\Theta_a(\zeta_1) = \{\pi\}.$$

8. We play "connect the dots," hoping that we will arrive at a decent approximation of the actual root-locus. The actual root locus, along with the skeleton produced by our procedure, is shown in Figure 11.7. Note that without the knowledge of which roots of D go where as $K \to \infty$, it is in actuality difficult to know the details of the character of the root locus. Nevertheless, in simple examples one can often figure out the character of the root locus. •

11.3 Design based on the root-locus

The root-locus method can be thought of as a design technique. As such, it is most useful for plants that can be stabilised using simple controllers. For such plants, the root-locus method allows one to evaluate a one-parameter family of controllers by investigating the root-locus plot as the parameter varies. This assumes that one can put the system into a form where as the parameter varies we produce an (N, D)-polynomial family. We argued in Section 11.1.1 that this can sometimes be done, although it will by no means always be the case.

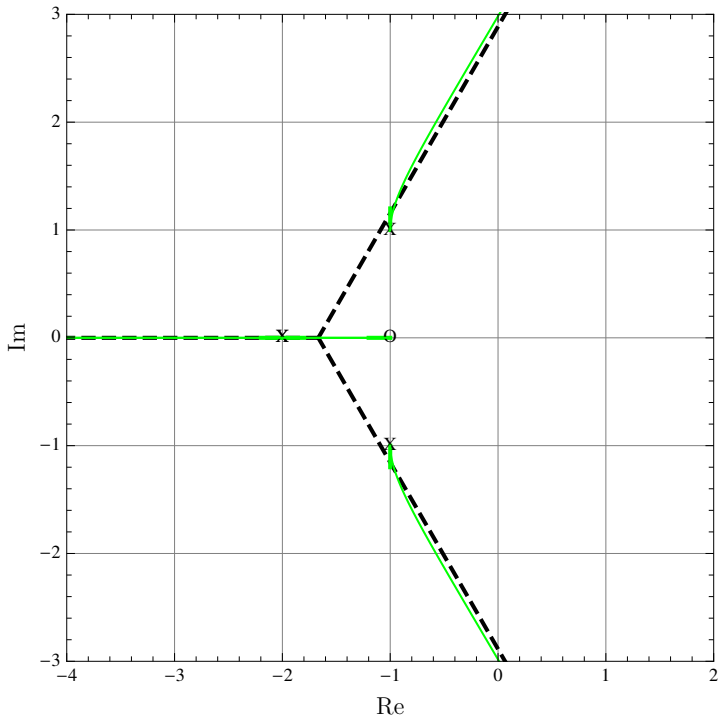


Figure 11.7 The skeleton and the actual root-locus for $(N(s), D(s) = (s+1, s^4+6s^3+14s^2+16s+8)$.

11.3.1 Location of closed-loop poles using root-locus

For systems that have simple enough behaviour, one can attempt to determine the quality of the performance of the system based on the location of the transfer function poles in the complex plane. For example, as we saw in Section 8.2, first and second-order transfer functions have performance attributes that are easily related to pole locations. Sometimes in practice one is able to design a controller that makes a system behave similarly to one of these two simple transfer functions, so one can use them as a basis for control design. With this as flimsy justification, we deal with interconnections like that depicted in Figure 11.8, and make the following assumption.

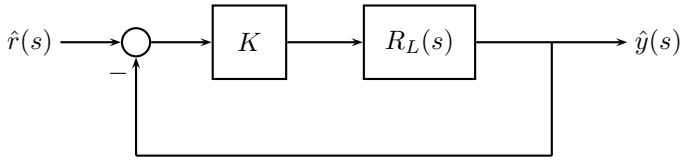


Figure 11.8 Interconnection for investigating closed-loop pole locations

In this section we will assume that the closed-loop interconnection of Figure 11.8 is designed so that for some values of K, the closed-loop transfer function is stable, and has a pair of complex conjugate poles whose real part is larger than that of all other poles. We call these complex conjugate poles the ***dominant poles****.*

The idea is that we pretend the dominant poles allow us to think of the closed-loop system as being second-order, and we base our choice of K on this consideration. In practice, one may wish to find a second-order transfer function that well approximates the system by, say, matching Bode plots as best one can.

Our approach is to carefully observe the relationship between system performance and pole location for second-order transfer functions. Thus we let

$$T_{\zeta,\omega_0}(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}.$$

For this transfer function, let us list some of the more important performance measures, some approximately, in terms of the parameters ζ and ω_0.

11.18 Performance measures for second-order transfer functions Consider the transfer function T_{ζ,ω_0}.

1.

11.3.2 Root sensitivity in root-locus

See Bishop and Dorf.

11.4 The relationship between the root-locus and the Nyquist contour

It turns out that there are some rather unobvious connections between the root-locus and the Nyquist contour. This is explained in a MIMO setting in [MacFarlane 1982], and cite info we only look at the SISO case here.

11.4.1 The symmetry between gain and frequency

We begin with a proper rational function R with c.f.r. (N, D) and canonical minimal realisation $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$. We next place R, represented by its canonical minimal realisation, into a *positive* feedback loop with feedback gain k^{-1} as shown in the upper block diagram in Figure 11.9. There are a few things to note here: (1) there is no negative sign where the feedback enters the summer with the reference r, (2) except for the sign, the block diagram is the same as that for static output feedback as studied in Section 6.4.2, and (3) the top block diagram in Figure 11.9 is equivalent to the bottom block diagram in the same figure. The second of these block diagrams has the advantage of making apparent a symmetry that exists between the parameter s and the parameter k. It is this symmetry which we study here.

Let us write

$$\begin{aligned}\boldsymbol{K}(s) &= \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} + \boldsymbol{D} \in \mathbb{R}(s)^{1\times 1}\\ \boldsymbol{S}(k) &= \boldsymbol{c}^t(k\boldsymbol{I}_1 - \boldsymbol{D})^{-1}\boldsymbol{b} + \boldsymbol{A} \in \mathbb{R}(k)^{n\times n}.\end{aligned}$$

The following lemma tells us how we should interpret these matrices of rational functions.

11.19 Lemma *The following statements hold:*

(i) $\boldsymbol{K}(s)$ *is the open-loop transfer function from* r *to* y *for the uppermost block diagram in Figure 11.9, i.e., the transfer function if the feedback loop is snipped;*

(ii) $\boldsymbol{S}(k)$ *is the "*$\boldsymbol{A}$*-matrix" for the closed-loop system in Figure 11.9, i.e., if the closed-loop transfer function has canonical minimal realisation* $\tilde{\Sigma} = (\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{c}}^t, \tilde{\boldsymbol{D}})$*, then* $\tilde{\boldsymbol{A}} = \boldsymbol{S}(k)$.

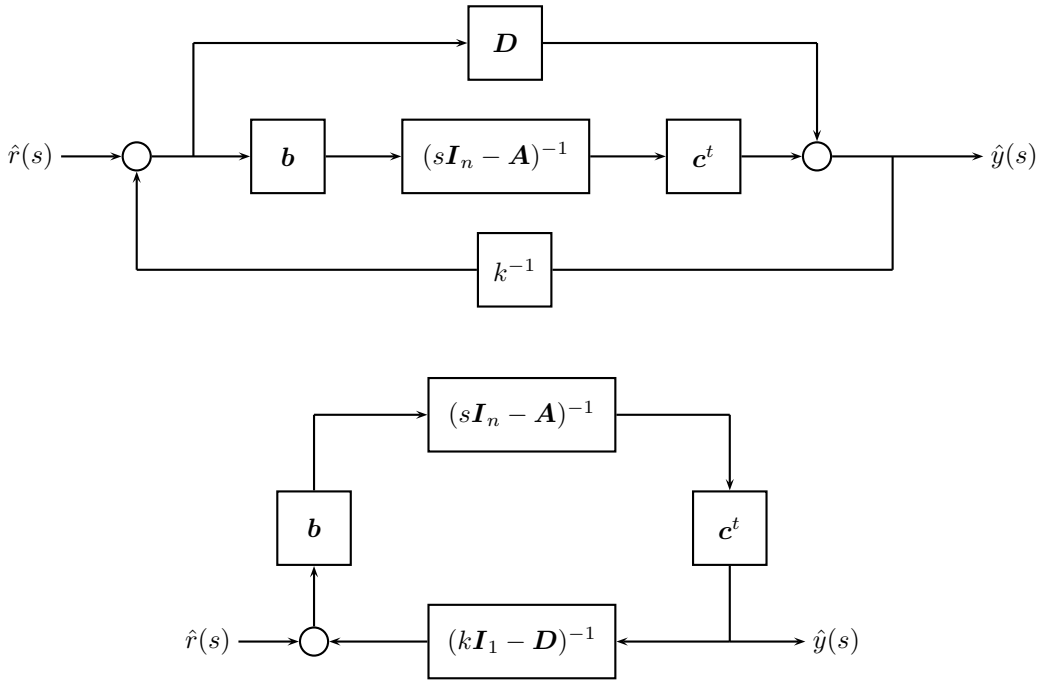


Figure 11.9 Equivalent block diagrams for studying the relationship between the root-locus and the Nyquist contour

Proof The first assertion is clear. The second follows from the same computations that gave the closed-loop system under static output feedback in Section 6.4.2. ■

Note that the closed-loop system is internally stable if and only if $\operatorname{spec}(\boldsymbol{S}(k)) \subset \mathbb{C}_-$. We wish to understand a similar relationship between closed-loop stability and $\boldsymbol{K}(s)$. The following lemma gives us the essence of this relationship

11.20 Lemma *If* $s \in \mathbb{C}\backslash\operatorname{spec}(\boldsymbol{A})$ *and* $k \in \mathbb{C}\backslash\operatorname{spec}(\boldsymbol{D})$*, then the following statements are equivalent:*

(i) $\det(s\boldsymbol{I}_n - \boldsymbol{S}(k)) = 0$*;*

(ii) $\det(k\boldsymbol{I}_1 - \boldsymbol{K}(s)) = 0$.

Proof First note that if s and k are as hypothesised, then $\boldsymbol{S}(k)$ and $\boldsymbol{K}(s)$ are well-defined. Since Σ is assumed controllable and observable, the poles of the closed-loop transfer function for the system in Figure 11.9 are exactly the eigenvalues of $\boldsymbol{S}(k)$, as a consequence of Lemma 11.19(ii). However, the closed-loop transfer function is

$$\frac{k^{-1}\boldsymbol{K}(s)}{1 - k^{-1}\boldsymbol{K}(s)} = \frac{\boldsymbol{K}(s)}{k\boldsymbol{I}_1 - \boldsymbol{K}(s)},$$

giving the lemma. ■

From this we have as a consequence the test for closed-loop stability in terms of $\boldsymbol{K}(s)$.

11.21 Corollary $\operatorname{spec}(\boldsymbol{S}(k)) \subset \mathbb{C}_-$ *if and only if* $\operatorname{spec}(\boldsymbol{K}(s)) \subset \mathbb{C}_-$.

Proof The closed-loop system of Figure 11.9 is BIBO stable if and only if $\operatorname{spec}(\boldsymbol{S}(k)) \subset \mathbb{C}_-$ if and only if the poles of the closed-loop transfer function lie in $\mathbb{C}_-$. The result then follows from Lemma 11.20. ■

What we have done to this point how the gain $k \in \mathbb{C}$ and the frequency $s \in \mathbb{C}$ have a symmetric relationship in determining the closed-loop stability of the system in Figure 11.9. Now let us see how this investigation bears fruit.

11.4.2 The characteristic gain and the characteristic frequency functions

What we saw in the preceding discussion was that the two meromorphic functions

$$F_k(s) = \det(s\boldsymbol{I}_n - \boldsymbol{S}(k)) \in \mathbb{R}(s)$$
$$F_s(k) = \det(k\boldsymbol{I}_1 - \boldsymbol{K}(s)) \in \mathbb{R}(k).$$

With this as motivation we have the following result giving us the precise manner in which the gain and frequency are related. First we make two important definitions. To do so, the reader will wish to recall our discussion in Section D.5 on algebraic functions and Riemann surfaces.

Exercises

E11.1 Let $\mathscr{P}(N,D) = \{D + KN \mid K \geq 0\}$ be a (N,D)-polynomial family.

(a) Show that there exists $R_L \in \mathbb{R}[s]$ so that the closed-loop characteristic polynomial for the interconnection of Figure 11.1 is $D + KN$. Explicitly give R_L in terms of N and D.

(b) Provide R_L for the (N,D)-polynomial family of Example 11.1–3.

In Section 11.1.1 we saw that the problem of static output feedback leads naturally to a (N,D)-polynomial family. In the next exercise, you will show that the converse also happens, i.e., that a (N,D)-polynomial family leads naturally to a static output feedback problem.

E11.2 Let $\mathscr{P}(N,D) = \{D+KN \mid K \geq 0\}$ be a (N,D)-polynomial family. Show that there exists a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ so that the closed-loop characteristic polynomial for the static output feedback interconnection of Figure E11.1 is exactly

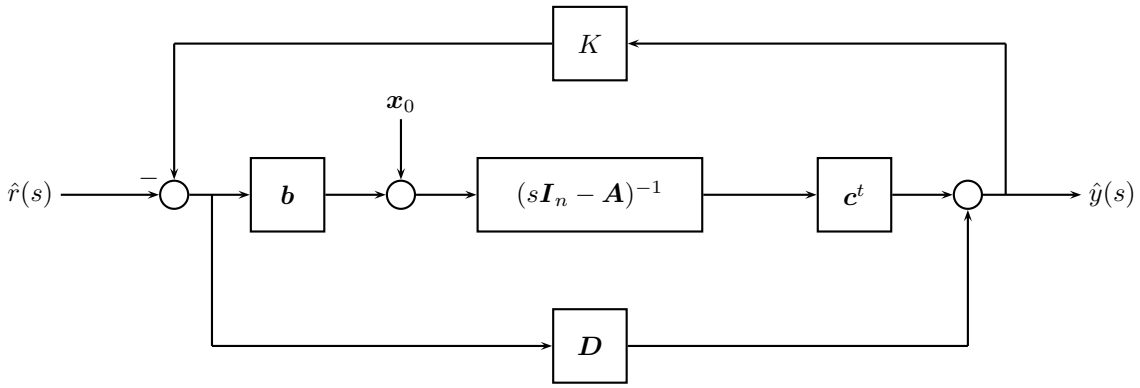


Figure E11.1 Static output feedback and (N,D)-polynomial families

P_K.

E11.3 Prove part (ii) of Lemma 11.4 using the implicit function theorem.

The development of the root-locus properties for an (N,D)-polynomial family assumed that $\deg(N) < \deg(D)$. It is also possible to proceed when confronted with the case when $\deg(N) > \deg(D)$.

E11.4 Let $N, D \in \mathbb{R}[s]$ and assume that D is monic, that either N or $-N$ is monic, and that $\deg(N) > \deg(D)$. Define

$$\tilde{\mathscr{P}}(N,D) = \{D + KN \mid K \geq 0\},$$

$P_K = D + KN$, and let

$$\tilde{\mathrm{RL}}(\tilde{\mathscr{P}}(N,D)) = \overline{\{z \in \mathbb{C} \mid P_K(z) = 0 \text{ for some } K \geq 0\}}.$$

Show that there exists $\tilde{N}, \tilde{D} \in \mathbb{R}[s]$ satisfying the conditions of Definition 11.2 so that $\mathrm{RL}(\mathscr{P}(\tilde{N}, \tilde{D})) = \tilde{\mathrm{RL}}(\tilde{\mathscr{P}}(N,D))$.

E11.5 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \mathbf{0}_1)$ where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 & \\ -1 & 1 \end{bmatrix}.$$

Show using root-locus method that it is not possible to stabilise the system using static output feedback.

Chapter 12

Ad hoc methods II: Simple frequency response methods for controller design

In Chapter 7 we introduced a frequency domain technique, the Nyquist criterion, for studying closed-loop stability. The stability margins introduced in Section 7.2 suggest that the frequency domain techniques may be useful in control design, as they offer tangible objectives for the properties of the closed-loop system. The design ideas we discuss in this chapter are essentially *ad hoc*, but are frequently useful in situations where controller design is not essentially difficult—at least as concerns obtaining closed-loop stability—but where one would like guidance in improving the performance of a controller. The methods we use in this chapter fall under the broad heading of ***loop shaping***, as the objective is to shape the Nyquist plot to have desired properties. In this chapter the emphasis is on choosing *a priori* a form for the controller, then using the design methodology to improve the closed-loop response. In Chapter 15, the design method does not assume a form for the controller, but rather the form of the controller is decided upon by the objectives of the problem.

Contents

12.1 Compensation in the frequency domain

The discussion of the previous few sections has focused on analysis of closed-loop systems using frequency domain techniques. In this discussion, certain key contributors to stability were identified, principally gain and phase margin. Faced with a plant transfer function R_P, one will want to design a controller rational function R_C which has certain desirable characteristics. The term ***compensation*** is used for the process of designing a controller rational function, reflecting the fact that one is "compensating" for deficiencies in the innate properties of the plant. In Section 6.5 we investigated the PID controller, and discussed some of its properties. In this section we introduce another class of controller transfer functions

which are useful in shaping the frequency response. We also discuss PID compensation in a manner slightly different from our previous discussion.

12.1.1 Lead and lag compensation

The first type of compensation we look at deals essentially with the manipulation of the phase of the closed-loop system.

12.1 Definition A ***lead/lag compensator*** is a rational function of the form

$$R_C(s) = \frac{K(1+\alpha\tau s)}{1+\tau s}$$

for $K, \alpha, \tau \in \mathbb{R}$ with $\alpha \notin \{0, 1\}$ and $K, \tau \neq 0$. It is a ***lead compensator*** when $|\alpha| > 1$ and a ***lag compensator*** when $|\alpha| < 1$. •

Note that any rational function of the form $A\frac{s+z}{s+p}$ can be put into the form of a lead/lag compensator by defining

$$\tau = \frac{1}{p}, \quad \alpha = \frac{p}{z}, \quad K = \frac{A}{\alpha}.$$

The form for the lead/lag compensator as we define it is convenient for analysing the nature of the compensator, as we shall see.

The following result gives the essential properties of a lead/lag compensator.

12.2 Proposition *If $R_C(s) = \frac{1+\alpha\tau s}{1+\tau s}$ is a lead/lag compensator then, with $H_C(\omega) = R_C(i\omega)$, the following statements hold:*

(i) $\lim_{\omega\to 0} \angle H_C(\omega) = 0$*;*

(ii) *(a)* $\lim_{\omega\to\infty} \angle H_C(\omega) = 0$ *if* $\alpha > 0$*;*

(b) $\lim_{\omega\to\infty} \angle H_C(\omega) = 180^\circ$ *if* $\alpha < 0$ *and* $\tau < 0$*;*

(c) $\lim_{\omega\to\infty} \angle H_C(\omega) = -180^\circ$ *if* $\alpha < 0$ *and* $\tau > 0$*;*

(iii) $\lim_{\omega\to 0} |H_C(\omega)| = 0\text{dB}$*;*

(iv) $\lim_{\omega\to\infty} |H_C(\omega)| = 20 \log \alpha \text{dB}$*;*

(v) *for $\alpha > 0$, the function $\omega \mapsto |\angle H_C(\omega)|$ achieves its maximum at $\omega_m = (|\tau|\sqrt{\alpha})^{-1}$, and this maximum value is $\phi_m = \arcsin\frac{\alpha-1}{\alpha+1}$. Furthermore,*

(a) *if $\alpha > 1$, $\angle H_C(\omega)$ has a maximum at ω_m and the maximum value is between 0° and 90°, and*

(b) *if $\alpha < 1$, $\angle H_C(\omega)$ has a minimum at ω_m and the minimum value is between -90° and 0°;*

(vi) $|H_C(\omega_m)| = \sqrt{\alpha}$.

Proof (i) This is evident since $\lim_{\omega\to 0} H_C(\omega) = 1$.

(ii) In this case, the assertion follows since $\lim_{\omega\to\infty} H_C(\omega) = \frac{\alpha\tau}{\tau}$. Keeping track of the signs of the real and imaginary parts gives the desired result.

(iii) This follows since $\lim_{\omega\to 0} H_C(\omega) = 1$.

(iv) This is evident since $\lim_{\omega\to\infty} H_C(\omega) = \alpha$.

(v) We compute

$$\text{Re}(H_C(\omega)) = \frac{1+\alpha\tau^2\omega^2}{1+\tau^2\omega^2}, \quad \text{Im}(H_C(\omega)) = \frac{\tau\omega(\alpha-1)}{1+\tau^2\omega^2},$$

so that

$$\measuredangle H_C(\omega) = \measuredangle\big(\tau\omega(\alpha-1) + i(1+\alpha\tau^2\omega^2)\big).$$

In differentiating this with respect to ω, we need to take care of the various branches of arctan. Let us first consider $\tau > 0$ and $\alpha > 1$ so that

$$\measuredangle H_C(\omega) = \arctan \frac{\tau\omega(\alpha-1)}{1+\alpha\tau^2\omega^2}.$$

Differentiating this with respect to ω gives

$$\frac{\mathrm{d}\measuredangle H_C(\omega)}{\mathrm{d}\omega} = \frac{\tau(\alpha(1+\tau^2\omega^2-\alpha\tau^2\omega^2)-1)}{(1+\tau^2\omega^2)(1+\alpha^2\tau^2\omega^2)},$$

which has a zero at $\omega = \sqrt{\alpha\tau^2}^{-1}$. If $\tau > 0$ and $\alpha < 1$, let us define $\beta = \frac{1}{\alpha}$ so that

$$\measuredangle H_C(\omega) = \arctan \frac{\tau\omega(1-\frac{1}{\beta})}{1+\frac{1}{\beta}\tau^2\omega^2}.$$

Differentiating this with respect to ω gives

$$\frac{\mathrm{d}\measuredangle H_C(\omega)}{\mathrm{d}\omega} = \frac{\tau(\beta(1+\tau^2\omega^2-\beta)-\tau^2\omega^2)}{(1+\tau^2\omega^2)(\beta^2+\tau^2\omega^2)},$$

which has a zero at $\omega = \frac{\sqrt{\beta}}{\sqrt{\tau^2}}$. The computations are the same for $\tau < 0$, except that the phase angle has the opposite sign. One may also check that the second derivative is not zero at $\omega = \frac{1}{\sqrt{\alpha}|\tau|}$, and so the function must have a maximum or minimum at this frequency. A straightforward substitution gives

$$\tan\phi_m = \frac{\alpha-1}{2\sqrt{\alpha}}.$$

If $\alpha > 1$ this angle is positive, and if $\alpha < 1$ it is negative. That the value is bounded in magnitude by $90°$ is a consequence of the maximum argument of $1+i\alpha\tau\omega$ being $90°$ for $\tau > 0$ and the minimum argument of $(1+i\tau\omega)^{-1}$ being $-90°$ for $\tau > 0$. A similar statement holds for $\tau < 0$. To get the final form for ϕ_m given in the statement of the proposition, we recall that if $\tan\theta = x$ then $\sin\theta = \frac{x}{\sqrt{1+x^2}}$. Taking $x = \tan\phi_m$ gives the result.

(vi) This is a simple matter of substituting the expression for ω_m into $H_C(\omega)$. ■

12.3 Remarks

1. One sees from part (v) of Proposition 12.2 why the names lead and lag compensator are employed. For a lead compensator with $\alpha, \tau > 0$ the phase angle has a frequency window in which the phase angle is increased, and similarly for a lag compensator with $\alpha, \tau > 0$ there is a frequency window in which the phase is decreased.
2. A plot of the maximum phase lead ϕ_m for $\alpha > 1$ is shown in Figure 12.1. When $\alpha < 1$ we can define $\beta = \frac{1}{\alpha}$ and then see that

$$\phi_m = -\arcsin\frac{\beta-1}{\beta+1}.$$

Therefore, to determine the maximum phase lag when $\alpha < 1$, one can reads off the phase angle from Figure 12.2 at $\frac{1}{\alpha}$, and changes the sign of the resulting angle.

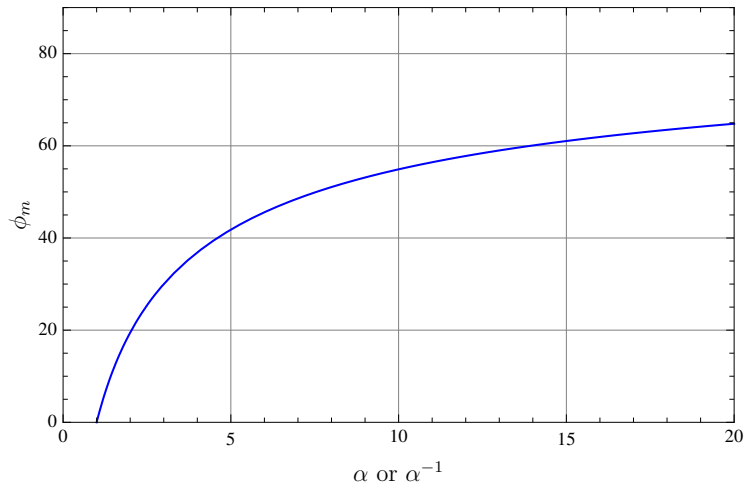


Figure 12.1 Maximum phase shift versus α for lead compensator and versus $\frac{1}{\alpha}$ for lag compensator

Note that the equation $\sin\phi_m = \frac{\alpha-1}{\alpha+1}$ can be solved for α given ϕ_m, and the result is

$$\alpha = \frac{1+\sin\phi_m}{1-\sin\phi_m}. \tag{12.1}$$

3. Although in Proposition 12.2 we have stated the characterisation of lead/lag compensators in fairly general terms, typically one deals with the case when $\alpha, \tau > 0$.
4. If one plots the Bode plot for a lead/lag compensator by hand using the methods of Section 4.3.2, there are two break frequencies, one at $\omega_1 = \frac{1}{\tau}$ and another at $\omega_2 = \frac{1}{\alpha\tau}$. Depending on whether the compensator is lead or lag, one or the other of these frequencies will be the smaller. In either case, the geometric mean of these two break frequencies is, by definition, $\sqrt{\omega_1\omega_2}$ which is equal to ω_m. On a Bode plot for the lead/lag compensator we therefore have $\log\omega_m = \frac{1}{2}(\log\omega_1 + \log\omega_2)$, showing that on the Bode plot, the maximum phase shift occurs exactly between the two break frequencies.
5. The behaviour of the magnitude portion of the frequency response at ω_m is also nice on a Bode plot. By (vi) the magnitude at ω_m is $20\log\sqrt{\alpha} = 10\log\alpha$, and so is half of the limiting value of the magnitude as $\omega \to \infty$. •

Let us look at some simple examples of lead/lag compensators. We illustrate the various things which can happen when choosing parameters in the compensator. But do keep in mind that one normally uses a lead/lag compensator with $\tau > 0$, and either $0 < \alpha < 1$ (lag compensator) or $\alpha > 1$ (lead compensator).

12.4 Examples

1. We look at a lead compensator with transfer function

$$R_C(s) = \frac{1+10s}{1+s},$$

meaning that $\tau = 1$ and $\alpha = 10$. Based on Proposition 12.2 and Remark 12.3–4, it is actually easy to guess what the Bode plot for this transfer function will look like. The two break frequencies are $\omega_1 = 1$ ad $\omega_2 = \frac{1}{10}$. Thus the maximum phase lead will occur at $\log\omega_m = \frac{1}{2}(\log 1 + \log\frac{1}{10}) = -\frac{1}{2}$. The magnitude part of the Bode plot will start at

0dB and turn up at about $\log\omega = -1$ until it reaches $\log\omega = 0$ where it will level off at 20dB. Given that $\alpha = 10$, from Figure 12.1 we determine that the maximum phase lead is about 55° (the calculation gives $\phi_m \approx 54.9°$). The actual Bode plot can be found in Figure 12.2. One should note that the maximum phase shift does indeed occur between

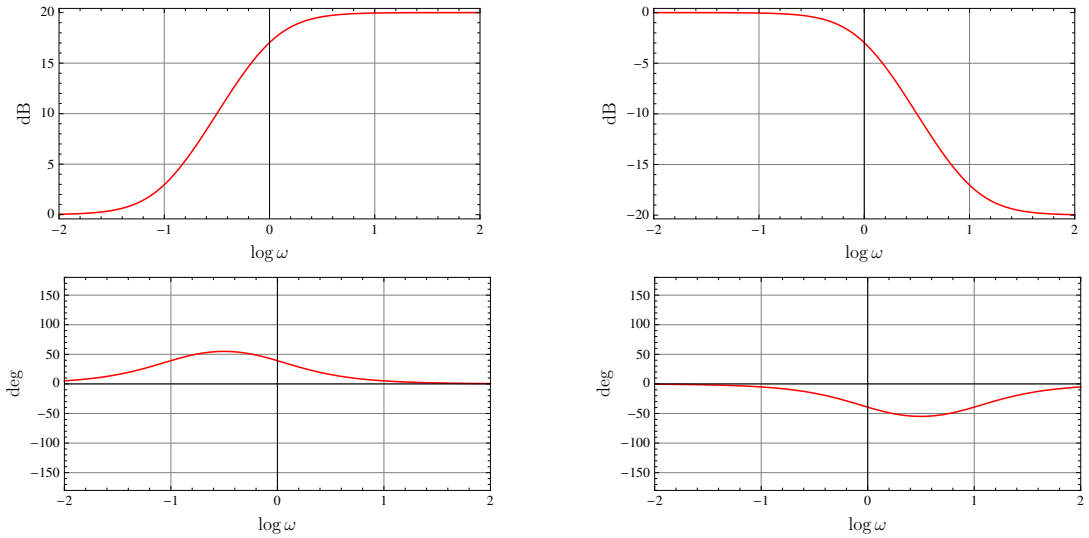


Figure 12.2 Bode plots for a lead compensator $R_C(s) = \frac{1+10s}{1+s}$ (left) and lag compensator $R_C(s) = \frac{1+\frac{1}{10}s}{1+s}$ (right)

the two break frequencies, for example.

2. Here we take the lag compensator

$$R_C(s) = \frac{1 + \frac{1}{10}s}{1 + s},$$

so that $\tau = 1$ and $\alpha = \frac{1}{10}$. Here on determines that the two break frequencies are $\omega_1 = 1$ and $\omega_2 = 10$. Thus $\log\omega_m = \frac{1}{2}(\log 1 + \log 10) = \frac{1}{2}$. The magnitude portion of the Bode plot will start at 0dB and turn down at $\log\omega = 0$ until it reaches $\log\omega = 1$ where it levels off at -20dB. The maximum phase lag can be determined from Figure 12.1. Since $\alpha < 1$ we need to work with $\frac{1}{\alpha}$ which is 10 in this case. Thus, from the graph, we determine that the maximum phase lag is about $-55°$. The actual Bode plot is shown in Figure 12.2. •

12.1.2 PID compensation in the frequency domain

We now look at how a PID controller looks in the frequency domain. First we need to modify just a little the type of PID controller which we introduced in Section 6.5.

12.5 Definition A ***PID compensator*** is a rational function of the form

$$R_C(s) = \frac{K}{s}(1 + T_D s)\big(s + \tfrac{1}{T_I}\big) \tag{12.2}$$

for $K, T_I, T_D > 0$. •

If we expand this form of the PID compensator we get

$$R_C(s) = K\Big(1 + \frac{T_D}{T_I} + \frac{1}{T_I s} + T_D s\Big),$$

and this expression is genuinely different from the PID compensator we investigated in Section 6.5. However, the alteration is not one of any substance (the difference is merely a constant factor $K\frac{T_D}{T_I}$), and it turns out that the form (12.2) is well-suited to frequency response methods.

The following result gives some of the essential features of the transfer function R_C in (12.2).

12.6 Proposition *Let R_C be a PID compensator of the form (12.2) and define $H_C(\omega) = R_C(i\omega)$ for $\omega > 0$. The following statements hold:*

(i) $\lim_{\omega\to 0} \omega|H_C(\omega)| = \frac{K}{T_I}$;

(ii) $\lim_{\omega\to\infty} \frac{1}{\omega}|H_C(\omega)| = KT_D$;

(iii) $\lim_{\omega\to 0} \measuredangle H_C(\omega) = -90°$;

(iv) $\lim_{\omega\to\infty} \measuredangle H_C(\omega) = 90°$;

(v) the function $\omega \mapsto |H_C(\omega)|$ has a minimum of $K\frac{T_D+T_I}{T_I}$ at $\omega_m = \sqrt{T_D T_I}^{-1}$;

(vi) $\measuredangle H_C(\omega_m) = 0$;

Proof The first four assertions are readily ascertained from the expression

$$R_C(s) = K\Big(1 + \frac{T_D}{T_I} + \frac{1}{T_I s} + T_D s\Big)$$

for PID compensator. The final two assertions are easily derived from the decomposition

$$H_C(\omega) = \frac{K}{T_I}\Big((T_D + T_I) + i\frac{T_D T_I \omega^2 - 1}{\omega}\Big)$$

of $H_C(\omega)$ into its real and imaginary parts, and then differentiating the magnitude to find extrema. ■

12.7 Remarks

1. The frequency ω_m at which the magnitude of the frequency response achieves its minimum is the geometric mean of the frequencies $\frac{1}{T_D}$ and $\frac{1}{T_I}$. Thus on a Bode plot, this minimum will occur at the midpoint of these two frequencies.
2. If one redefines a normalised frequency $\tilde{\omega} = T_I\omega$ and a normalised derivative time $\tilde{T}_D = \frac{T_D}{T_I}$, the frequency response for a PID compensator, as a function of the normalised frequency, is given by

$$H_C(\omega) = K\Big(\tilde{T}_D + 1) + i\frac{\tilde{T}_D\tilde{\omega}^2 - 1}{\tilde{\omega}}\Big).$$

This identifies the normalised derivative time and the gain K as the essential parameters in describing the behaviour of the frequency response of a PID compensator. Of these, of course the dependence on the gain is straightforward, so it is really the normalised derivative time $\tilde{T}_D = \frac{T_D}{T_I}$ which is most relevant. •

We have a good idea of what the Bode plot will look like for a PID compensator based on the structure outline in Proposition 12.6. Let's look at two "typical" examples to see how it looks in practice.

12.8 Example In each of these examples, we take $K = 1$ since the effect of changing K is merely reflected in a shift of the magnitude Bode plot by $\log K$.

1. We take $T_D = 1$ and $T_I = 10$. In this case, Proposition 12.6 tells us that the magnitude Bode plot will have its minimum at $\omega_m = \sqrt{10}^{-1}$ or $\log \omega_m = -\frac{1}{2}$. The value of the minimum will be $\frac{T_D+T_I}{T_I} = \frac{11}{10}$. At the frequency ω_m, the phase will be zero. Below this frequency the phase is negative and above it the phase is positive. Putting this together gives the frequency response shown in Figure 12.3.

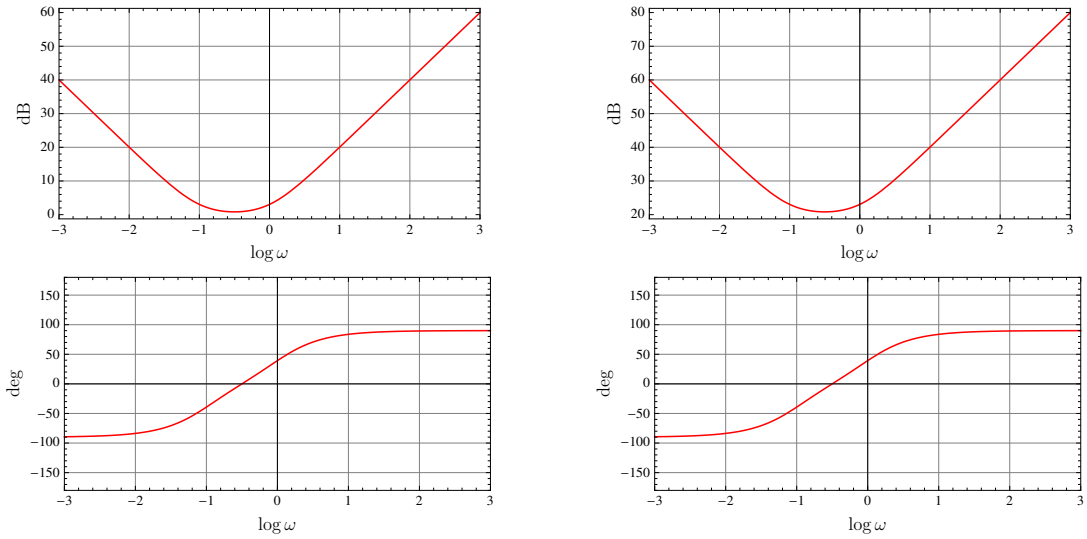


Figure 12.3 The Bode plots for a PID compensator with $K = 1$, $T_D = 1$, and $T_I = 10$ (left) and with $K = 1$, $T_D = 10$, and $T_I = 1$ (right)

2. Now we take $T_D = 10$ and $T_I = 1$. Again we have $\omega_m = \sqrt{10}^{-1}$. The magnitude Bode plot has its minimum at $\omega_m = \sqrt{10}^{-1}$ and the value of the minimum is $\frac{T_D+T_I}{T_I} = 11$. At the frequency ω_m the phase is $0°$, and it is negative below this frequency and positive above it. The Bode plot is shown in Figure 12.3. •

12.2 Design using controllers of predetermined form

In this section we indicate how the procedures of the preceding sections can be helpful in controller design. The basic idea of the frequency response methods we discuss here are that one designs controllers which do various things to the system's frequency response, and so alter both the Bode plot for the loop gain as well as the Nyquist plot. The technique of using the frequency response to design a controller is called ***loop shaping*** since, as we shall see, the idea is to shape the Nyquist plot to have a shape we like.

In Sections 12.2.2 and 12.2.3 we discuss ways to design controller rational functions which shape the Nyquist contour, or equivalently the Bode plot, to have desired characteristics. Let us begin, however, with a simple design situation where one wants to determine the effects of adjusting a gain.

12.2.1 Using the Nyquist plot to choose a gain

We work with the unity gain feedback setup of Figure 6.21 which we reproduce here in Figure 12.4. In this scenario, we have chosen R_C (perhaps it is a PID controller) and

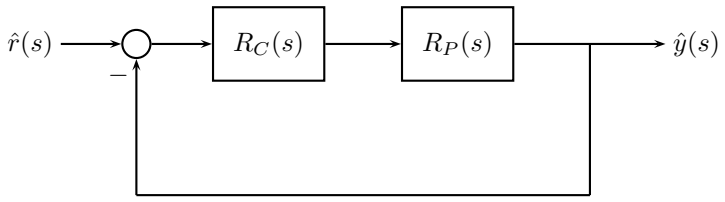


Figure 12.4 Closed-loop system

we wish to tune the gain K. To determine whether a given gain K produces an IBIBO stable closed-loop system, we could, for example, plot the Nyquist contour for the loop gain $R_L = KR_CR_P$. However, if we have to do this for very many gains, it might get a bit tedious. The following result relieves us of this tedium.

12.9 Proposition *For the closed-loop interconnection of Figure 12.4, define $R_L = R_CR_P$, and let $R, r > 0$ be selected so that the (R, r)-Nyquist contour is defined. Then, for $K \neq 0$, the number of encirclements of $-1 + i0$ by the (R, r)-Nyquist contour for KR_L is equal to the number of encirclements of $-\frac{1}{K} + i0$ by the (R, r)-Nyquist contour for R_L.*

Proof A point $s \in \mathbb{C}$ is on the (R, r)-Nyquist contour for KR_L if and only if $\frac{1}{K}s$ is on the (R, r)-Nyquist contour for R_L. That is, the (R, r)-Nyquist contour for KR_L is the (R, r)-Nyquist contour for R_L scaled by a factor of K. From this the result follows. If $K < 0$ the result still holds as the (R, r)-Nyquist contour for KR_L is reflected about the imaginary axis. ■

This result allows one to simply plot the Nyquist contour for R_CR_P, and then determine stability for the closed-loop system with gain K merely by counting the encirclements of $-\frac{1}{K} + i0$. The following example illustrates this.

12.10 Example Let us look at the open-loop unstable system with $R_P(s) = \frac{1}{s-1}$, and we wish to stabilise this using a proportional controller, i.e., $R_C(s) = 1$. Let us use the Nyquist criterion to determine for which values of the gain the closed-loop system is stable. The Nyquist contour for $R_L(s) = R_C(s) = R_P(s) = \frac{1}{s-1}$ is shown in Figure 12.5. Since the loop gain has 1 pole in $\mathbb{C}_+$, by Proposition 12.9, for IBIBO stability of the closed-loop system we must have one counterclockwise encirclement of $-\frac{1}{K} + i0$, provided $K \neq 0$. From Figure 12.5 we see that this will happen if and only if $K > 1$.

Of course, we can determine this directly also. The closed-loop transfer function is

$$T_L(s) = \frac{KR_L(s)}{1 + KR_L(s)} = \frac{K}{s + K - 1}.$$

The Routh/Hurwitz criterion (if you wish to hit this with an oversized hammer) says that we have closed-loop IBIBO stability if and only if $K > 1$. •

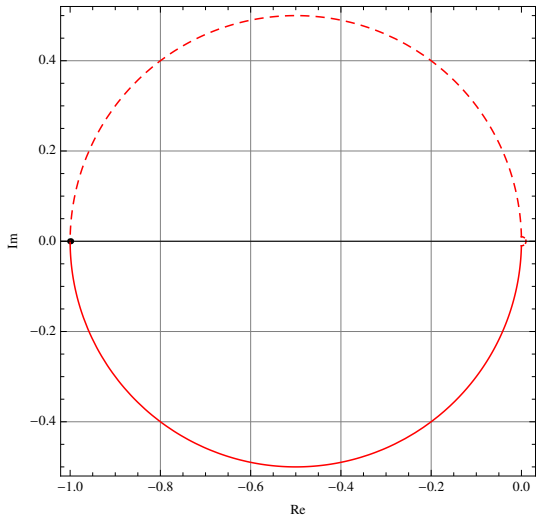


Figure 12.5 Stabilising an open-loop unstable system using the Nyquist criterion

12.2.2 A design methodology using lead and integrator compensation

In the previous section we assumed that we had in hand the nature of the controller we would be using, and that the only matter to account for was the gain. In doing this, we glossed over how one can choose the controller rational function R_C. There are, in actuality, many ways in which one can use PID control elements, in conjunction with lead/lag compensators, to make a frequency response look like what we need it to look like. We shall explore just a few design methodologies, and these can be seen as representative of how one can approach the problem of controller design.

We look at a methodology which first employs a lead compensator to obtain a desired phase margin for stability, and then combines this with an integrator for low frequency performance and disturbance rejection. The rough idea is described in the following "algorithm."

12.11 A design methodology Consider the closed-loop system of Figure 12.6. To design a controller

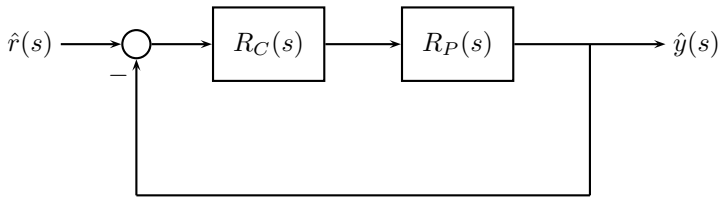


Figure 12.6 Closed-loop system for loop shaping

transfer function R_C, proceed as follows.

(i) Select a gain crossover frequency for your system. This choice of frequency is based upon requirements for transient response since larger gain crossover frequencies are related to shorter rise times. One cannot expect to specify an arbitrary gain crossover

frequency, however. The bandwidth of a system will be limited by the bandwidth of the system components.

(ii) Design a phase lead controller rational function $R_{C,1}$ so that the closed-loop system will meet the phase and gain margin requirements.

(iii) Use phase lag compensation or integral control to boost the low frequency gain of the controller transfer function. These terms will assist in the tracking of step inputs, and the rejection of disturbances. Take care here not to degrade the stability of the system.

(iv) Should the plant not have sufficient attenuation at high frequencies, add a term which "rolls off" at high frequencies to alleviate the systems susceptibility to high frequency noise. Most plants will have sufficient attenuation at high frequencies, however. •

We can illustrate this design methodology with an example. Although this example is illustrative, it is simply not possible to come up with a design methodology, or an example, which will work in all cases.

12.12 Example We consider the problem of controlling a mass in space in the absence of gravity. We again suppose that the motive force is applied through a propeller, as in Example 6.60. The governing differential equations are

$$m\ddot{y}(t) = f(t) + d(t),$$

where f is the input force from the propeller, and $d(t)$ is a disturbance force on the system. In the Laplace transform domain, the block diagram is as depicted in Figure 12.7. In this

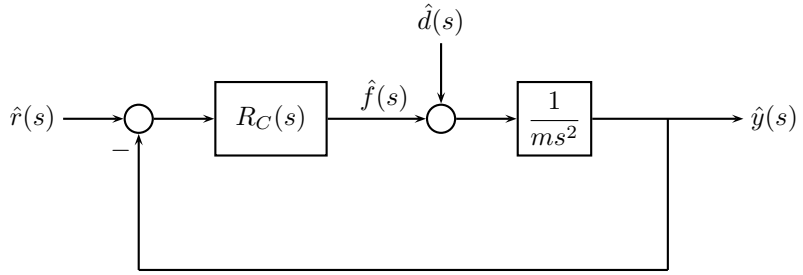


Figure 12.7 Block diagram for controller design for mass

example we fix $m = 1$. Let us suppose that we have determined that a respectable value for the gain crossover frequency is ω_{gc} (in stating this, we are implicitly assuming that we will have just one gain crossover frequency). For suitable stability we require that the phase margin for the system be at least 40°. We also ask that the system have zero steady-state error to a step disturbance.

Now we can go about ascertaining how to employ our design methodology to obtain the specifications. First we look at the Bode plot for the plant transfer function $R_P(s) = \frac{1}{ms^2}$. This is shown in Figure 12.8. At the gain crossover frequency $\omega_{\text{gc}} = 10\text{rad/ sec}$ (i.e., $\log \omega_{\text{gc}} = 1$), the phase of the plant transfer function is 180°. Indeed, the phase of the plant transfer function is 180° for *all* frequencies. To get the desired phase margin, we employ a lead compensator. Indeed, we should choose a lead compensator which gives us the phase margin we desire, and hopefully with some to spare. We need a minimum phase lead of 40°, and from (12.1) we can see that this means that $\alpha \geq \frac{1+\sin 40}{1-\sin 40} \approx 4.60$. Let us be really conservative and choose $\alpha = 20$. Now, we not only need for the phase shift to be at least 40°, we need it to

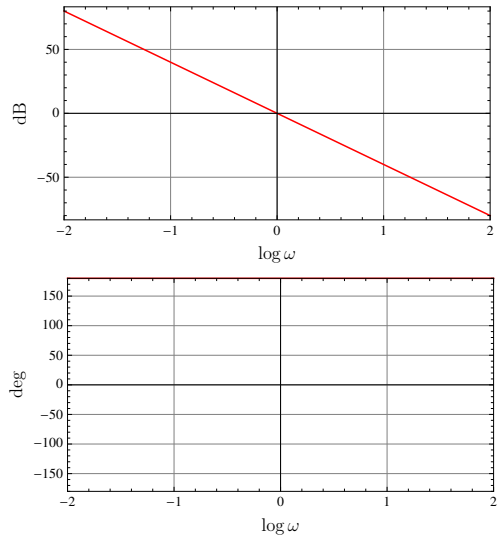


Figure 12.8 Plant Bode plot for mass system

exceed this value at a specific frequency. But Proposition 12.2 tells us how to manage this: we need to specify ω_m. We have $\omega_m = (|\tau|\sqrt{\alpha})^{-1}$, and solving this for τ with $\alpha = 20$ and $\omega_m = 10$ gives $\tau = \frac{1}{20\sqrt{5}} \approx 0.022$. We have been sufficiently conservative with our choice of α that we can afford to make τ a nice number, so let's take $\tau = \frac{1}{50}$. With this specification, the lead compensator is

$$R_{C,1}(s) = \frac{1+2s}{1+\frac{1}{50}s}.$$

Let's evaluate where we are for the moment. In Figure 12.9 we give the Nyquist and Bode

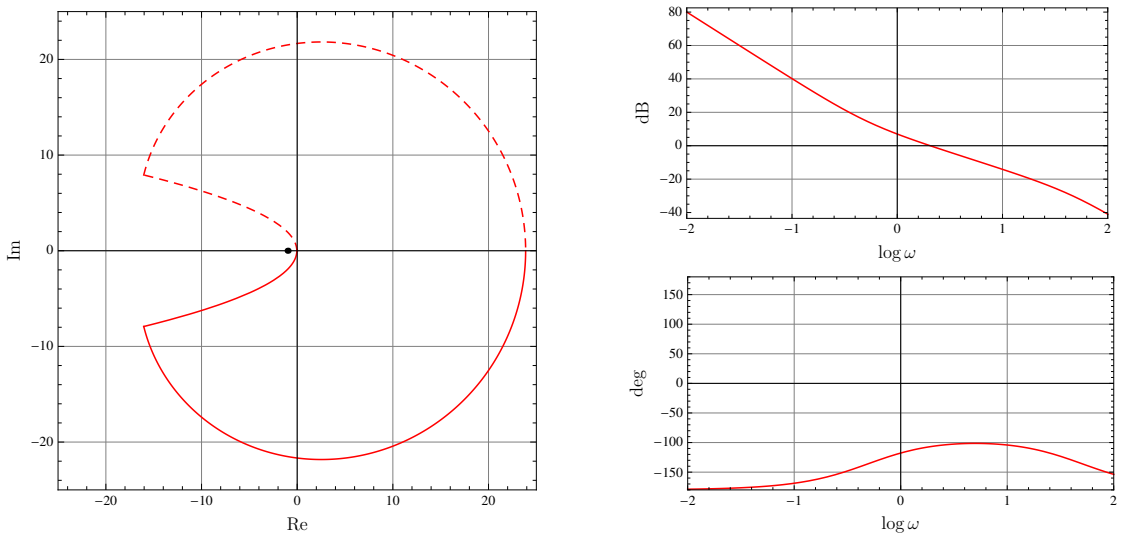


Figure 12.9 The $(100, 0.25)$-Nyquist contour and the Bode plot for mass and lead compensator

plots for $R_{C_1} R_P$. From the Nyquist plot we see that there are no encirclements of $-1 + i0$, and so the system is IBIBO stable. From the Bode plot, we see that the phase is about $-105°$

(about $-104.2°$ to be more precise) at ω_{gc}, which gives a phase margin of about $75°$, well in excess of what we need. There is a problem at the moment, however, in that the desired gain crossover frequency is, in fact, not a gain crossover frequency since the transfer function has less magnitude than we desire at ω_{gc}. We can correct this by boosting the gain of the lead compensator. The matter of just how much to boost the gain is easily resolved. At $\omega = \omega_{gc}$, from the Bode plot of Figure 12.9 we see that the magnitude is about -15dB (about -14.1dB to be more precise). So we need to adjust K so that $20 \log K = 15$ or $K \approx 5.62$. Let us take $K = 5\frac{1}{2}$. The Nyquist and Bode plots for the gain boosted led compensator are provided in Figure 12.10. Also note from the Bode plot that the magnitude is falling off at

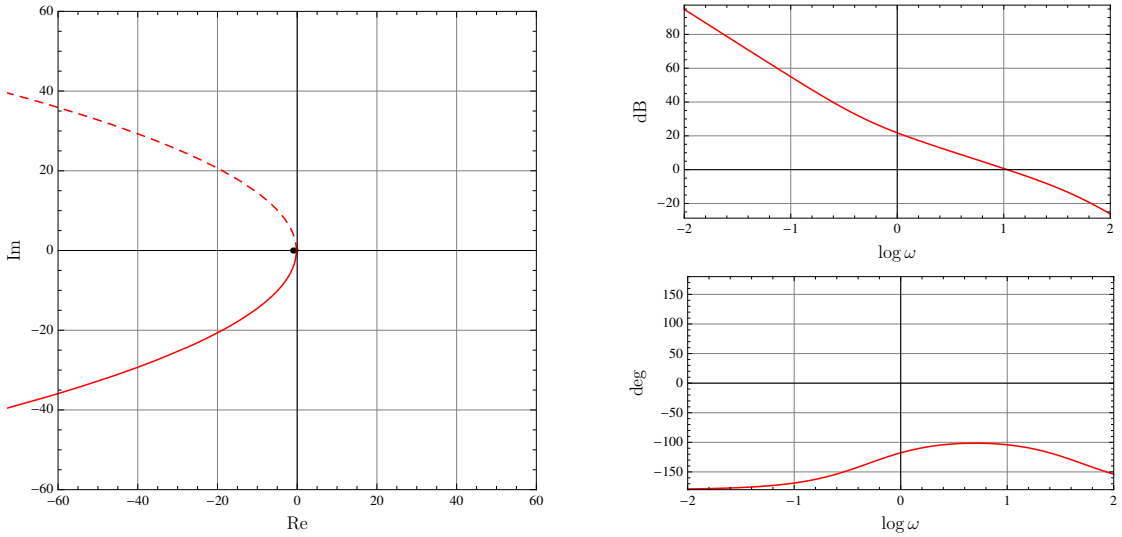


Figure 12.10 The $(100, 0.25)$-Nyquist contour and the Bode plot for mass and lead compensator with increased gain

40dB/decade, which means that the system is type 2 by Proposition 8.14. Therefore this system meets our objective to track step inputs. In Figure 12.11 we plot the response of the

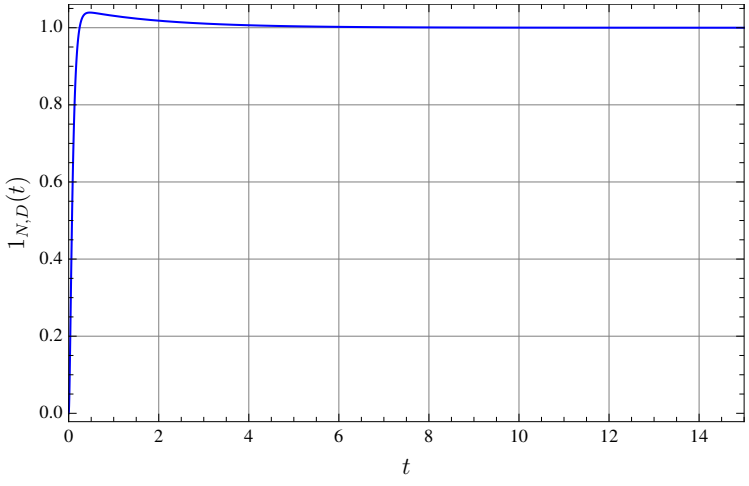


Figure 12.11 Step response of mass with lead compensator

closed-loop system to a unit step input which we obtained by using Proposition 3.40. Note that the error decays to zero as predicted.

Let's now look at how the system handles step disturbances since we have asked that these be handled gracefully (specifically, we want no error to step disturbances). In Figure 12.12 we display the response of the system to a unit step disturbance (no input). Clearly this is

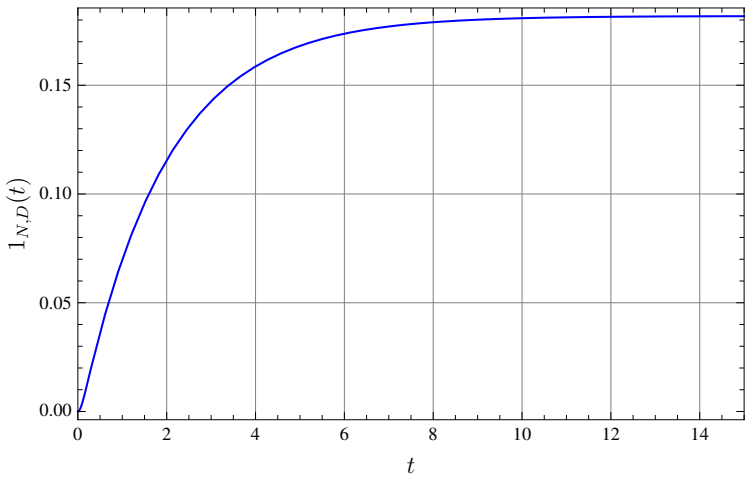


Figure 12.12 Step disturbance response of mass with lead compensator

not satisfactory, given our design specifications. To see what is going on here, we determine the transfer function from the disturbance to the output to be

$$T_d(s) = \frac{R_P(s)}{1 + R_{C,1}(s)R_P(s)}.$$

We compute

$$\lim_{s\to 0} T_d(s) = \frac{1}{\lim_{s\to 0} \frac{1}{R_P(s)} + \lim_{s\to 0} R_{C,1}(s)} = \frac{1}{\lim_{s\to 0} R_{C,1}(s)} = 1,$$

and so the system has disturbance type 0 with respect to this disturbance. This is not satisfactory for the purpose of eliminating error resulting from a step input, so we need to repair this in some way. As we have seen in the past, a good way to do this is to add an integrator to the controller transfer function. Let's see how this works. We work with the new controller rational function

$$R_{C,2}(s) = \frac{K_I}{s} + \frac{1 + 2s}{1 + \frac{1}{10}s}.$$

First note that $\lim_{s\to 0} R_{C,2}(s) = \infty$ and so this immediately means that the system type with respect to the disturbance is at least 1 (and is in fact exactly 1). We look to ascertain how changing the integrator gain K_I affects system stability. In Figure 12.13 the Bode plots are shown for various values of K_I. We can see that, as expected since an integrator will decrease the phase, the phase margin becomes worse as K_I increases. What's more, one can check the Nyquist criterion to show that the system is in fact unstable for K sufficiently large. Thus we choose a not too large integrator gain of $K_I = \frac{1}{10}$. Let's see how this choice of controller

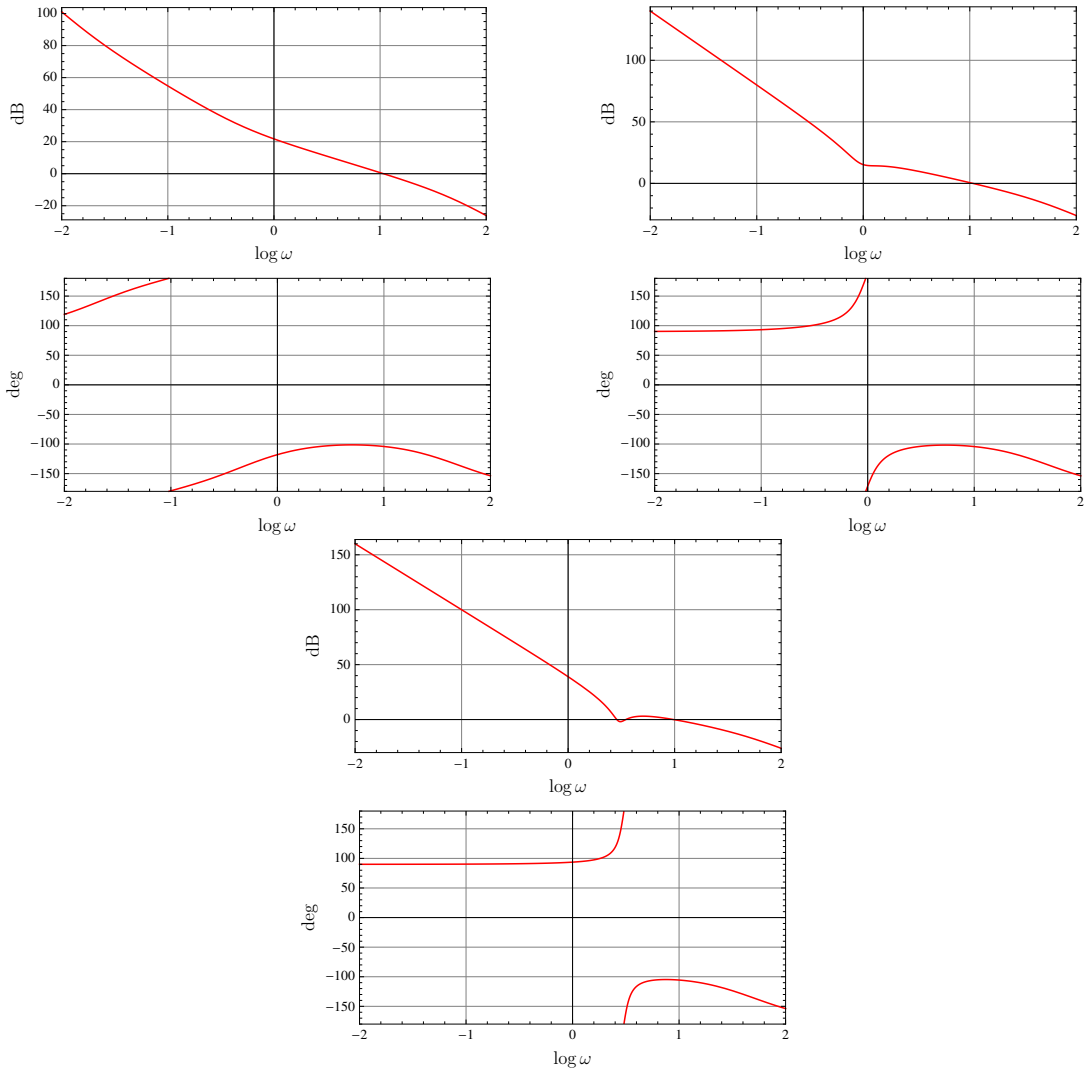


Figure 12.13 Bode plots for mass with lead and integrator compensation: $K_I = \frac{1}{10}$ (top left), $K_I = 10$ (top right), and $K_I = 100$ (bottom)

performs. First, in Figure 12.14 we display the step response (no disturbance) of the system with the integrator added to the controller. Note that the overshoot and the settling time have increased from the situation when we simply employed the lead compensator, but may be considered acceptable. If they are not, then one should iterate the design methodology until satisfactory performance is achieved. Also in Figure 12.14 we display the response of the system to a step disturbance (no input). Note that this response ends up at zero as we have ensured by designing the controller as we have done. The decay of the effect of the error is quite slow, however. One may wish to improve this by increasing the integrator gain.

Now that we have added the integrator to reject disturbances, we need to ensure that the other performance specifications are still met. A look at Figure 12.13 shows that our gain crossover frequency has decreased. We may wish to repair this by further boosting the gain on the lead compensator. •

This example, even though simple, shows some of the tradeoffs which must take place in designing a controller to meet sometimes conflicting performance requirements.

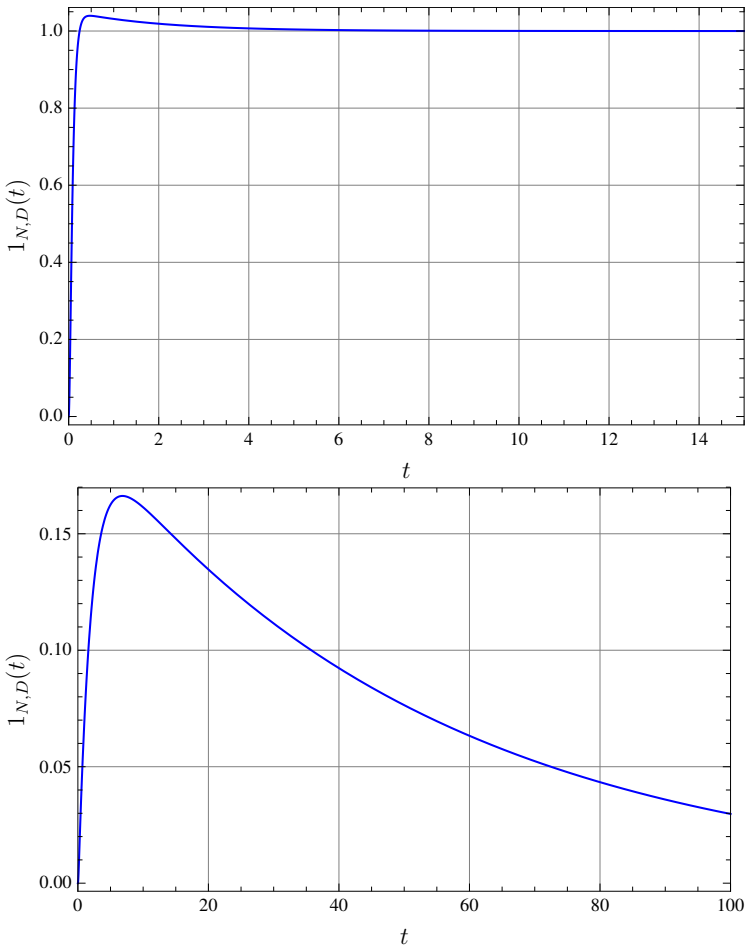


Figure 12.14 Step response of mass (top) and response to step disturbance (bottom) with lead and integrator compensation

12.2.3 A design methodology using PID compensation

Let's see in this section how one can employ a PID controller in a systematic way in the frequency domain to obtain a desired response. The rough idea of PID design is outlined as follows.

12.13 A design methodology Consider the closed-loop system of Figure 12.6. To design a controller transfer function R_C, proceed as follows.

(i) Select a gain crossover frequency consistent with the demands for quick response and the limitations of the system components.

(ii) Commit to T_I being quite a bit larger than T_D.

(iii) Adjust T_D so that the phase margin requirements are met.

(iv) Adjust K so that the gain crossover frequency is as desired. •

Let us illustrate this in an example.

12.14 Example **(Example 12.12 cont'd)** We again take the plant transfer function

$$R_P(s) = \frac{1}{ms^2}.$$

The Bode plot for this transfer function is reproduced in Figure 12.15. We suppose that we

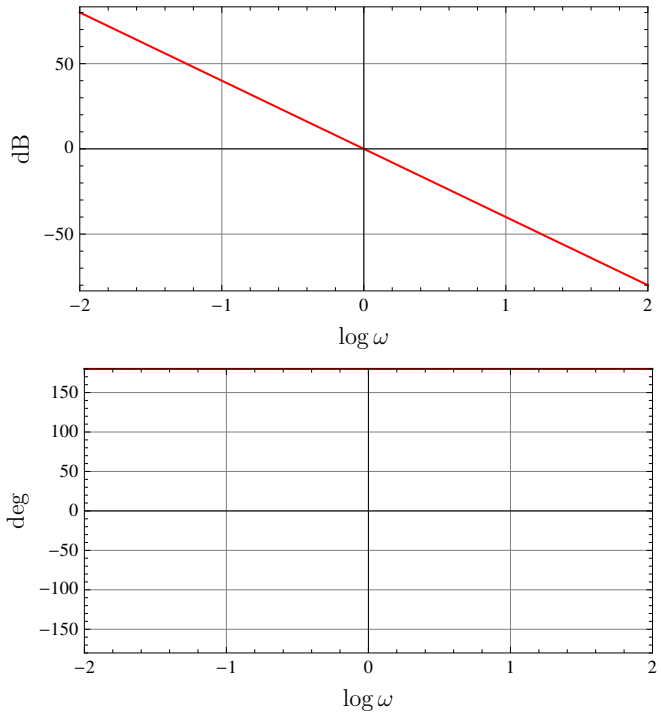


Figure 12.15 Plant Bode plot for mass system

are given the criterion of having a phase margin of at least 65° at as high a gain crossover frequency as possible.

Let us decide to go with $\frac{T_D}{T_I} = \frac{1}{5}$. In order to achieve the prescribed phase margin, we must use the PID compensator to produce a phase for the loop gain $R_C R_P$ which is somewhere negative and greater than $-115°$. Note that given the relative degree of the plant and the PID controller, the phase for large frequencies will approach $-90°$. The contribution to the phase made by the controller changes sign at $\omega_m = \sqrt{T_D T_I}^{-1}$. Given our choice of $\frac{T_D}{T_I} = \frac{1}{5}$, this means that $\omega_m = \frac{1}{\sqrt{5}} T_D^{-1}$. Thus choosing a large derivative time will decrease the frequency at which the phase becomes negative. In Figure 12.16 we show the situation for $T_D = \frac{1}{5}$ and $T_D = 2$. For the smaller value of T_D, we see that the phase margin requirements are met at a quite high frequency (around 20rad/ sec). However, a peek at the Nyquist contour for $T_D = \frac{1}{5}$ shows that there are 2 clockwise encirclements of $-1 + i0$. Since $R_C R_P$ has no poles in $\mathbb{C}_+$, this means the system is not IBIBO stable for the given PID parameters. When the derivative time is $T_D = 2$, the phase requirement is met at roughly 2rad/ sec, but now the Nyquist contour is predicting IBIBO stability. Thus we see that we should expect there to be a tradeoff between stability and performance in this design methodology. Let us fix $T_D = 1$ (and hence $T_I = 5$), for which we plot the Bode plot and Nyquist contour in Figure 12.17.

We now choose the gain K in order to make the gain crossover frequency large. From Figure 12.17 we see that when the phase is $-115°$ the frequency is roughly given by $\log \omega = 0.5$, so we take $\omega_{\text{gc}} = 10^{0.5} \approx 3.16$. At this frequency the magnitude of the frequency response is about -10dB. Thus we need to choose K so that $20 \log K = 10$, or $K \approx 3.16$. Let's make

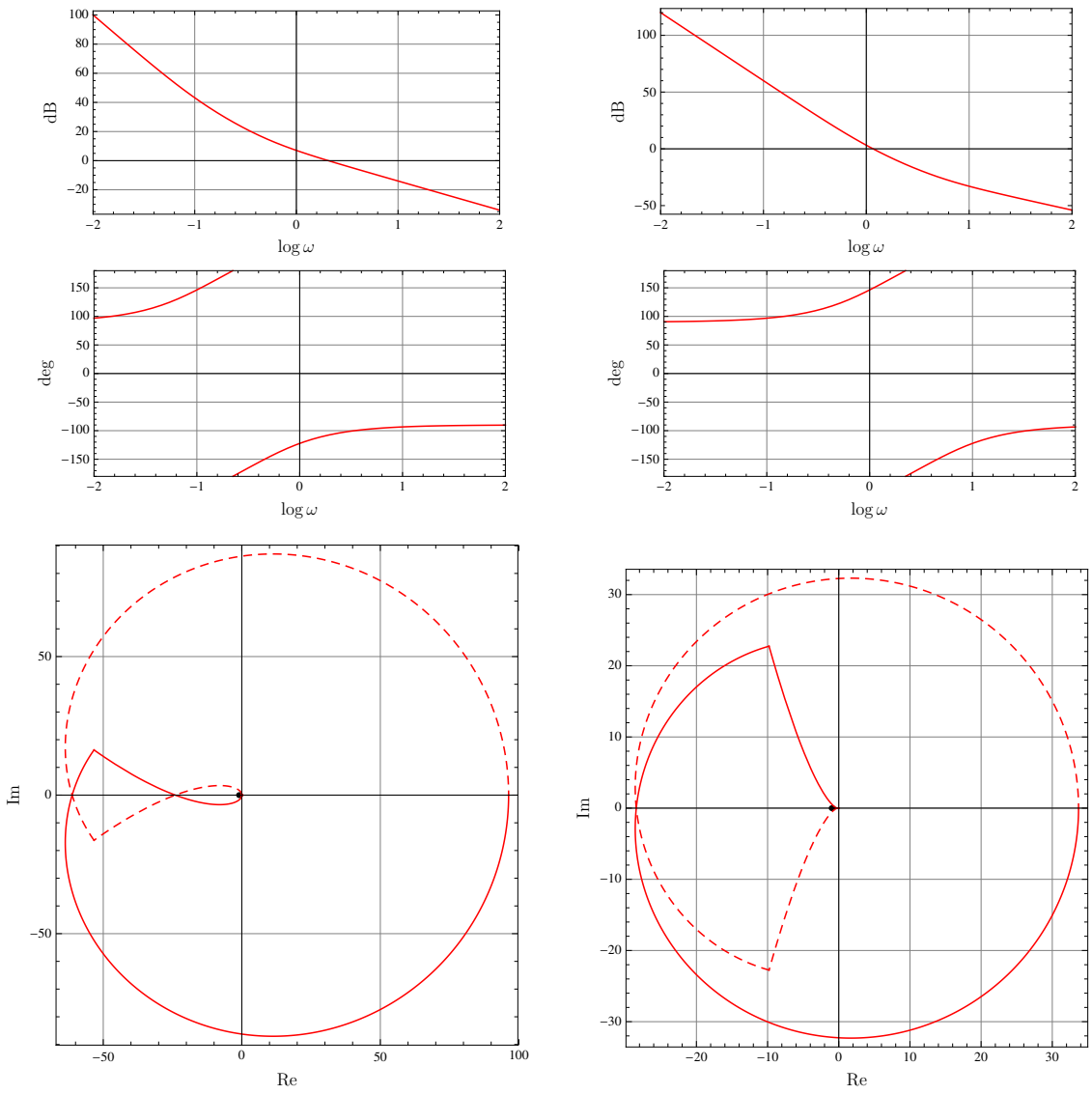


Figure 12.16 Effect of varying PID derivative time for $R_P = \frac{1}{s^2}$: $T_D = \frac{1}{5}$ (left) and $T_D = 2$ (right). In both cases $K = 1$ and $T_I = 5T_D$, and Nyquist plots are on top and Bode plots are on the bottom

this $K = 3\frac{1}{4}$ so that our final controller is given by

$$R_C(s) = \frac{3\frac{1}{4}}{s}(1 + T_D s)\big(s + \tfrac{1}{T_I}\big).$$

The Bode and Nyquist plots for the corresponding loop gain are shown in Figure 12.18. From the Nyquist plot we see that the system is IBIBO stable, and we also see that out phase margin requirements are satisfied.

In Figure 12.19 is the response of the system to a unit step input. We may also compute the effect of a disturbance entering between the controller and the plant. The corresponding transfer function is

$$T_d(s) = \frac{R_P(s)}{1 + R_C(s)R_P(s)} = \frac{s}{s^3 + 3\frac{1}{4}s^2 + 3\frac{9}{10}s + \frac{13}{20}},$$

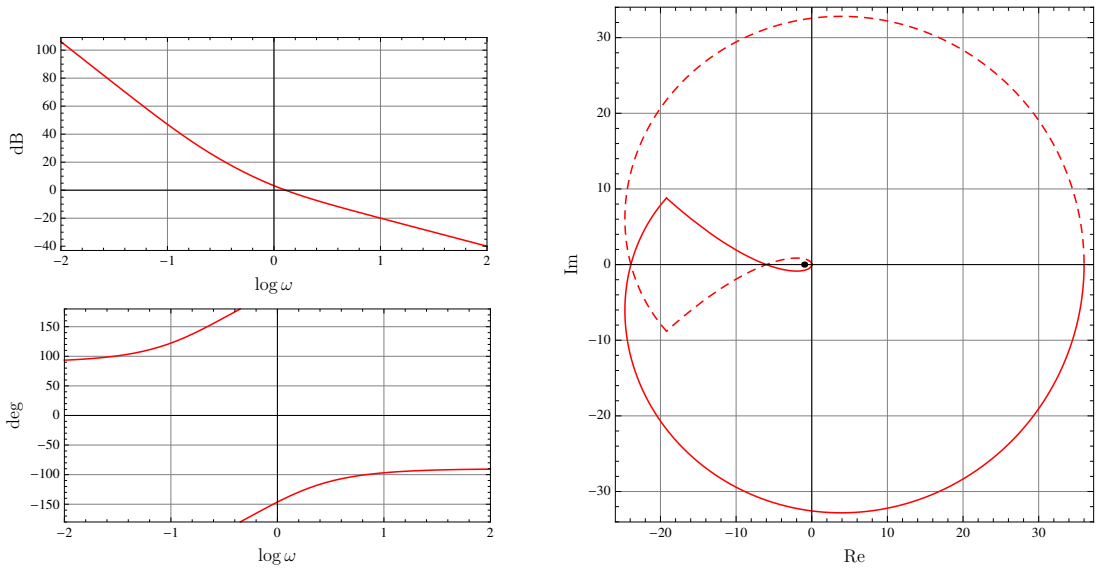


Figure 12.17 Nyquist plot (left) and Bode plot (right) for mass with PID controller and $K = 1$, $T_D = 1$, and $T_I = 5$

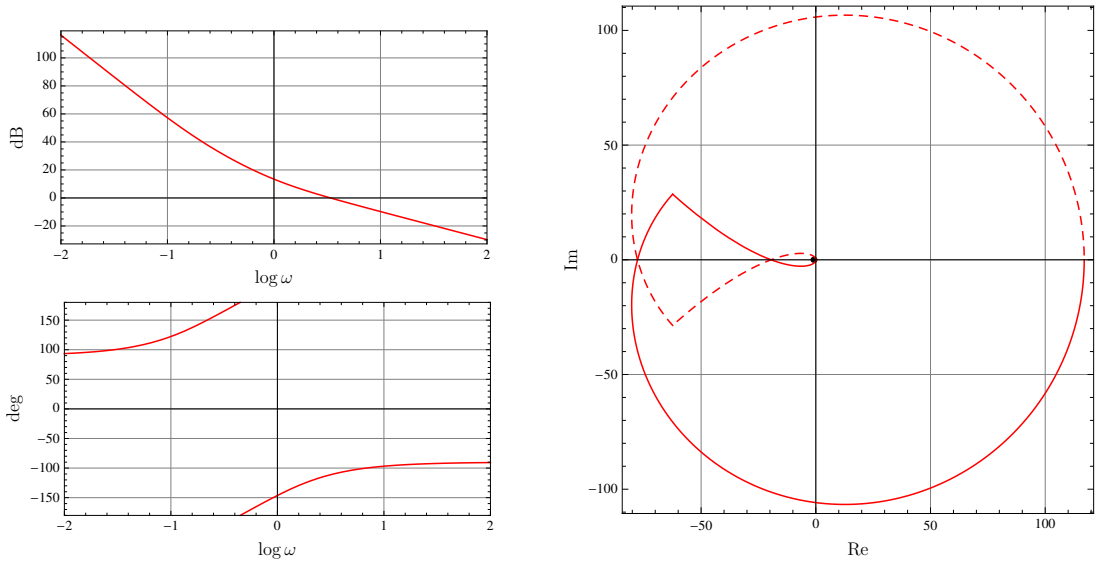


Figure 12.18 Nyquist plot (left) and Bode plot (right) for mass with PID controller and $K = 3\frac{1}{4}$, $T_D = 1$, and $T_I = 5$

and the response to a unit step is shown in Figure 12.20. •

12.2.4 A discussion of design methodologies

The above examples give an idea of how one can use ideas of frequency response in designing controller rational functions. Indeed, it is interesting to compare the examples since they use the same plant with different control design strategies. For example, the one thing we notice straight away is the better disturbance response for the PID compensator.

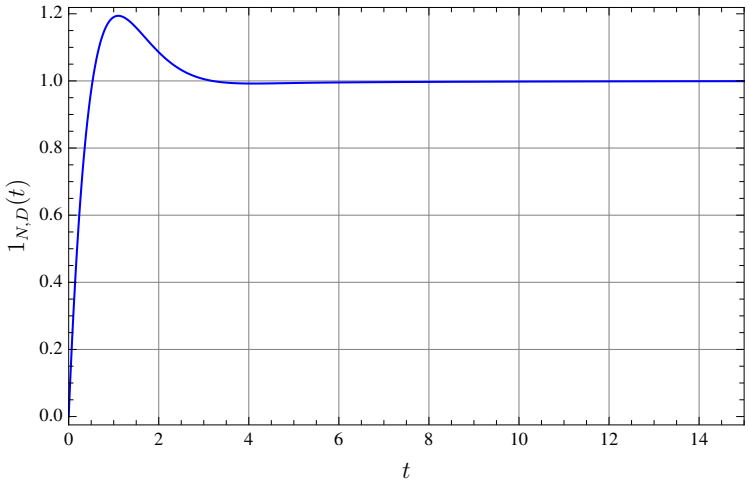


Figure 12.19 Step response of mass with PID controller

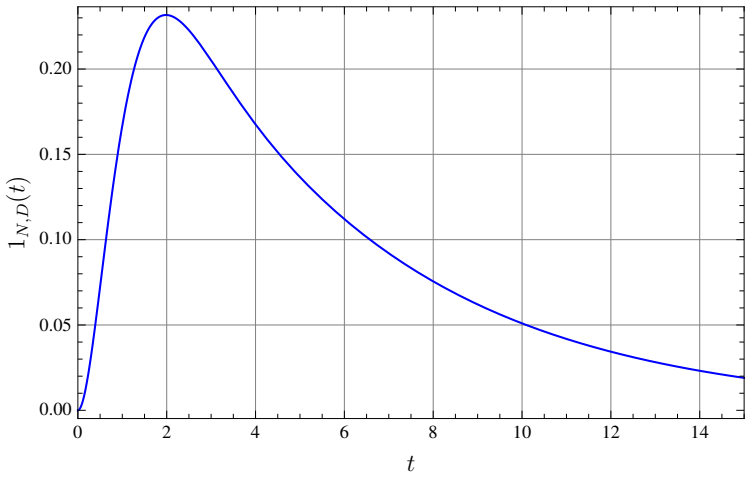


Figure 12.20 Unit step disturbance of mass with PID controller

This should come as no surprise since the reset time for the integrator is significantly larger for the PID compensator.

We should emphasise that the process is rarely as straightforward as the examples suggest, and that in practice one will usually have to iterate the design process to account for the various tradeoffs which exist between stability and performance. Also, the methodologies we discuss above can only be expected to have a reasonable chance at success when the plant transfer function R_P has no poles or zeros in $\mathbb{C}_+$. If R_P *does* have poles in $\mathbb{C}_+$, then one must design the controller R_C so that the point $-1 + i0$ is encircled. If R_P has zeros in $\mathbb{C}_+$, then it typically turns out to be more difficult to design a stabilising controller that meets goals for stability margins. In such cases, mundane considerations of gain and phase margin become less satisfactory measures of a good design unless weighed against other factors.

12.3 Design with open controller form

In the previous section, the emphasis was on tuning controllers of a specified type. This can be a difficult exercise for plants which are unstable and/or nonminimum phase (see Exercise E12.6). The difficulty is that be fixing the form on the controller, one also limits what can be done to the sensitivity function and the closed-loop transfer function. It may be possible that the plant will not allow stabilisation by a controller of a certain form.

12.4 Summary

In this section, we have been able to assimilate our knowledge gained to this point to generate some design methods for controller rational functions. One should note that our methods are not guaranteed to work, but for "simple" applications, they provide a starting point for serious controller design. Let us review some of the basic ideas.

1. The basis for the frequency response design methodology is the Nyquist criterion. This can be a somewhat subtle notion, so it would be best to understand it. A good way to do this is to study the proof of the Nyquist criterion since it is quite simple, given the Principle of the Argument.
2. We have presented two design methodologies using frequency response: one for lead compensation and the other for PID compensation. One should make sure to understand the *idea* behind these design methodologies.

Exercises

The next three exercises have to do with designing circuits to implement various controllers. Although we look at only a three specific controller transfer functions, it is possible, in principle, to design a circuit using only passive resistors, capacitors, and inductors, to realise any transfer function in RH_∞^+. A means of doing this was first pointed out in the famous paper of Bott and Duffin [1949].

E12.1 Consider the circuit of Figure E12.1.

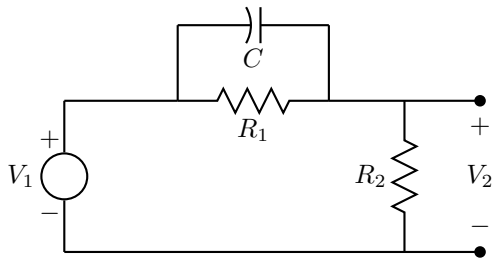


Figure E12.1 Circuit for lead compensation

(a) Determine the differential equation governing the output voltage V_2 given the input voltage V_1.
(b) Convert this differential equation to a transfer function, and show that the resulting transfer function is that of a lead compensator.
(c) Determine expressions for K, α, and τ in the standard form for a lead compensator in terms of R_1, R_2, and C.

E12.2 Consider the circuit of Figure E12.2.

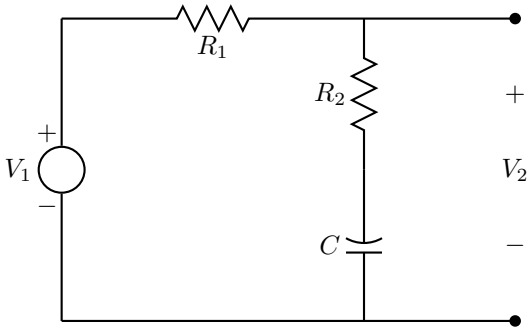


Figure E12.2 Circuit for lag compensation

(a) Determine the differential equation governing the output voltage V_2 given the input voltage V_1.
(b) Convert this differential equation to a transfer function, and show that the resulting transfer function is that of a lag compensator.
(c) Determine expressions for K, α, and τ in the standard form for a lag compensator in terms of R_1, R_2, and C.

E12.3 Consider the circuit of Figure E12.3.

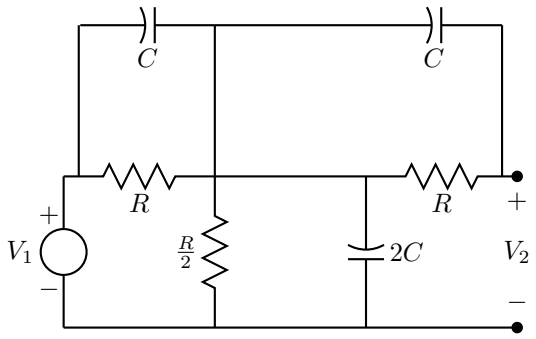


Figure E12.3 Circuit for notch filter

(a) Determine the differential equation governing the output voltage V_2 given the input voltage V_1.

(b) Convert this differential equation to a transfer function, and show that the resulting transfer function is that of a notch filter.

(c) Determine expressions for K, α, and τ in the standard form for a lag compensator in terms of $R$and C.

E12.4 Consider the plant transfer function $R_P(s) = \frac{2}{s^2(s+2)}$. Design a PID controller using frequency domain methods which produces an IBIBO stable system with a phase margin of at least 65° at as large a gain crossover frequency as possible. Check the stability of your design using the Nyquist criterion, and produce the step response, and the response to a step disturbance which enters the loop between the controller and the plant.

E12.5 Consider a controller transfer function $R_C(s) = K\left(\frac{1+\alpha\tau s}{1+\tau s} + \frac{1}{T_I s}\right)$ where K, α, τ, and T_I are all finite and nonzero. Determine a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ so that $T_\Sigma = R_C$.

E12.6 In this problem you will design a controller for the unstable, nonminimum phase plant shown in Figure E12.4.

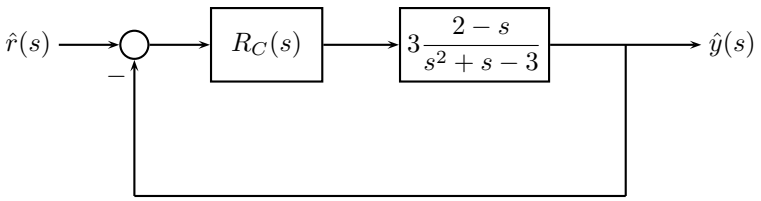


Figure E12.4 A closed-loop system with an unstable, nonminimum phase plant

(a) Why is the plant unstable? nonminimum phase?

(b) Produce the Nyquist and Bode plots for the plant transfer function. Is the closed-loop system stable with $R_C(s) = 1$?

(c) Is it possible to design a proportional controller $R_C(s) = K$ which renders the closed-loop system IBIBO stable?

(d) We will design a stabilising controller by manipulating the Nyquist contour to look like the cartoon in Figure E12.5. Sketch the Bode plot for such a Nyquist

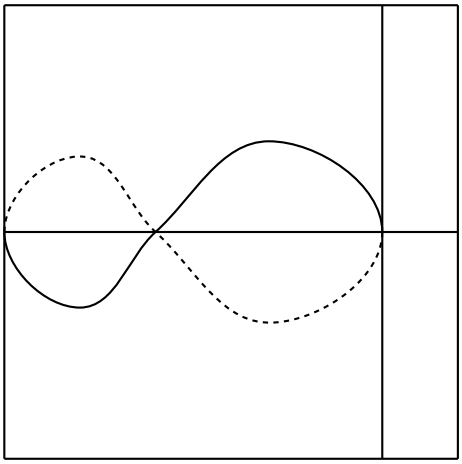


Figure E12.5 The desired Nyquist contour

contour, being particularly concerned with the phase of the Bode plot. Assume that the loop gain $R_L = R_C R_P$ is strictly proper and that its maximum magnitude occurs at $\omega = 0$.

(e) Design a phase lead controller R_C with the property that the Bode plot for $R_L = R_C R_P$ has the Bode plot phase which looks qualitatively like that you drew in part (d). For the moment, design R_C to that $R_C(0) = 1$. Plot the Nyquist contour for R_L to verify that it has the shape shown in Figure E12.5.

(f) With the lead compensator you have designed, is your closed-loop system IBIBO stable? By adding a gain K to the lead compensator from part (e), it can be ensured that the system will be IBIBO stable for some values of K. Using your Nyquist contour from part (e), for which values of K will the closed-loop system be IBIBO stable? Pick one such value of K and produce the Nyquist contour to verify that the controller you have designed does indeed render the closed-loop system IBIBO stable.

Congratulations, you have just designed a controller for an unstable, nonminimum phase plant, albeit a contrived one. Let us see how good this controller is.

(g) Comment on the gain and phase margins for your system.

(h) Produce the step response for the closed-loop system, and comment on its performance.

E12.7 In this exercise you will be given a plant transfer function for which it is not possible to achieve arbitrary design objectives. The transfer function is

$$R_P(s) = \tilde{R}_P(s)\frac{s-z}{s-p}$$

for $s, p \in \mathbb{R}$ positive. Thus, we know that our plant has an unstable real pole p, and a nonminimum phase real zero z. All other poles and zeros are assumed to be stable and minimum phase, respectively. Let R_C be a stable, minimum phase controller transfer function and define $R_L = R_C R_P$.

(a) Define a Finish this

E12.8 Consider the plant transfer function

$$R_P(s) = \frac{1}{s^3 - s^2 + s - 1},$$

and the lead/lag controller transfer function

$$R_C(s) = \frac{K(1+\alpha\tau s)}{1+\tau s}.$$

Answer the following two questions.

(a) Show that it is not possible to design a lead/lag controller R_C for which $R_C \in \mathscr{S}(R_P)$.

(b) Show that for any proper, second-order plant transfer function $R_P \in \mathbb{R}(s)$ there exists a lead/lag controller R_C for which $R_C \in \mathscr{S}(R_P)$.

E12.9 Consider the plant transfer function

$$R_P(s) = \frac{1}{s^4 - s^3 + s^2 - s + 1},$$

and the PID controller transfer function

$$R_C(s) = K\Big(1 + T_D s + \frac{1}{T_I s}\Big).$$

Answer the following two questions.

(a) Show that it is not possible to design a PID controller R_C for which $R_C \in \mathscr{S}(R_P)$.

(b) Show that for any proper, third-order plant transfer function $R_P \in \mathbb{R}(s)$ there exists a PID controller R_C for which $R_C \in \mathscr{S}(R_P)$.

E12.10 Consider the coupled tank system of Exercises E1.11, E2.6, and E3.17. Take as system parameters $\alpha = \frac{1}{3}$, $\delta_1 = 1$, $A_1 = 1$, $A_2 = \frac{1}{2}$, $a_1 = \frac{1}{10}$, $a_2 = \frac{1}{20}$, and $g = 9.81$. As output, take the level in tank 2.

(a) Design a PID controller for the system that achieves a phase margin of at least 75° with as large a gain crossover frequency as possible.

We now wish to simulate the behaviour of the system with the controller that you have designed. The states of the system are nominally h_1 and h_2. However, since the controller involves an integration, an additional state will need to be defined. With this in mind, answer the following question.

(b) Suppose that a reference output $h_{2,\mathrm{ref}}$ has been specified, and that the control u is specified by the PID controller you designed in part (a). Develop the *nonlinear* state differential equations for the system, starting with the equations you linearised in Exercise E2.6. As indicated in the preamble to this part of the problem, you will have *three* state equations.

(c) With the controller that you have designed, simulate the differential equations from part (b) with initial conditions equal to the equilibrium initial conditions, and subject to a reference step input of size $\frac{1}{4}$.

(d) Plot the height in both tanks, and comment on the behaviour of your controller.

Chapter 13

Advanced synthesis, including PID synthesis

We saw in Chapters 11 and 12 that one can use *ad hoc* methods to choose PID parameters that can serve as acceptable starting points for final designs of such controllers. These classical methods, while valuable in terms of providing some insight into the process of control design, can often be surpassed in effectiveness by more modern methods. In this chapter we survey some of these, noting that they rely on some of the more sophisticated ideas in the text to this point. This explains why they may not form a part of the typical introductory text dealing with PID control.

Contents

13.1 Ziegler-Nichols tuning for PID controllers

The ideas we discuss here are the result of an empirical investigation by Ziegler and Nicols [1942]. We give two methods for specifying PID parameters. The first will be applicable quite often, especially for BIBO stable plants, whereas the second makes some assumptions about the nature of the system. In each case, the criterion for optimisation was the minimisation of the integral of the absolute value of the error due to a unit step input. Thus one minimises

$$\int_0^\infty |e(t)|\,\mathrm{d}t,$$

where $e(t)$ is the difference between the step response and the desired response to a step input. In particular, it is assumed that this integral is finite. Since the methods in this section are *ad hoc*, they should not be thought of as being guarantees, but rather as a good

starting point for beginning a final tuning of the parameters. A methodology for this is the subject of [Hang, Åstrom, and Ho 1990].

13.1.1 First method

We shall work with systems in input/output form. Thus let R_P be a plant transfer function with c.f.r. (N, D), and let $1_{N,D}(t)$ be the step response for the system. We assume that (N, D) is BIBO stable. Define a parameter $\sigma \in \mathbb{R}_+$ by

$$\sigma = \sup_{t \geq 0} |\dot{1}_{N,D}(t)|.$$

Thus σ is the maximum slope of the step response. Let $t_\sigma \in \mathbb{R}$ be the smallest time satisfying $|\dot{1}_{N,D}(t)| = \sigma$. Thus t_σ is the time at which the slope of the step response reaches its maximum value. Typically this time is unique. We then define $\tau \in \mathbb{R}$ by

$$\tau = t_\sigma - \frac{1_{N,D}(t_\sigma)}{\sigma}.$$

The meaning of τ is as shown in Figure 13.1. With the parameters σ and τ at hand, we can

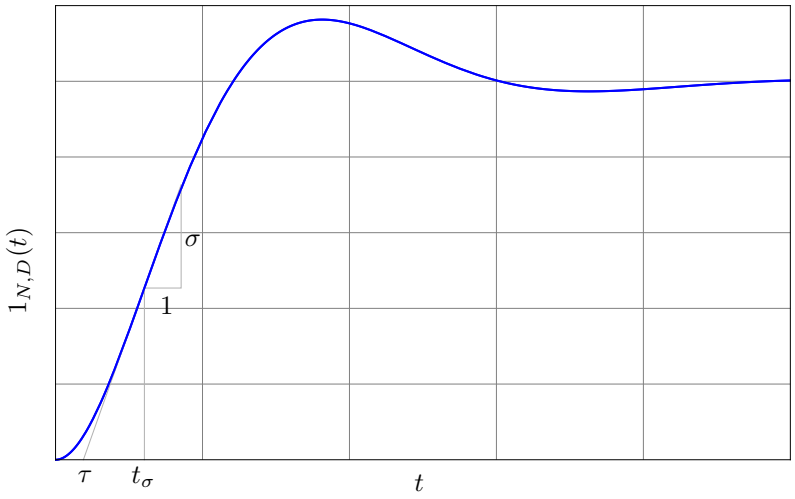


Figure 13.1 The definitions of σ and τ for Ziegler-Nicols PID tuning

specify the parameters in a P, PI, or PID control law of the form

$$\begin{aligned} R_C(s) &= K && \text{(P)} \\ R_C(s) &= K\Big(1 + \frac{1}{T_I s}\Big) && \text{(PI)} \\ R_C(s) &= K\Big(1 + T_D s + \frac{1}{T_I s}\Big) && \text{(PID).} \end{aligned} \tag{13.1}$$

In Table 13.1 we tabulate choices of the parameter values for various types of controllers.

13.1.2 Second method

We resume the setting above with a plant R_P with c.f.r. (N, D). In this method, we make the following assumption.

Table 13.1 Controller parameters for the first Ziegler-Nicols tuning method

Controller type	Controller parameters
P	$K = \frac{1}{\sigma\tau}$
PI	$K = \frac{9}{10\sigma\tau}$, $T_I = \frac{10}{3\tau}$
PID	$K = \frac{6}{5\sigma\tau}$, $T_I = 2\tau$, $T_D = \frac{\tau}{2}$

13.1 Assumption The closed-loop transfer function with proportional control,

$$T = \frac{KR_P}{1 + KR_P},$$

is BIBO stable for K very near zero. Furthermore, if the gain K is increased from $K = 0$, there exists a critical gain K_u where exactly one pair of the poles of the transfer function crosses the imaginary axis, with the remaining poles in $\mathbb{C}_-$. •

Under the conditions of the assumption, at the gain K_u the step response will exhibit an oscillatory behaviour for sufficiently large times. If the poles on the imaginary axis are at $\pm i\omega_u$, the period of this oscillation will be $T_u = \frac{2\pi}{\omega_u}$. With this information, the criterion for choosing the parameters in the controller (13.1) are as given in Table 13.2.

Table 13.2 Controller parameters for the second Ziegler-Nicols tuning method

Controller type	Controller parameters
P	$K = \frac{K_u}{2}$
PI	$K = \frac{9K_u}{20}$, $T_I = \frac{5}{6}T_u$
PID	$K = \frac{6K_u}{10}$, $T_I = \frac{T_u}{2}$, $T_D = \frac{T_u}{8}$

Note that there are some simple cases in which the Ziegler-Nicols criterion will not apply (see Exercise E13.1). However, for cases where the method does apply, it can be a useful starting point. It also has the advantage that it can be applied to an experimentally obtained step response.

13.1.3 An application of Ziegler-Nicols tuning

Let us apply the Ziegler-Nicols tuning methods to an example. Suppose that we have a rotor spinning on a shaft supported by bearings. The angular position of the rotor will satisfy a differential equation of the form

$$J\ddot{\theta} + d\dot{\theta} + k\theta = u,$$

where J is the inertia of the rotor, d accounts for the viscous friction in the bearings, k is the shaft spring constant, and u is the torque applied to the shaft. This then gives a plant transfer function

$$R_P(s) = \frac{1}{Js^2 + ds + k}.$$

Let us take $J = 1$, $d = \frac{1}{10}$, and $k = 2$.

Let us look at the first Ziegler-Nicols method. The step response for the plant is shown in Figure 13.2. One may compute σ and τ graphically. However, in this case it is possible to

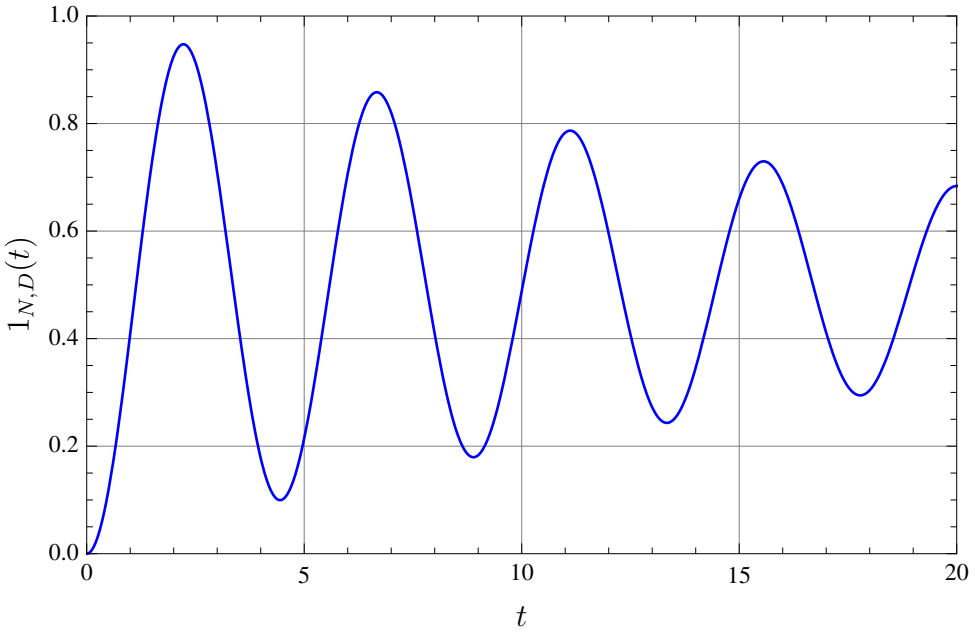


Figure 13.2 Step response for rotor on shaft

compute these numerically since the step response is a known function on t. To find t_σ one determines where $\ddot{1}_{N,D}(t_\sigma) = 0$, where (N, D) is the c.f.r. for R_P. We compute $t_\sigma \approx 1.08$ and then easily compute $\sigma \approx 0.70$ and $\tau \approx 0.40$. The values for the P, PI, or PID parameters are shown in Table 13.3. The three corresponding step responses for the closed-loop transfer

Table 13.3 Controller parameters for the rotor example first Ziegler-Nicols tuning method

Controller type	Controller parameters
P	$K \approx 3.83$
PI	$K \approx 3.45$, $T_I \approx 8.55$
PID	$K \approx 4.60$, $T_I = 0.78$, $T_D \approx 0.19$

function, normalised so that they have the same steady state value, are shown in Figure 13.3. Notice that the closed-loop performance is actually rather abysmal. Furthermore, it is quite evident that what is needed in more derivative time. If we arbitrarily set $T_D = 1$ in the PID controller, the resulting step response is shown in Figure 13.4. Note that the response is now more settled. This exercise points out that the Ziegler-Nicols tuning method does not produce guaranteed effective control laws. Indeed, the system we have utilised is quite benign, and it still needed some adjustment to work well.

Let us now move onto the second of the Ziegler-Nicols methods. We cannot use the rotor example, because it does not satisfy Assumption 13.1. So let us come up with a plant

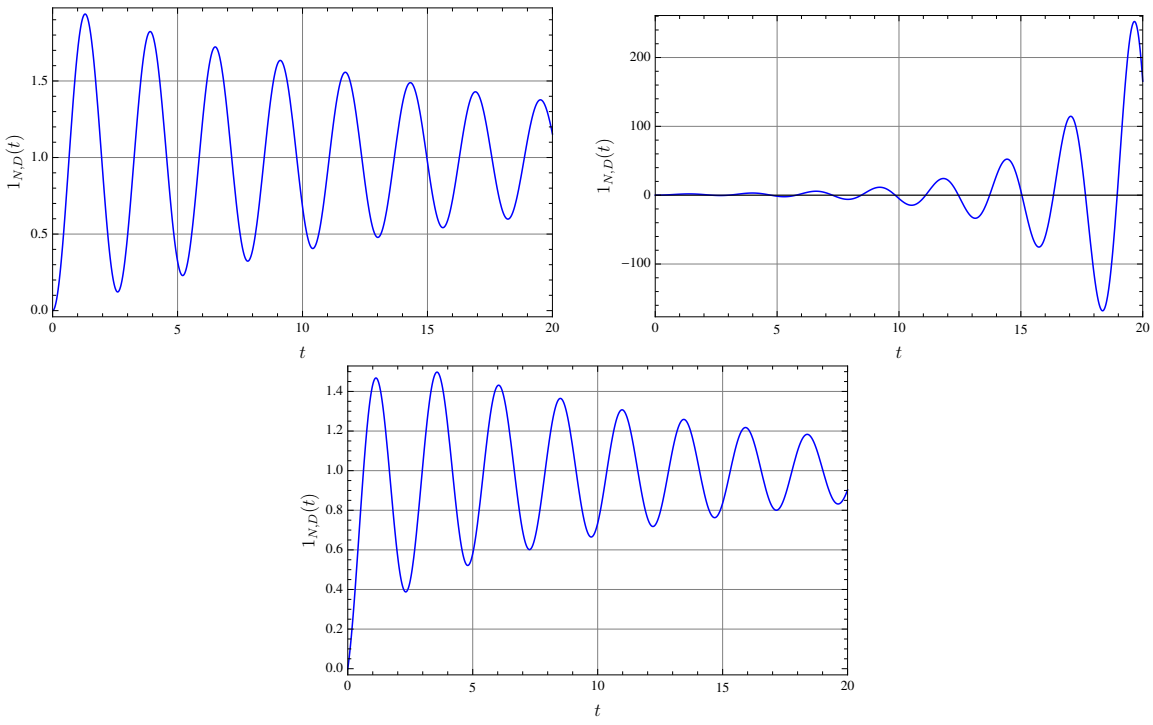


Figure 13.3 Normalised step responses for rotor example using first Ziegler-Nicols tuning method: P (top left), PI (top right), and PID (bottom)

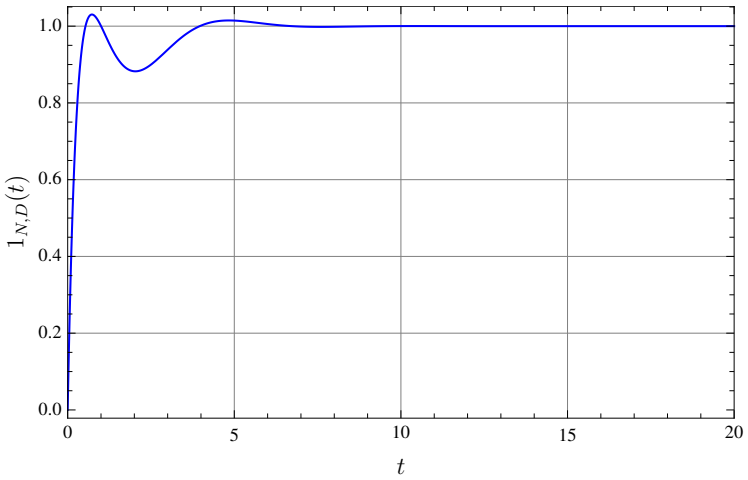


Figure 13.4 Normalised step response for rotor example using first Ziegler-Nicols tuning method and derivative time adjusted to $T_D = 1$

transfer function that *does* work. An example is

$$R_P(s) = \frac{1}{s^3 + 3s^2 + 4s + 1}.$$

In Figure 13.5 is a plot of the behaviour of the poles of the closed-loop system as a function of K with $R_C(s) = K$. As we can see, the roots behave as specified in Assumption 13.1, so

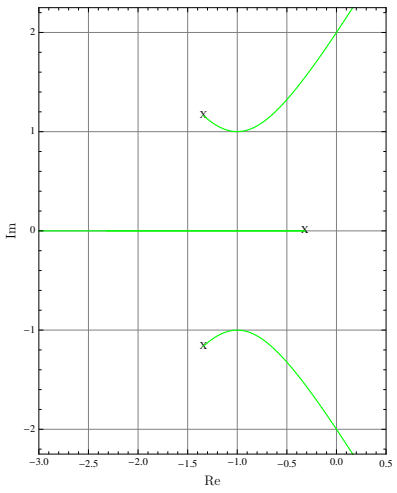


Figure 13.5 Behaviour of poles for plant $R_P(s) = \frac{1}{s^3+3s^2+4s+1}$ and $R_C(s) = K$ as K varies

we can proceed with that design methodology. The method asks that we find the critical gain K_u for which the roots cross the imaginary axis. One can do this by trial and error, looking at the step response. For example, in Figure 13.6 we plot the step response for two

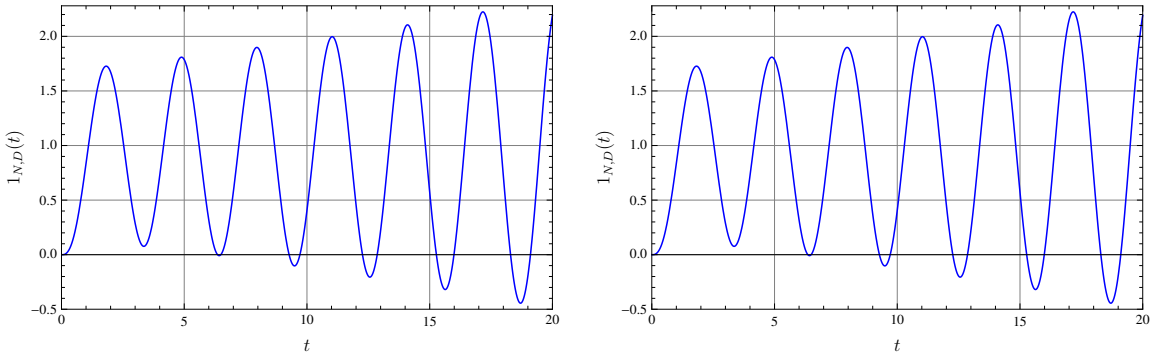


Figure 13.6 The step response of the plant $R_P(s) = \frac{1}{s^3+3s^2+4s+1}$ and $R_C(s) = K$ for $K = 10$ (left) and $K = 12$ (right)

values of K. As we can see, for the plot on the left $K < K_u$ and for the plot on the right $K > K_u$. One can imagine iteratively finding something quite close to K_u by looking at such plots. However, I found K_u by numerically determining when the real part of the poles for the closed-loop transfer function

$$T(s) = \frac{KR_P(s)}{1 + KR_P(s)} = \frac{1}{s^3 + 3s^2 + 4s + 1 + K}$$

are zero. The answer is approximately $K_u \approx 11.0$. With this value of K the imaginary part of the poles is then $\omega \approx 2.0$. Thus we have $T_u = \frac{2\pi}{\omega_u} \approx 3.14$. The corresponding values for the PID parameters are displayed in Table 13.4, and the normalised closed-loop step responses are shown in Figure 13.7. The PID response is respectable, but might benefit from more derivative time given its largish overshoot.

Table 13.4 Controller parameters for example using second Ziegler-Nicols tuning method

Controller type	Controller parameters
P	$K \approx 5.5$
PI	$K \approx 4.95$, $T_I \approx 2.62$
PID	$K \approx 6.6$, $T_I = 1.57$, $T_D \approx 0.39$

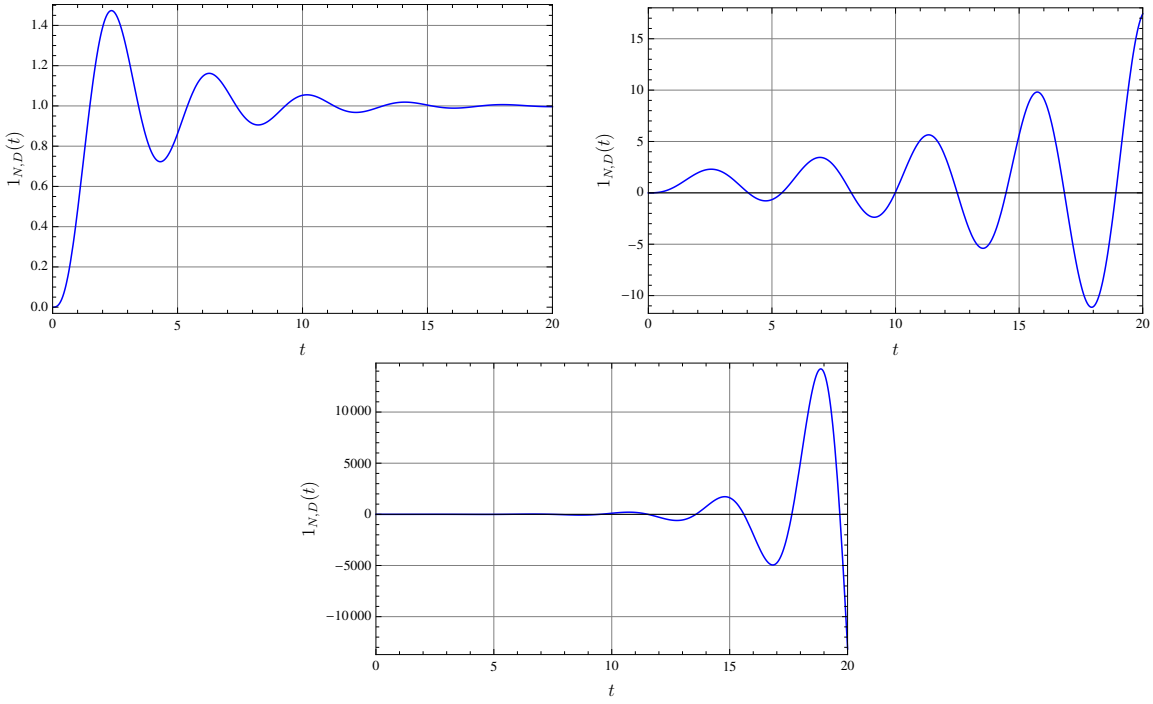


Figure 13.7 Normalised step responses for example using second Ziegler-Nicols tuning method: P (top left), PI (top right), and PID (bottom)

13.2 Synthesis using pole placement

In this section we use a form of pole placement to indicate how to select PID parameters based upon the location of poles. Clearly one cannot choose a PID controller to place the poles anywhere for a general controller. However, in this section we see exactly how well we can do.

13.2.1 Pole placement using polynomials

In this section we engage in a rather general discussion of the closed-loop poles using purely polynomial methodology. We consider the standard unity gain feedback loop of Figure 13.8. Our objective is to characterise some of the possible closed-loop characteristic polynomials for the system. The main result is the following.

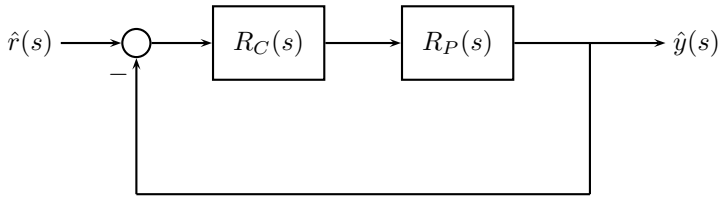


Figure 13.8 Feedback loop for polynomial pole placement

13.2 Theorem *Consider the interconnection of Figure 13.8 and let (N_P, D_P) be the c.f.r. for R_P, supposing $\deg(D_P) = n$. The following statements hold:*

(i) if $\deg(N_P) \leq n-1$ and if $P \in \mathbb{R}[s]$ is monic and degree $2n-1$, then there exists a proper $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) so that the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P;

(ii) if $\deg(N_P) \leq n$ and if $P \in \mathbb{R}[s]$ is monic and degree $2n$, then there exists a strictly proper $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) so that the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P.

Proof The following result contains the essential part of the proof.

1 Lemma **(Sylvester's theorem)** *For polynomials*

$$\begin{aligned} P(s) &= p_n s^n + p_{n-1} s^{n-1} + \cdots + p_1 s + p_0 \\ Q(s) &= q_n s^n + q_{n-1} s^{n-1} + \cdots + q_1 s + q_0, \end{aligned}$$

*with $p_n^2 + q_n^2 \neq 0$, define their **eliminant** as the $2n \times 2n$ matrix*

$$\boldsymbol{M}(P,Q) = \begin{bmatrix} p_n & 0 & \cdots & 0 & q_n & 0 & \cdots & 0 \\ p_{n-1} & p_n & \cdots & 0 & q_{n-1} & q_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \\ p_1 & p_2 & \cdots & p_n & q_1 & q_2 & \cdots & q_n \\ p_0 & p_1 & \cdots & p_{n-1} & q_0 & q_1 & \cdots & q_{n-1} \\ 0 & p_0 & \cdots & p_{n-2} & 0 & q_0 & \cdots & q_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & p_0 & 0 & 0 & \cdots & q_0 \end{bmatrix}.$$

Then P and Q are coprime if and only if $\det \boldsymbol{M}(P,Q) \neq 0$.

Proof Note that if P and Q are not coprime then there exists $z \in \mathbb{C}$ so that

$$\begin{aligned} P(s) &= (s-z)(\tilde{p}_{n-1} s^{n-1} + \cdots + \tilde{p}_1 s + \tilde{p}_0) \\ Q(s) &= (s-z)(\tilde{q}_{n-1} s^{n-1} + \cdots + \tilde{q}_1 s + \tilde{q}_0). \end{aligned}$$

This implies that

$$(\tilde{q}_{n-1} s^{n-1} + \cdots + \tilde{q}_1 s + \tilde{q}_0)P(s) - (\tilde{p}_{n-1} s^{n-1} + \cdots + \tilde{p}_1 s + \tilde{p}_0)Q(s) = 0.$$

We now balance the coefficients of powers of s in this expression:

$$\begin{aligned}
s^{2n-1}&: \ p_n\tilde{q}_{n-1} - q_n\tilde{p}_{n-1} = 0\\
s^{2n-2}&: \ p_{n-1}\tilde{q}_{n-1} + p_n\tilde{q}_{n-2} - q_{n-1}\tilde{p}_{n-1} - q_n\tilde{p}_{n-2} = 0\\
&\vdots\\
s^{1}&: \ p_0\tilde{q}_1 + p_1\tilde{q}_0 - q_0\tilde{p}_1 - q_1\tilde{p}_0 = 0\\
s^{0}&: \ p_0\tilde{q}_0 - q_0\tilde{q}_0 = 0.
\end{aligned}$$

One readily ascertains that this is exactly equivalent to

$$\boldsymbol{M}(P,Q)\begin{bmatrix} \tilde{q}_{n-1}\\ \vdots\\ \tilde{q}_1\\ \tilde{q}_0\\ -\tilde{p}_{n-1}\\ \vdots\\ -\tilde{p}_1\\ -\tilde{p}_0 \end{bmatrix} = \boldsymbol{0}.$$

This then implies that $\det \boldsymbol{M}(P,Q) = 0$ since not all of the coefficients $\tilde{q}_{n-1},\dots,\tilde{q}_0,\tilde{p}_{n-1},\dots,\tilde{p}_0$ can vanish.

Now suppose that $\det \boldsymbol{M}(P,Q) = 0$. This implies that there is a nonzero vector $\boldsymbol{x} \in \mathbb{R}^{2n}$ so that $\boldsymbol{M}(P,Q)\boldsymbol{x} = \boldsymbol{0}$. Let us write

$$\boldsymbol{x} = \begin{bmatrix} \tilde{q}_{n-1}\\ \vdots\\ \tilde{q}_1\\ \tilde{q}_0\\ -\tilde{p}_{n-1}\\ \vdots\\ -\tilde{p}_1\\ -\tilde{p}_0 \end{bmatrix}.$$

Now reversing the argument for the preceding part of the proof shows that

$$\begin{aligned}
&(\tilde{q}_{n-1}s^{n-1} + \cdots + \tilde{q}_1 s + \tilde{q}_0)P(s) = (\tilde{p}_{n-1}s^{n-1} + \cdots + \tilde{p}_1 s + \tilde{p}_0)Q(s)\\
\Longrightarrow\quad &\frac{Q(s)}{P(s)} = \frac{\tilde{q}_{n-1}s^{n-1} + \cdots + \tilde{q}_1 s + \tilde{q}_0}{\tilde{p}_{n-1}s^{n-1} + \cdots + \tilde{p}_1 s + \tilde{p}_0}.
\end{aligned}$$

Since either P or Q has degree n, it must be the case that P and Q have a common factor. ▼

Now we proceed with the proof.

(i) Since N_P and D_P are coprime, their eliminant $\boldsymbol{M}(D_P,N_P)$ is invertible by Lemma 1. Now let $P \in \mathbb{R}[s]$ be monic and of degree $2n-1$ and write

$$P(s) = s^{2n-1} + p_{2n-2}s^{2n-2} + \cdots + p_1 s + p_0.$$

Now define a vector in $\mathbb{R}^{2n}$ by

$$\begin{bmatrix} b_{n-1} \\ \vdots \\ b_1 \\ b_0 \\ a_{n-1} \\ \vdots \\ a_1 \\ a_0 \end{bmatrix} = \boldsymbol{M}(D_P, N_P)^{-1} \begin{bmatrix} p_{2n-1} \\ \vdots \\ p_1 \\ p_0 \end{bmatrix}. \tag{13.2}$$

One verifies by direct computation that the resulting equation

$$\boldsymbol{M}(D_P, N_P) \begin{bmatrix} b_{n-1} \\ \vdots \\ b_1 \\ b_0 \\ a_{n-1} \\ \vdots \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} p_{2n-1} \\ \vdots \\ p_1 \\ p_0 \end{bmatrix}$$

is exactly the result of equating $D_C D_P + N_C N_P = P$, provided that we define

$$R_C(s) = \frac{a_{n-1}s^{n-1} + \cdots + a_1 s + a_0}{s^{n-1} + b_{n-2}s^{n-2} + \cdots + b_1 s + b_0}.$$

To complete the proof, we must also show that the numerator and denominator in this expression for R_C are coprime.

finish

(ii) In this case we take

$$P(s) = s^{2n} + p_{2n-1}s^{2n-1} + \cdots + p_1 s + p_0,$$

and define a vector in $\mathbb{R}^{2n+1}$ by

$$\begin{bmatrix} b_n \\ b_{n-1} \\ \vdots \\ b_1 \\ b_0 \\ a_{n-1} \\ \vdots \\ a_1 \\ a_0 \end{bmatrix} = \bar{\boldsymbol{M}}(D_P, N_P)^{-1} \begin{bmatrix} p_{2n} \\ p_{2n-1} \\ \vdots \\ p_1 \\ p_0 \end{bmatrix}. \tag{13.3}$$

Here the matrix $\bar{\boldsymbol{M}}(D_P, N_P)$ is defined by

$$\bar{\boldsymbol{M}}(D_P, N_P) = \left[\begin{array}{c|c} p_{2n} & \mathbf{0}^t \\ \hline \boldsymbol{m}(D_P) & \boldsymbol{M}(D_P, N_P) \end{array}\right], \tag{13.4}$$

and where $\boldsymbol{m}(D_P) \in \mathbb{R}^{2n}$ is a vector containing the coefficients of D_P, with the coefficient of s^{n-1} in the first entry, and with the last n entries being zero. Clearly $\bar{\boldsymbol{M}}(D_P, N_P)$ is invertible since $\boldsymbol{M}(D_P, N_P)$ is invertible. Now one checks that if

$$R_C(s) = \frac{a_{n-1}s^{n-1} + \cdots + a_1 s + a_0}{s^n + b_{n-1}s^{n-1} + \cdots + b_1 s + b_0},$$

then this controller satisfies the conclusions of this part of the proposition. ■

13.3 Remarks

1. This result is analogous to Theorem 10.27 in that it provides an *explicit* formula, in this case either (13.2) or (13.3), for a stabilising controller for the feedback loop of Figure 13.8. In fact, in each case we can achieve a prescribed characteristic polynomial of a certain type.
2. In Theorem 10.27 the characteristic polynomial had to be of degree $2n$ and had to be writable as a product of two polynomials (this latter restriction is only a restriction when n is odd). However, in Theorem 13.2 we go this one better because the characteristic polynomial had degree one less, $2n-1$, at least in cases when R_P is strictly proper. This means we have a controller whose denominator has one degree less than that of Theorem 10.27. This can be advantageous.
3. One of the things we loose in Theorem 13.2 is the separation principle interpretation available for Theorem 10.27 (cf. Theorem 10.48). •

An example illustrates how to explicitly apply Theorem 13.2.

13.4 Example Let us consider the unstable, nonminimum phase plant

$$R_P(s) = \frac{1-s}{s^2+1}.$$

We first wish to design a proper controller that produces the closed-loop characteristic polynomial

$$P(s) = s^3 + 3s^2 + 4s + 2.$$

We determine the eliminant $\boldsymbol{M}(D_P, N_P)$ to be

$$\boldsymbol{M}(D_P, N_P) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

An application of (13.2) gives

$$\boldsymbol{M}(D_P, N_P)^{-1} \begin{bmatrix} 1 \\ 3 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 1 \\ -2 \end{bmatrix}.$$

This then gives the controller

$$R_C(s) = \frac{s-2}{s+4}.$$

$$T(s) = \frac{R_C(s)R_P(s)}{1 + R_C(s)R_P(s)} = \frac{(s-1)(s-2)}{s^3 + 3s^2 + 4s + 2}.$$

Note that R_C is indeed a proper, but not a strictly proper, controller.

To achieve a proper controller we use part (ii) of Theorem 13.2. To do this, we must specify a closed-loop characteristic polynomial of degree 4. Let us go with

$$P(s) = s^4 + 4s^3 + 7s^2 + 6s + 2.$$

Next we must determine the matrix $\bar{\boldsymbol{M}}(D_P, N_P)$ in the equation (13.3). Using the description of this provided in the proof we get

$$\bar{\boldsymbol{M}}(D_P, N_P) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

An application of (13.3) gives

$$\bar{\boldsymbol{M}}(D_P, N_P) \begin{bmatrix} 1 \\ 4 \\ 7 \\ 6 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 5 \\ -1 \\ -3 \end{bmatrix}.$$

This then gives the strictly proper controller

$$R_C(s) = \frac{-s-3}{s^2 + 4s + 5}.$$

The corresponding closed-loop transfer function is

$$T(s) = \frac{R_C(s)R_P(s)}{1 + R_C(s)R_P(s)} = \frac{(s+3)(s-1)}{s^4 + 4s^3 + 7s^2 + 6s + 2}.$$

Note that the order of the two controllers we have designed is as predicted by Theorem 13.2, and are comparable to the order of the plant.

The closed-loop characteristic polynomials are designed to be stable. In Figure 13.9 we show the Nyquist plots for both systems. One can get some idea of the stability margins of the closed-loop system from these. •

Let us discuss this a little further by looking into a couple of related results. The first is that if we wish to increase the denominator degree of the controller, we may.

13.5 Corollary *Consider the interconnection of Figure 13.8 and let (N_P, D_P) be the c.f.r. for R_P, supposing $\deg(D_P) = n \geq \deg(N_P)$. If $P \in \mathbb{R}[s]$ is monic, of degree $k \geq 2n-1$, and if the coefficient of s^{2n-1} in P is nonzero, then there exists $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) so that the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P.*

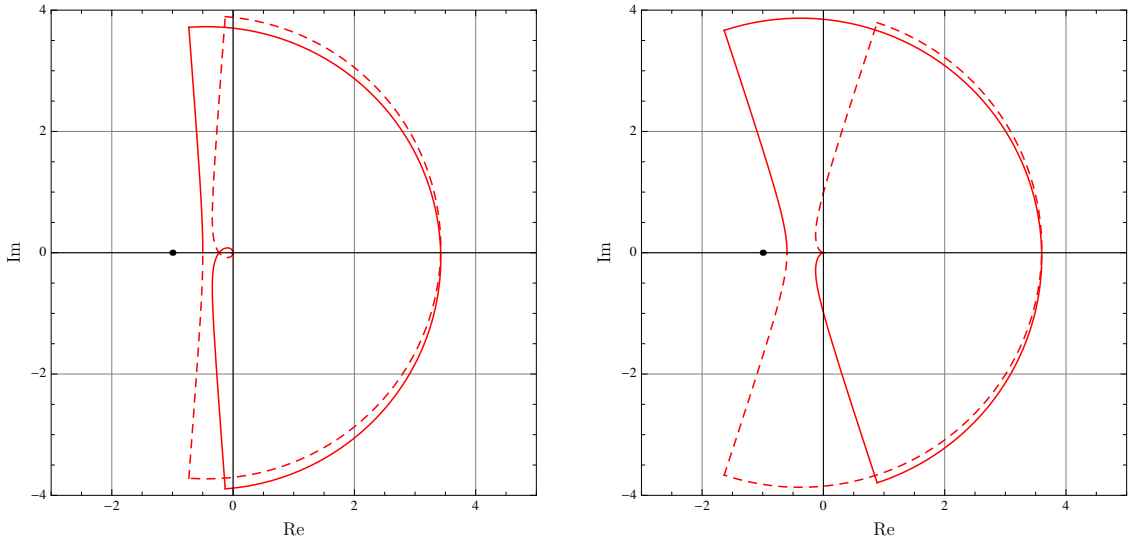


Figure 13.9 Nyquist plots for polynomial pole placement for $R_P(s) = \frac{1}{s^2+1}$ for a proper controller (left) and a strictly proper controller (right)

Proof Define

$$D_{C,1}(s) = s^k + b_{k-1}s^{k-1} + \cdots + b_n s^n$$

by asking that the polynomial $D_{C,1}D_P - P$ have degree $2n-1$. Thus $D_{C,1}$ is obtained by equating the coefficients of $s^k, \ldots, s^{2n}$ in the polynomials $D_{C,1}D_P$ and P. Now define $\tilde{P} = \frac{1}{p_{2n-1}}(P - D_{C,1}D_P)$. By construction of $D_{C,1}$, $\tilde{P}$ is monic and has degree $2n-1$. Thus, by Theorem 13.2, there exists

$$\tilde{R}_C(s) = \frac{\tilde{a}_{n-1}s^{n-1} + \cdots + \tilde{a}_1 s + \tilde{a}_0}{s^{n-1} + \tilde{b}_{n-2}s^{n-2} + \cdots + \tilde{b}_1 s + \tilde{b}_0}$$

with the property that $D_{C,2}D_P + N_{C,2}N_P = \tilde{P}$, if $(N_{C,2}, D_{C,2})$ is the c.f.r. for $\tilde{R}_C$. Taking

$$D_C = D_{C,1} + p_{2n-1}D_{C,2}, \quad N_C = p_{2n-1}N_{C,2}, \quad R_C = \frac{N_C}{D_C}$$

gives the corollary. ■

13.6 Remark The controller R_C in the corollary will not be unique as it will be, for example, in Theorem 13.2. •

Let us now see that the above result works in an example.

13.7 Example **(Example 13.4 cont'd)** Let us see what happens when we choose a closed-loop characteristic polynomial whose degree is "too high." Let us suppose that we wish to achieve the closed-loop characteristic polynomial

$$P(s) = s^5 + 5s^4 + 12s^3 + 16s^2 + 12s + 4.$$

We should look for a controller of order $5 - 2 = 3$. One may verify that the controller

$$R_C(s) = \frac{-5s^3 - 16s^2 - 8s - 20}{s^3 + 24}$$

achieves the desired characteristic polynomial. However, this is not the only controller that accomplishes this task. Indeed, one may verify that any controller of the form

$$\frac{-5s^3 - 16s^2 - 8s - 20 + a_1(s^3 + s^2 + s + 1) + a_2(s^2 + 1)}{s^3 + 24 + a_1(s^2 - 1) + a_2(s - 1)}$$

will achieve the same characteristic polynomial. Thus we see explicitly how the freedom of the high degree characteristic polynomial plays out in the controller. This may be helpful in practice to fine tune the controller do have certain properties. •

It is not surprising that we should be able to achieve a characteristic polynomial of degree higher than that of Theorem 13.2. It is also the case that we can generally expect to do not better. The following result gives this in precise terms.

13.8 Proposition *For the interconnection of Figure 13.8, let R_P be proper with c.f.r. (N_P, D_P) and $\deg(D_P) = n$. If $k < 2n - 1$ then there exists a monic polynomial P of degree k so that there is no proper controller R_C with the property that $\deg(D_C) = k - n$, where (N_C, D_C) is the c.f.r. for R_C.*

Proof The most general proper controller R_C for which $\deg(D_C) = k-n$ will have $2(n-k+1)$ undetermined coefficients. The equation $D_C D_P + N_C N_P = P$ gives a linear equation in these coefficients by balancing powers of s. This linear equation is one with $k + 1$ equations in $2(k - n + 1)$. However, we have

$$\begin{aligned} & k < 2n - 1 \\ \implies \quad & -k > -2n + 1 \\ \implies \quad & k + 1 > 2k - 2n + 2. \end{aligned}$$

This means that the linear equation for determining the coefficients of R_C has more equations than unknowns. Since a linear map from a vector space into one of larger dimension is incapable of being surjective, the proposition follows. ■

This means that we cannot expect to stabilise a general plant except with a controller whose denominator has comparable degree. This is easily illustrated with an example

13.9 Example **(Example 13.4 cont'd)** Let us continue with our example where

$$R_P(s) = \frac{1 - s}{s^2 + 1}.$$

We wish to show that we cannot find a proper controller R_C with the property that the closed-loop characteristic polynomial for the interconnection of Figure 13.8 is an arbitrary Hurwitz polynomial of degree 2. Clearly a proper controller of degree 1 or greater will lead to a closed-loop characteristic polynomial of degree 3 or greater. Thus we may only use a constant controller: $R_C(s) = K$. In this case the closed-loop characteristic polynomial is readily determined to be

$$P(s) = s^2 - Ks + 1 + K.$$

Note that this polynomial is Hurwitz only for $K \in (-1, 0)$. Thus the class of closed-loop characteristic polynomials of degree 2 that we may achieve is limited. Indeed, in Figure 13.10 we show the locus of roots for those values of K for which the closed-loop system is IBIBO stable. The placement of poles is clearly restricted. •

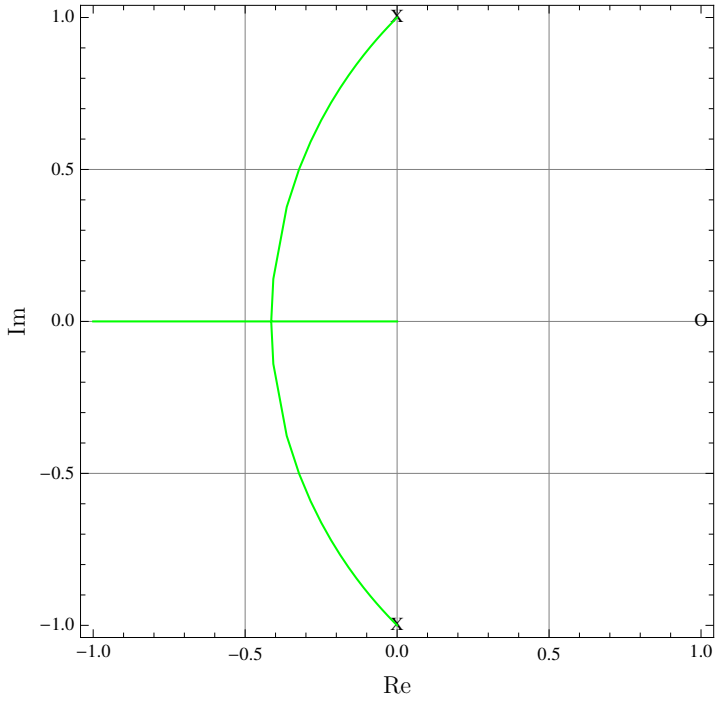


Figure 13.10 Locus of roots for $s^2 - Ks + 1 + K$ where $K \in (-1, 0)$

13.2.2 Enforcing design considerations

Sometimes one wants to design a controller with certain properties. One of the most often encountered of these is that R_C have a pole at $s = 0$ so as to ensure asymptotic tracking of step inputs and rejection of step disturbances. With this in mind, we have the following result.

13.10 Proposition *Consider the interconnection of Figure 13.8 and suppose that (N_P, D_P) if the c.f.r. for R_P. For $k \in \mathbb{N}$ the following statements hold:*

(i) if $\deg(N_P) \leq n - 1$ and if $P \in \mathbb{R}[s]$ is monic and degree $2n + k - 1$, then there exists a controller $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) with the following properties:

- *(a) R_C is proper;*
- *(b) R_C has a pole of order at least k at $s = 0$;*
- *(c) $\deg(D_C) = n + k - 1$;*
- *(d) the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P;*

(ii) if $\deg(N_P) \leq n$ and if $P \in \mathbb{R}[s]$ is monic and degree $2n + k$, then there exists a controller $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) with the following properties:

- *(a) R_C is strictly proper;*
- *(b) R_C has a pole of order at least k at $s = 0$;*
- *(c) $\deg(D_C) = n + k$;*
- *(d) the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P.*

Proof (i) Let us write

$$R_P(s) = \frac{c_{n-1}s^{n-1} + \cdots + c_1 s + c_0}{s^n + d_{n-1}s^{n-1} + \cdots + d_1 s + d_0}$$

and

$$P(s) = s^{2n+k-1} + p_{2n+k-2}s^{2n+k-2} + \cdots + p_1 s + p_0.$$

The idea is the same as in the proof of Theorem 13.2 in that the issue is matching coefficients. The content lies in ascertaining the form of the coefficient matrix. In this case we define a $(2n+k) \times k$ matrix $\boldsymbol{A}_k(N_P)$ by

$$\boldsymbol{A}_k(N_P) = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ c_{n-1} & 0 & \cdots & 0 & 0 \\ c_{n-2} & c_{n-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & c_0 & c_1 \\ 0 & 0 & \cdots & 0 & c_0 \end{bmatrix}.$$

The way to assemble $\boldsymbol{A}_k(N_P)$ in practice is to put the components of the polynomial N_C into the last column of $\boldsymbol{A}_k(N_P)$, starting with c_0 in the last row and working backwards. The next column to the left is made by shifting the last column up one row and placing a zero in the last row. One proceeds in this way until all k columns have been formed. Now define

$$\boldsymbol{M}_k(D_P, N_P) = \left[\begin{array}{c|c} \boldsymbol{M}(D_P, N_P) & \boldsymbol{A}_k(N_P) \\ \boldsymbol{0} & \end{array}\right] \in \mathbb{R}^{(2n+k)\times(2n+k)}.$$

One now defines a vector in $\mathbb{R}^{2n+k}$ by

$$\begin{bmatrix} b_{n+k-1} \\ \vdots \\ b_k \\ a_{n+k-1} \\ \vdots \\ a_0 \end{bmatrix} = \boldsymbol{M}_k(D_P, N_P)^{-1} \begin{bmatrix} p_{2n+k-1} \\ p_{2n+k-2} \\ \vdots \\ p_0 \end{bmatrix}.$$

One now checks that if

$$R_C(s) = \frac{a_{n+k-1}s^{n+k-1} + \cdots + a_1 s + a_0}{b_{n+k-1}s^{n+k-1} + \cdots + b_k s^k},$$

then R_C has the properties stated in the proposition.

(ii) Here we write

$$R_P(s) = \frac{c_n s^n + c_{n-1}s^{n-1} + \cdots + c_1 s + c_0}{s^n + d_{n-1}s^{n-1} + \cdots + d_1 s + d_0}$$

and

$$P(s) = s^{2n+k} + p_{2n+k-1}s^{2n+k-1} + \cdots + p_1 s + p_0.$$

The next step is to define $\bar{\boldsymbol{A}}_k(N_P) \in \mathbb{R}^{(2n+k+1)\times k}$ by

$$\bar{\boldsymbol{A}}_k(N_P) = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ c_n & 0 & \cdots & 0 & 0 \\ c_{n-1} & c_n & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & c_0 & c_1 \\ 0 & 0 & \cdots & 0 & c_0 \end{bmatrix}.$$

Then we let

$$\bar{\boldsymbol{M}}_k(D_P, N_P) = \left[\begin{array}{c|c} \bar{\boldsymbol{M}}(D_P, N_P) & \bar{\boldsymbol{A}}_k(N_P) \\ \boldsymbol{0} & \end{array}\right] \in \mathbb{R}^{(2n+k-2)\times(2n+k-2)},$$

where $\bar{\boldsymbol{M}}(D_P, N_P)$ is as defined in (13.4). Then we define a vector in $\mathbb{R}^{2n+k+1}$ by

$$\begin{bmatrix} b_{n+k} \\ \vdots \\ b_k \\ a_{n+k-1} \\ \vdots \\ a_0 \end{bmatrix} = \bar{\boldsymbol{M}}_k(D_P, N_P)^{-1} \begin{bmatrix} p_{2n+k} \\ p_{2n+k-1} \\ \vdots \\ p_0 \end{bmatrix}.$$

The rational function

$$R_C(s) = \frac{a_{n+k-1}s^{n+k-1} + \cdots + a_1 s + a_0}{b_{n+k}s^{n+k} + \cdots + b_k s^k}$$

has the desired properties. ■

Let us apply this in an example, as the proof, as was the proof of Theorem 13.2, is constructive.

13.11 Example **(Example 13.4 cont'd)** We take

$$R_P(s) = \frac{1-s}{s^2+1},$$

and design a controller with a pole at $s = 0$ that produces a desired characteristic polynomial. First let us achieve a pole of degree $k = 1$ at the origin. To derive a proper controller, P must have degree $2n + k - 1 = 4$. Let us take

$$P(s) = s^4 + 4s^3 + 7s^2 + 6s + 2.$$

The matrix $\boldsymbol{M}_1(D_P, N_P)$ is then constructed as in the proof of Proposition 13.10 as

$$\boldsymbol{M}_1(D_P, N_P) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

We then compute

$$\boldsymbol{M}_1(D_P, N_P)^{-1} \begin{bmatrix} 1 \\ 4 \\ 7 \\ 6 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 9 \\ 5 \\ -1 \\ 2 \end{bmatrix}.$$

Thus we take

$$R_C(s) = \frac{5s^2 - s + 2}{s^2 + 9s}.$$

The Nyquist plot for the loop gain $R_C R_P$ is shown in Figure 13.11.

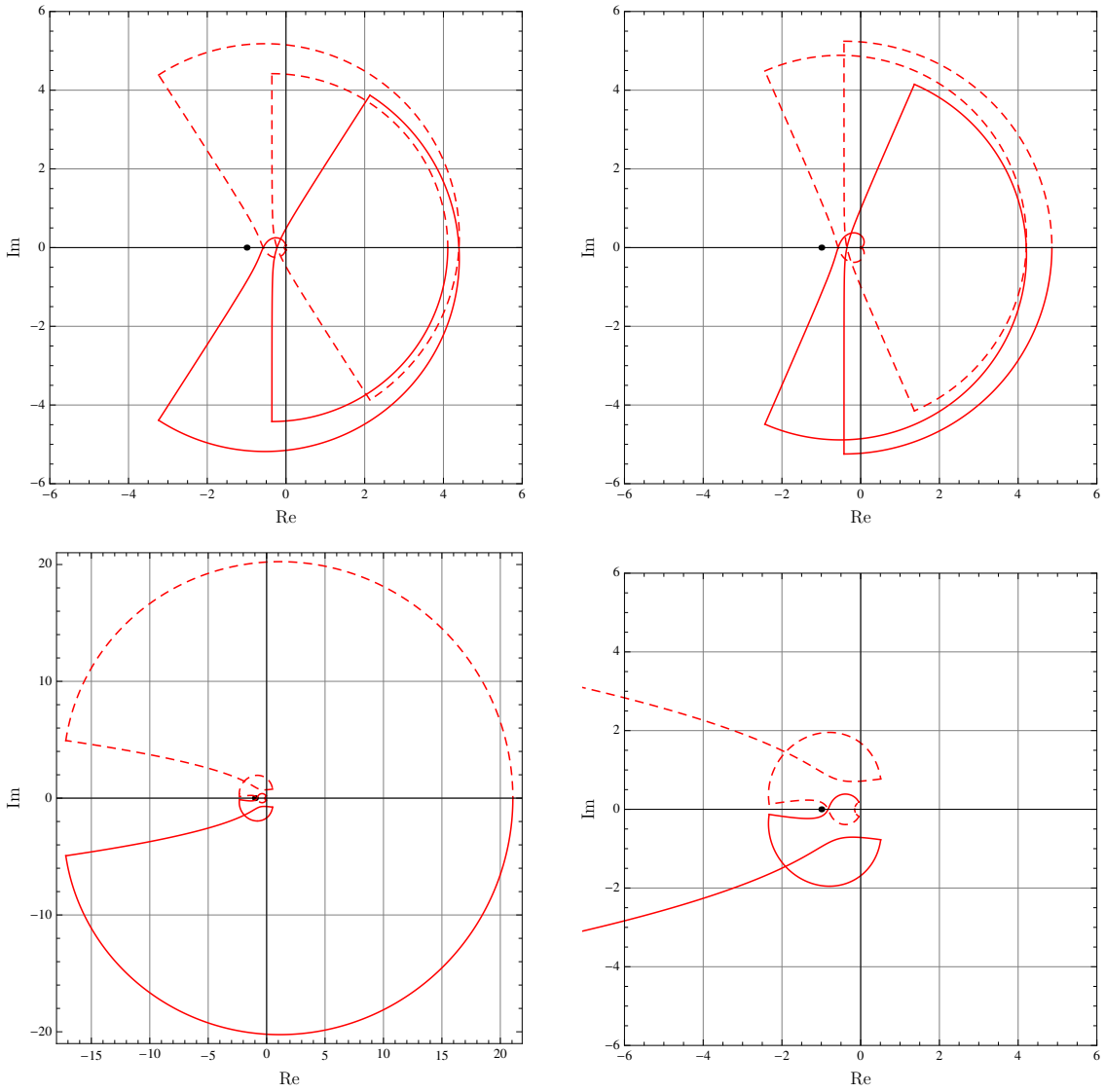


Figure 13.11 Nyquist plot for plant $R_P(s) = \frac{1-s}{s^2+1}$ and controllers $R_C(s) = \frac{5s^2-s+2}{s^2+9s}$ (top left), $R_C(s) = \frac{8s^2-3s+4}{s^3+5s^2+19s}$ (top right), $R_C(s) = \frac{19s^3+8s^2+16s+4}{s^3+24s^2}$ (bottom left and right)

If we demand that R_C be proper, then we use part (ii) of Proposition 13.10. We must

now specify a polynomial of degree $2n + k = 5$; let us take

$$P(s) = s^5 + 5s^4 + 12s^3 + 16s^2 + 12s + 4.$$

We compute

$$\bar{\boldsymbol{M}}_1(D_P, N_P) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

giving

$$\bar{\boldsymbol{M}}_1(D_P, N_P)^{-1} \begin{bmatrix} 1 \\ 5 \\ 12 \\ 16 \\ 12 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 19 \\ 8 \\ -3 \\ 4 \end{bmatrix}.$$

Thus we have

$$R_C(s) = \frac{8s^2 - 3s + 4}{s^3 + 5s^2 + 19s}.$$

One may see the Nyquist plot for the controller in Figure 13.11.

We may also design a controller giving a pole of degree $k = 2$ at $s = 0$. This requires the specification of a closed-loop characteristic polynomial of degree $2n + k - 1 = 5$; let us take

$$P(s) = s^5 + 5s^4 + 12s^3 + 16s^2 + 12s + 4.$$

again. We compute

$$\boldsymbol{M}_2(D_P, N_P) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$\boldsymbol{M}_2(D_P, N_P)^{-1} \begin{bmatrix} 1 \\ 5 \\ 12 \\ 16 \\ 12 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 24 \\ 19 \\ 8 \\ 16 \\ 4 \end{bmatrix}.$$

Thus we take

$$R_C(s) = \frac{19s^3 + 8s^2 + 16s + 4}{s^3 + 24s^2}.$$

The Nyquist plot for this plant/controller is shown in Figure 13.11. This is not strictly proper, and one may of course apply part (ii) of Proposition 13.10 to get a strictly proper controller. However, it is hopefully clear how to do this at this point. •

13.12 Remark It is interesting to look at the Nyquist plots for the controllers designed in Examples 13.4 and 13.11. First of all, the Nyquist plots get increasingly more complicated as we increase the controller order. One's ability to design controllers of this complexity with *ad hoc* methods is quite limited. On the other hand, one can see that the Nyquist plot for the third controller of Example 13.11 (the bottom Nyquist plots in Figure 13.11) has very bad gain and phase margins. One would be suspicious of such a controller, even though the theory tells us that it produces a desirable closed-loop characteristic polynomial. This points out the importance of using all the tools at one's disposal when designing a controller. •

One can also ask that more general polynomials appear in the denominator of R_C. Indeed, the proof of Proposition 13.10 can easily, if tediously, be adapted to prove the following result.

13.13 Corollary *Consider the interconnection of Figure 13.8 and suppose that (N_P, D_P) if the c.f.r. for R_P. For $F \in \mathbb{R}[s]$ a monic polynomial of degree k, the following statements hold:*

(i) if $\deg(N_P) \leq n-1$ and if $P \in \mathbb{R}[s]$ is monic and degree $2n+k-1$, then there exists a controller $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) with the following properties:

(a) R_C is proper;

(b) D_C has F as a factor;

(c) $\deg(D_C) = n+k-1$;

(d) the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P;

(ii) if $\deg(N_P) \leq n$ and if $P \in \mathbb{R}[s]$ is monic and degree $2n+k$, then there exists a controller $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) with the following properties:

(a) R_C is strictly proper;

(b) D_C has F as a factor;

(c) $\deg(D_C) = n+k$;

(d) the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P.

Just when one might wish to do this is a matter of circumstance. The following example gives one such instance.

13.14 Example Suppose we are given a plant R_P with c.f.r. (N_P, D_P) where $\deg(D_P) = n$. A reasonable design specification is that the closed-loop system be able to track well signals of a certain period ω. This would seem to demand that the transfer function from the input to the error should be zero at $s = i\omega$. For the interconnection of Figure 13.8 the transfer function from the input to the error is the sensitivity function for the loop,

$$S_L(s) = \frac{1}{1 + R_C(s)R_P(s)}.$$

By ensuring that R_C has a pole at $s = i\omega$, we can ensure that $S_L(i\omega) = 0$. Thus we should seek a controller of the form

$$R_C(s) = \frac{N_C(s)}{(s^2 + \omega^2)\tilde{D}_C(s)}.$$

Corollary 13.13 indicates that we can find a proper such controller provided that the closed-loop characteristic polynomial is specified to be of degree $2n + k - 1$. If the controller is

to be strictly proper, then we must allow the closed-loop characteristic polynomial to have degree $2n+k$. While we do not produce the explicit formula for the coefficients in N_C and $\tilde{D}_C$, one can easily produce such a formula by balancing polynomial coefficients, just as is done in Theorem 13.2 and Proposition 13.10. •

One may also wish to specify that the *numerator* of R_C have roots at some specified locations. The following result tells us when this can be done, and the degree of the closed-loop polynomial necessary to guarantee the required behaviour. We omit the details of the proof, as these go much like the proofs of Theorem 13.2 and Proposition 13.10, except that there are more complications.

13.15 Proposition *Consider the interconnection of Figure 13.8 and suppose that (N_P, D_P) if the c.f.r. for R_P. For $F \in \mathbb{R}[s]$ a monic polynomial of degree k, the following statements hold:*

(i) if $\deg(N_P) \leq n-1$ and if $P \in \mathbb{R}[s]$ is monic and degree $2n+k-1$, then there exists a controller $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) with the following properties:

(a) R_C is proper;

(b) N_C has F as a factor;

(c) $\deg(D_C) = n+k-1$;

(d) the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P;

(ii) if $\deg(N_P) \leq n$ and if $P \in \mathbb{R}[s]$ is monic and degree $2n+k$, then there exists a controller $R_C \in \mathbb{R}(s)$ with c.f.r. (N_C, D_C) with the following properties:

(a) R_C is strictly proper;

(b) N_C has F as a factor;

(c) $\deg(D_C) = n+k$;

(d) the closed-loop characteristic polynomial of the interconnection, $D_C D_P + N_P N_C$, is exactly P.

Note that if the GCD of N_C and D_P is F, then F is guaranteed to appear as a factor in the closed-loop characteristic polynomial.

13.2.3 Achievable poles using PID control

We begin by noting that in this section, as in the rest of this chapter, we use a PID controller that renders proper the derivative term in the controller. Thus we take a controller transfer function of the form

$$R_C(s) = K\Big(1 + \frac{T_D s}{\tau_D s + 1} + \frac{1}{T_I s}\Big). \tag{13.5}$$

The advantage of this from our point of view in this section is contained in the following result.

13.16 Lemma *Let $R \in \mathbb{R}[s]$ be any rational function of the form*

$$R(s) = \frac{a_2 s^2 + a_1 s + a_0}{s^2 + b_1 s}.$$

If one defines

$$K = \frac{a_1 b_1 - a_0}{b_1^2}, \quad T_D = \frac{a_0 - a_1 b_1 + a_2 b_1^2}{b_1(a_1 b_1 - a_0)}, \quad T_I = \frac{a_1 b_1 - a_0}{a_0 b_1}, \quad \tau_D = \frac{1}{b_1},$$

then $R = R_C$, where R_C is as in (13.5).

Proof We compute

$$R_C(s) = \frac{K(1+\frac{T_D}{\tau_D})s^2 + K(\frac{1}{\tau_D}+\frac{1}{T_I})s + \frac{K}{\tau_D T_I}}{s^2 + \frac{1}{\tau_D}s}$$

The lemma follows by setting

$$a_2 = K(1+\tfrac{T_D}{\tau_D}), \quad a_1 = K(\tfrac{1}{\tau_D}+\tfrac{1}{T_I}), \quad a_0 = \frac{K}{\tau_D T_I}, \quad b_1 = \frac{1}{\tau_D},$$

and solving for a_0, a_1, a_2, and b_1. ■

It is appropriate to employ a PID controller for first and second-order plants. To design PID controllers using the machinery of Sections 13.2.1 and 13.2.2, we use Proposition 13.10 with $k = 1$ to take the integrator into account. Let us see explicitly how to do this for general first and second-order plants.

naïve PID controllers with positive parameters

13.17 Proposition *Consider the three strictly proper plant transfer functions*

$$R_\tau(s) = \frac{1}{\tau s + 1}, \quad R_{\zeta,\omega_0,\tau}(s) = \frac{\omega_0^2(\tau s+1)}{s^2 + 2\zeta\omega_0 s + \omega_0^2}, \quad R_{\tau_1,\tau_2,\tau}(s) = \frac{\tau s + 1}{(\tau_1 s+1)(\tau_2 s+1)},$$

and the two polynomials

$$P_1(s) = s^3 + as^2 + bs + c, \quad P_2(s) = s^4 + as^3 + bs^2 + cs + d.$$

The following statements hold:

(i) if $R_P = R_\tau$ and if

$$R_C(s) = \frac{(a\tau - 1)s^2 + b\tau s + c\tau - \alpha s(\tau s+1)}{s^2 + \alpha s}$$

for $\alpha \in \mathbb{R}$, then the closed-loop polynomial of the interconnection of Figure 13.8 is P_1;

(ii) if $R_P = R_{\zeta,\omega_0,\tau}$ and if

$$\begin{aligned}
R_C(s) &= \frac{a_2 s^2 + a_1 s + a_0}{s^2 + b_1 s} \\
a_2 &= \frac{b - c\tau + d\tau^2 - \omega_0^2 + a\tau\omega_0^2 - 2a\omega_0\zeta - 2\tau\omega_0^3\zeta + 4\omega_0^2\zeta^2}{\omega_0^2(1 + \tau^2\omega_0^2 - 2\tau\omega_0\zeta)} \\
a_1 &= \frac{c}{\omega_0^2} - \frac{d\tau}{\omega_0^2} - \frac{a - b\tau + c\tau^2 - d\tau^3 + \tau\omega_0^2 - 2\omega_0\zeta}{1 + \tau^2\omega_0^2 - 2\tau\omega_0\zeta} \\
a_0 &= \frac{d}{\omega_0^2} \\
b_1 &= \frac{a - b\tau + c\tau^2 - d\tau^3 + \tau\omega_0^2 - 2\omega_0\zeta}{1 + \tau^2\omega_0^2 - 2\tau\omega_0\zeta}
\end{aligned}$$

then the closed-loop polynomial of the interconnection of Figure 13.8 is P_2;

(iii) if $R_P = R_{\tau_1,\tau_2,\tau}$ *and if*

$$
\begin{aligned}
R_C(s) &= \frac{(a_2/b_2)s^2 + (a_1/b_2)s + (a_0/b_2)}{s^2 + (b_1/b_2)s} \\
a_2 &= -\tau_2^2 - d\tau^2\tau_1^2\tau_2^2 + \tau_1\tau_2(-1 + a\tau_2) + \tau_1^2(-1 + a\tau_2 - b\tau_2^2) + \tau(\tau_1 + \tau_2 - a\tau_1\tau_2 + c\tau_1^2\tau_2^2) \\
a_1 &= -\tau_1 - \tau_2 + a\tau_1\tau_2 - c\tau_1^2\tau_2^2 - d\tau^2\tau_1\tau_2(\tau_1 + \tau_2) + \tau(1 - b\tau_1\tau_2 + d\tau_1^2\tau_2^2 + c\tau_1\tau_2(\tau_1 + \tau_2)) \\
a_0 &= -(d(\tau - \tau_1)\tau_1(\tau - \tau_2)\tau_2) \\
b_2 &= -(\tau - \tau_1)(\tau - \tau_2) \\
b_1 &= \tau_1 + \tau_2 - a\tau_1\tau_2 - c\tau^2\tau_1\tau_2 + d\tau^3\tau_1\tau_2 + \tau(-1 + b\tau_1\tau_2),
\end{aligned}
$$

then the closed-loop polynomial of the interconnection of Figure 13.8 is P_2.

13.18 Remarks

1. Using Lemma 13.16 one can turn the controllers of Proposition 13.17 into PID controllers of the form (13.5).
2. Note that the three plants R_τ, $R_{\zeta,\omega_0,\tau}$ and $R_{\tau_1,\tau_2,\tau}$ cover all possible strictly proper first and second-order plants.
3. The essential point is not so much the formulae themselves as their existence. That is to say, the main point is that for a first-order plant, a PID controller can be explicitly found that produces a desired third-order characteristic polynomial, and that for a second-order plant, a PID controller can be explicitly found that produces a desired fourth-order characteristic polynomial.
4. In practice, one would not use the formulae of Proposition 13.17, but would simply design a controller of the type $R_C(s) = \frac{a_2s^2+a_1s+a_0}{s(s+b_1)}$ by enforcing a pole at $s = 0$ as in Proposition 13.10.
5. Note that for first-order plants, there is some freedom in the design of an appropriate PID controller (characterised by the presence of the parameter α in part (i) of the proposition). However, the PID controller of part (ii) is uniquely specified by the desired characteristic polynomial. •

Let us illustrate a PID design using Proposition 13.17.

13.19 Example **(Example 12.14 cont'd)** So that we may contrast our design with that of the *ad hoc* PID design of Section 12.2.3, we take $RP = \frac{1}{s^2}$. Let us design a controller with poles at $\{-5, -5, -2 \pm 2i\}$. Thus we require the characteristic polynomial

$$P = s^4 + 14s^3 + 73s^2 + 180s + 200.$$

By contrast, the closed-loop characteristic polynomial of Example 12.14 is

$$s^3 + \tfrac{13}{4}s^2 + \tfrac{39}{10}s + \tfrac{13}{20},$$

which has roots of approximately $\{-0.197, -1.53 \pm 0.984i\}$. The lower degree of the characteristic polynomial is Example 12.14 is a consequence of the derivative term not being proper as in (13.5). In any event, an application of Proposition 13.17, or more conveniently of Proposition 13.10, gives

$$R_C(s) = \frac{73s^2 + 180s + 200}{s(s + 14)}.$$

Converting this to the PID form of (13.5) using Lemma 13.16 gives

$$K = \frac{580}{49}, \quad T_D = -\frac{2997}{8120}, \quad T_I = \frac{29}{35}, \quad \tau_D = \frac{1}{14}.$$

In Figure 13.12 we show the Nyquist plot for the loop gain $R_C R_P$. Note that this controller

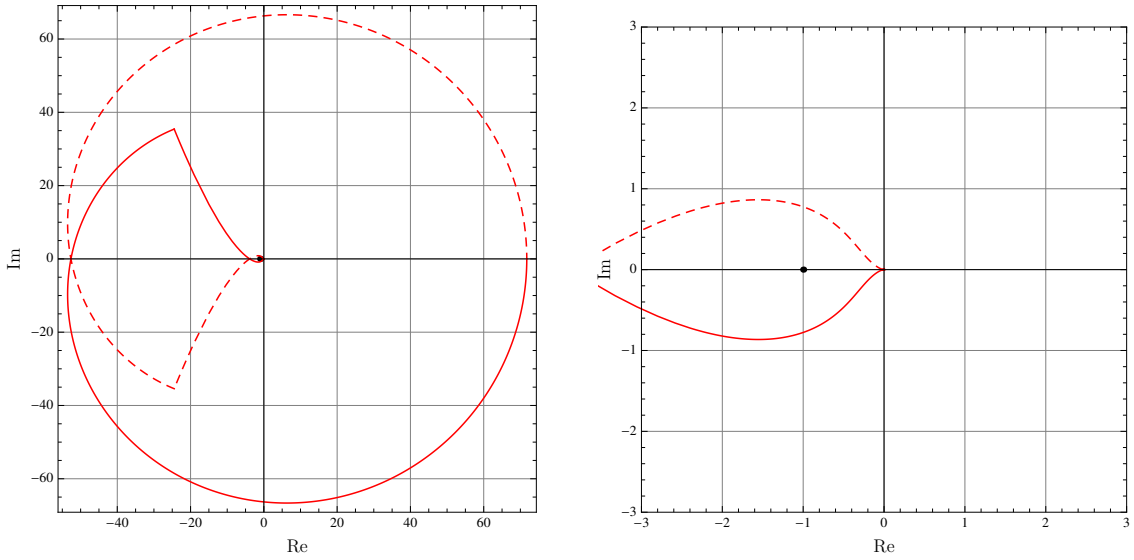


Figure 13.12 Nyquist plot for plant $R_P(s) = \frac{1}{s^2}$ and controller $R_C(s) = \frac{73s^2+180s+200}{s(s+14)}$

has off the bat presented us with respectable gain and phase margins. If one wished, this controller could be used as a starting point for further refinements to the stability margins. In Figure 13.13 is shown the step response and the response to a step disturbance between

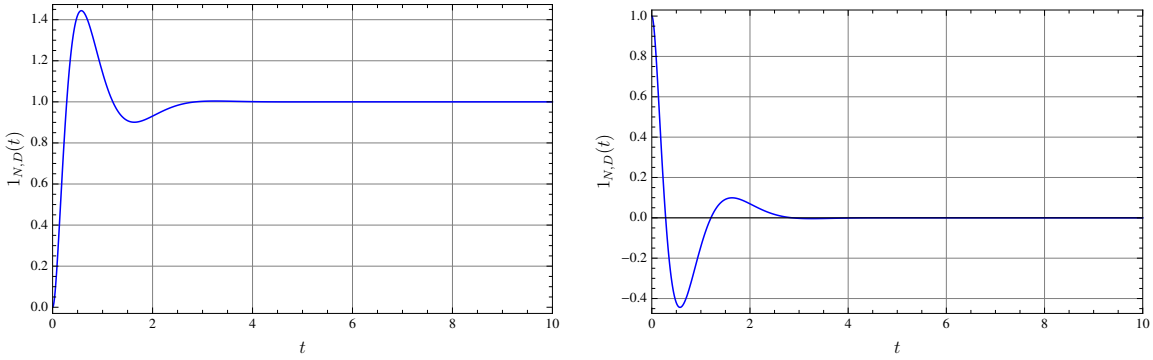


Figure 13.13 Step response (top) and response to step disturbance (bottom)

which enters the system at the output. The response time is quite good, although the overshoot is a bit large, and could be improved with a larger derivative time, perhaps. •

13.3 Two controller configurations

In the preceding section we considered the achieving of a specified closed-loop characteristic polynomial using a unity gain feedback configuration as in Figure 13.8. Next we turn our attention to a richer specification of the closed-loop transfer function to allow the determination of not only the closed-loop characteristic polynomial, but also additional features of the closed-loop transfer function. To do this we must consider a more complicated interconnection, and we consider the interconnection of Figure 13.14. There are now two

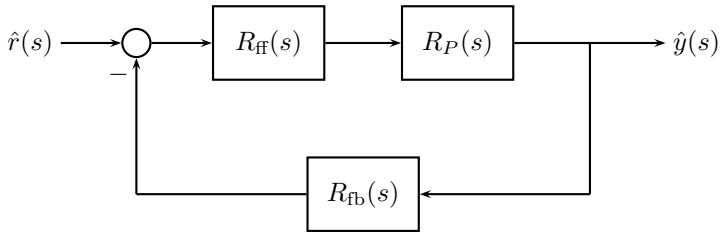


Figure 13.14 A two controller feedback loop

controllers we may specify, and we call R_{fb} the ***feedback controller*** and R_{ff} the ***feedforward controller***. It certainly makes sense that having two controllers makes it possible to do more than was possible in Section 13.2. The objective of this section is to quantify how much can be done with the richer configuration, and to detail exactly how to do what is possible.

13.3.1 Implementable transfer functions

First let us be clear about what we are after.

13.20 Definition Let $R_P \in \mathbb{R}(s)$ be a proper plant. $R \in \mathbb{R}(s)$ is ***implementable*** for R_P if there exists proper feedback and feedforward controllers $R_{\text{fb}}, R_{\text{ff}} \in \mathbb{R}(s)$ so that

(i) R is the closed-loop transfer function for the interconnection of Figure 13.14,

(ii) the interconnection is IBIBO stable, and

(iii) the only zeros of R in $\overline{\mathbb{C}}_+$ are those of R_P, including multiplicities.

We denote by $\mathscr{I}(R_P)$ the collection of implementable closed-loop transfer functions for the plant R_P. •

The next result tells us that it is possible to achieve an implementable transfer function with an interconnection of the form Figure 13.14.

13.21 Theorem *Let R_P be a proper plant with c.f.r. (N_P, D_P), and let $R \in \mathbb{R}(s)$ have c.f.r. (N, D). The following conditions are equivalent.*

(i) R is implementable;

(ii) $R, \frac{R}{R_P} \in \mathrm{RH}^+_\infty$ and the strictly nonminimum phase zeros of R and R_P agree, including multiplicities;

(iii) the following three conditions hold:

(a) D is Hurwitz;

(b) $\deg(D) - \deg(N) \geq \deg(D_P) - \deg(N_P)$;

(c) the roots of N in $\overline{\mathbb{C}}_+$ are exactly the roots of N in $\overline{\mathbb{C}}_+$, including multiplicities.

Proof (i) $\Longrightarrow$ (ii) If R is implementable, the interconnection of Figure 13.14 is IBIBO stable. In particular, the closed-loop transfer function is IBIBO stable, and since this is by hypothesis R, we have $R \in \mathrm{RH}_\infty^+$. The transfer function from the plant input $\hat{u}$ to the reference $\hat{r}$ should also be BIBO stable. This means that

$$\frac{\hat{u}}{\hat{r}} = \frac{\hat{y}/\hat{r}}{\hat{y}/\hat{u}} = \frac{R}{R_P} \in \mathrm{RH}_\infty^+.$$

Thus (ii) holds.

(ii) $\Longrightarrow$ (iii) As $R = \frac{N}{D}$ obviously D is Hurwitz so (iii a) holds. Since $\frac{R}{R_P} \in \mathrm{RH}_\infty^+$, $\frac{R}{R_P}$ must be proper. Since

$$\frac{R}{R_P} = \frac{ND_P}{DN_P}, \tag{13.6}$$

this implies that $\deg(DN_P) \geq \deg(ND_P)$. Since the degree of the product of polynomials is the sum of the degrees, it follows that

$$\deg(D) + \deg(N_P) \geq \deg(N) + \deg(D_P),$$

from which (iii b) follows. If $R, \frac{R}{R_P} \in \mathrm{RH}^+ +_\infty$ it follows from (13.6) that all roots of N_P in $\overline{\mathbb{C}}_+$ must also be roots of N, and vice versa. Thus (iii c) holds.

(iii) $\Longrightarrow$ (i) Suppose that R has c.f.r. (N, D) satisfying the conditions of (iii). Let F be the GCD of N and N_P and write $N = F\tilde{N}$ and $N_P = F\tilde{N}_P$. Then define $P_1 = D\tilde{N}_P$. Note that by (iii a) and (iii c), P_1 is Hurwitz. Now let P_2 be an arbitrary Hurwitz polynomial having the property that $\deg(P_1P_2) = 2n - 1$, where $n = \deg(D_P)$. By Corollary 13.5 we may find polynomials D_{ff} and N_{fb} so that

$$D_{\mathrm{ff}}D_P + N_{\mathrm{fb}}N_P = P_1P_2. \tag{13.7}$$

We also take $N_{\mathrm{ff}} = D_{\mathrm{fb}} = \tilde{N}P_2$. We now claim that if we take $R_{\mathrm{ff}} = \frac{N_{\mathrm{ff}}}{D_{\mathrm{ff}}}$ and $R_{\mathrm{fb}} = \frac{N_{\mathrm{fb}}}{D_{\mathrm{fb}}}$, then the interconnection of Figure 13.14 is IBIBO stable. The relevant transfer functions that must be checked as belonging to RH_∞^+ are

$$T_1 = \frac{R_PR_{\mathrm{ff}}R_{\mathrm{fb}}}{1 + R_PR_{\mathrm{ff}}R_{\mathrm{fb}}}, \quad T_2 = \frac{R_PR_{\mathrm{ff}}}{1 + R_PR_{\mathrm{ff}}R_{\mathrm{fb}}}$$
$$T_3 = \frac{R_PR_{\mathrm{fb}}}{1 + R_PR_{\mathrm{ff}}R_{\mathrm{fb}}}, \quad T_4 = \frac{R_{\mathrm{ff}}R_{\mathrm{fb}}}{1 + R_PR_{\mathrm{ff}}R_{\mathrm{fb}}}$$
$$T_5 = \frac{R_P}{1 + R_PR_{\mathrm{ff}}R_{\mathrm{fb}}}, \quad T_6 = \frac{R_{\mathrm{ff}}}{1 + R_PR_{\mathrm{ff}}R_{\mathrm{fb}}}$$
$$T_7 = \frac{R_{\mathrm{fb}}}{1 + R_PR_{\mathrm{ff}}R_{\mathrm{fb}}}.$$

NEED T_3 and T_7

To see that all of the transfer functions are have no poles in $\overline{\mathbb{C}}_+$ first note that they can all be written as rational functions with denominator

$$D_{\mathrm{ff}}D_{\mathrm{fb}}D_P + N_{\mathrm{ff}}N_{\mathrm{fb}}N_P = \tilde{N}P_2(D_{\mathrm{ff}}D_P + N_{\mathrm{fb}}N_P) = \tilde{N}P_1P_2^2. \tag{13.8}$$

P_2 is Hurwitz by design, P_1 is Hurwitz by (iii a) and (iii c), and $\tilde{N}$ is Hurwitz by (iii c). Thus all transfer functions T_1 and T_7 are analytic in $\overline{\mathbb{C}}_+$.

Let us now give the numerator for each of the transfer functions T_1 through T_7 when put over the denominator (13.8), and use this to ascertain that these transfer functions are all proper. In doing this, it will be helpful to have a bound on $\deg(\tilde{N}P_2)$, which we now obtain. We suppose that $\deg(P_1) = k$ so that $\deg(P_2) = 2n - k - 1$. Now,

$$\begin{aligned}\deg(P_1) &= \deg(D) + \deg(\tilde{N}_P)\\ &= \deg(D) + \deg(N_P) - \deg(F)\\ &\geq \deg(D_P) + \deg(N) - \deg(F)\\ &= n + \deg(\tilde{N}),\end{aligned}$$

using (iii b). Therefore

$$\begin{aligned}&\deg(P_2) = 2n - 1 - \deg(P_1) \leq 2n - 1 - n - \deg(\tilde{N})\\ \Longrightarrow\quad &\deg(\tilde{N}P_2) \leq n - 1. \qquad (13.9)\end{aligned}$$

Now we proceed with our calculations.

1. T_1: The numerator is $N_P N_{\text{ff}} N_{\text{fb}}$. Since $N_{\text{ff}} = \tilde{N}P_2$ this gives

$$T_1 = \frac{N_P N_{\text{fb}}}{P_1 P_2}.$$

   We have $\deg(P_1P_2) = 2n - 1$ and $\deg(N_P) \leq n$. Since N_{fb} is obtained from (13.7), $\deg(T_{\text{fb}}) \leq n - 1$. From this we deduce that T_1 is proper.

2. T_2: The numerator of T_2 is $N_P N_{\text{ff}} D_{\text{fb}}$. Since $N_{\text{ff}} = \tilde{N}P_2$ this gives

$$T_2 = \frac{N_P D_{\text{fb}}}{P_1 P_2}. \qquad (13.10)$$

   Again, $\deg(P_1P_2) = 2n - 1$ and $\deg(N_P) \leq n$. In (13.9) we showed that $\deg(D_{\text{fb}}) \leq n - 1$, thus T_2 is proper.

3. T_3: The numerator of T_3 is $N_P N_{\text{fb}} D_{\text{ff}}$, giving

$$T_3 = \frac{N_P N_{\text{fb}} D_{\text{ff}}}{\tilde{N}P_1P_2^2}$$

   We have $\deg(N_P) \leq n$ and $\deg(P_1P_2) = 2n - 1$. Since D_{ff} and N_{fb} satisfy (13.7), $\deg(D_{\text{ff}}), \deg(N_{\text{fb}}) \leq n - 1$. finish

4. T_4: The numerator is $D_P N_{\text{ff}} N_{\text{fb}}$. Since $N_{\text{ff}} = \tilde{N}P_2$ this gives

$$T_4 = \frac{D_P N_{\text{fb}}}{P_1 P_2}.$$

   We have $\deg(D_P) = n$ and $\deg(P_1P_2) = 2n - 1$. Since N_{fb} satisfies (13.7), $\deg(N_{\text{fb}}) \leq n - 1$. This shows that T_4 is proper.

5. T_5: The numerator is $N_P D_{\text{ff}} D_{\text{fb}}$. Since $D_{\text{fb}} = \tilde{N}P_2$ this gives

$$T_5 = \frac{N_P D_{\text{ff}}}{P_1 P_2}.$$

   We have $\deg(D_P) = n$ and $\deg(P_1P_2) = 2n - 1$. Since D_{ff} satisfies (13.7), $\deg(D_{\text{ff}}) \leq n - 1$. This shows that T_5 is proper.

6. T_6: The numerator is $D_P D_{\text{fb}} N_{\text{ff}}$. Since $N_{\text{ff}} = \tilde{N} P_2$ this gives
$$T_6 = \frac{D_P D_{\text{fb}}}{P_1 P_2}.$$
We have $\deg(D_P) = n$ and $\deg(P_1 P_2) = 2n - 1$. By (13.9) $\deg(D_{\text{fb}}) \leq n - 1$, showing that T_6 is proper.
7. T_7: The numerator is $D_P D_{\text{ff}} N_{\text{fb}}$, giving
$$T_7 = \frac{D_P D_{\text{ff}} N_{\text{fb}}}{\tilde{N} P_1 P_2^2}$$
We have $\deg(D_P) = n$ and $\deg(P_1 P_2) = 2n - 1$. Since D_{ff} and N_{fb} satisfy (13.7) we have $\deg(D_{\text{ff}}), \deg(N_{\text{fb}}) \leq n - 1$.

Finally we check that the closed-loop transfer function is R. Indeed, the closed-loop transfer function is T_2 and we have
$$R = \frac{N}{D} = \frac{\tilde{N} N_P}{D \tilde{N}_P} = \frac{\tilde{N} N_P P_2}{P_1 P_2} = \frac{N_P D_{\text{fb}}}{P_1 P_2} = T_2,$$
using (13.10).

Thus we have shown that the interconnection of Figure 13.14 is IBIBO stable with transfer function R. This means that R is implementable. ■

13.22 Remarks

1. The conditions $R, \frac{R}{R_P} \in \text{RH}_\infty^+$ are actually the weakest possible for a class of interconnections more general than that of Figure 13.14. Indeed, if $(\mathcal{S}, \mathcal{G})$ is any interconnected SISO linear system with the property that every forward path from the reference $\hat{r}$ to the output $\hat{y}$ passes through the plant, then one readily sees that the transfer function from the reference $\hat{r}$ to the input $\hat{u}$ to the plant is exactly $\frac{R}{R_P}$ (cf. the proof that (i) implies (ii) in Theorem 13.21 below). Thus we have the following statement:

   > *If $(\mathcal{S}, \mathcal{G})$ is an interconnected SISO linear system with the property that every forward path from the input to the output passes through the plant, then the interconnection is IBIBO stable with transfer function $R \in \mathbb{R}(s)$ only if $R, \frac{R}{R_P} \in \text{RH}_\infty^+$.*

   (This is Exercise E6.14.) This indicates that the conditions $R, \frac{R}{R_P} \in \text{RH}_\infty^+$ are the weakest one can impose on a closed-loop transfer function if it is to be realisable by some "reasonable" interconnection. The additional hypothesis that R have no nonminimum phase zeros other than those of R_P is reasonable: the nonminimum phase zeros of R_P *must* appear in R, and we would not want any more of these than necessary, given the discussions of Chapters 8 and 9 concerning the effects of nonminimum phase zeros.
2. As with our results of Section 13.2, Theorem 13.21 is constructive.
3. Note that the interconnection of Figure 13.14 is the same as that of Figure 10.9. Indeed, there is a relationship between the controllers constructed in Theorem 13.21 and the combination of the observer combined with static state feedback in Theorem 10.48. The procedure of Theorem 13.21 is a bit more flexible in that the order of the closed-loop characteristic polynomial is not necessarily $2n$.

Let us first illustrate via an example that implementability is indeed different that stabilisation of the closed-loop system.

13.23 Example Let $R_P(s) = \frac{s-2}{s^2-1}$. One can verify that $R(s) = \frac{2-s}{s^2+2s+2}$ is implementable. Indeed, clearly $R \in \mathrm{RH}_\infty^+$ and we also have

$$\frac{R(s)}{R_P(s)} = \frac{1-s^2}{s^2+2s+2} \in \mathrm{RH}_\infty^+.$$

Since we have

$$R(s) = \frac{R_C(s)R_P(s)}{1+R_C(s)R_P(s)},$$

we can solve this for R_C to get

$$R_C(s) = \frac{R}{(R-1)RP} = \frac{1-s^2}{s(s+3)}.$$

Note that although R is implementable, the interconnection of Figure 13.8 is *not* IBIBO stable. Thus we see that implementability *is* different from an IBIBO interconnection of the form Figure 13.8. •

13.3.2 Implementations that meet design considerations

As we saw in Section 13.2.2, one often wants to enforce more than IBIBO stability on one's feedback loop. Let us see how these considerations can be enforced in the two controller configuration of Figure 13.14.

13.4 Synthesis using controller parameterisation

In Section 10.3 we described the set of proper controllers that stabilise a proper plant R_P, and the parameterisation came to us in terms of a free function in RH_∞^+. In this section we turn to using this parameterisation to provide a useful guide to controller design. We will be concerned with the standard plant/controller unity feedback loop, and we reproduce this in Figure 13.15 for easy reference.

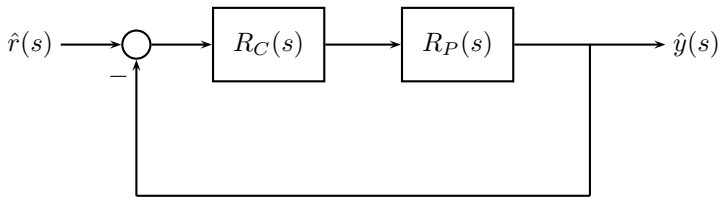


Figure 13.15 Interconnection for studying controller parameterisation

13.4.1 Properties of the Youla parameterisation

We first recall the Youla parameterisation of Theorem 10.37. Given a proper plant R_P, the set $\mathscr{S}_{\mathrm{pr}}(R_P)$ of proper controllers that render the interconnection of Figure 13.15 IBIBO stable is given by

$$\mathscr{S}_{\mathrm{pr}}(R_P) = \{\frac{\rho_1+\theta P_2}{\rho_2-\theta P_1} \mid \theta \text{ admissible}\}.$$

Recall that (P_1, P_2) is a coprime fractional representative of R_P and (ρ_1, ρ_2) is a coprime factorisation of (P_1, P_2). Also recall that the collection of admissible functions θ are defined by

1. $\theta \neq \dfrac{\rho_2}{P_1}$, and
2. $\lim\limits_{s\to\infty} \big(\rho_2(s) - \theta(s)P_1(s)\big) \neq 0$.

We wish to determine how the choice of the parameter θ affects the properties of the closed-loop system. For this purpose, let us suppose that we decide that we wish to place all poles of the closed-loop system in some region $\mathbb{C}_{\text{des}}$ in the complex plane. Note that we can always do this, since by, for example, Theorem 10.27 or Theorem 13.2, we may construct controllers that achieve any closed-loop characteristic polynomial, provided it has sufficiently high degree (twice the order of the plant for a strictly proper controller, and one less than this for a proper controller). Thus we are permitted to talk about the set of all proper controllers for which the closed-loop poles lie in $\mathbb{C}_{\text{des}}$, and let us denote this set of controllers by $\mathscr{S}_{\text{pr}}(R_P, \mathbb{C}_{\text{des}})$. The following result describes these controllers using the Youla parameterisation.

13.24 Proposition *Let R_P be a proper plant and $\mathbb{C}_{\text{des}} \subset \mathbb{C}$ be the set of desirable closed-loop pole locations. Also let (P_1, P_2) be a coprime fractional representative for R_P with (ρ_1, ρ_2) a coprime factorisation of (P_1, P_2). Then*

$$\mathscr{S}_{\text{pr}}(R_P, \mathbb{C}_{\text{des}}) = \{\frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1} \mid \theta \text{ admissible and } \theta \text{ has all poles in } \mathbb{C}_{\text{des}}\}.$$

Proof ■

The result is perhaps surprisingly "obvious," at least in statement, and clearly provides a useful tool for controller design.

13.5 Summary

1. Ziegler-Nichols tuning is available for doing PID control design.

Exercises

E13.1 In this exercise, you will show that the Ziegler-Nicols tuning method cannot be applied, even in some quite simple cases.

(a) Show that for any first-order plant, say $R_P = \frac{a}{s+b}$, both of the Ziegler-Nicols methods will not yield PID parameter values.

(b) Show that there exists a second-order plant, say of the form $R_P(s) = \frac{a}{s^2+bs+c}$, so that Assumption 13.1 will not be satisfied.

E13.2 In this exercise, you will apply the Ziegler-Nicols tuning method to design PID controllers for the plant transfer function

$$R_P(s) = \frac{1}{s^3 + s^2 + 2s + 1}.$$

First, the first method.

(a) Produce the step response numerically using Proposition 3.40.

(b) Graphically (or with the computer, if you can) determine the values of σ and τ, and so determine the parameters K, T_I, and T_D for the three cases of a P, PI, and PID controller.

(c) Using Proposition 3.40, determine the closed-loop step response.

(d) Comment on the performance of the three controllers.

Now the second method.

(e) Verify that Assumption 13.1 is satisfied for the plant transfer function R_P.
Hint: *A good way to do this might be to use a computer package to solve the equation obtained by setting the real part of the poles to zero.*

(f) Determine by trial and error (or with the computer, if you can) the value K_u of the proportional gain that makes two poles of the transfer function negative, and the period of the oscillatory part of the step response. Use these numbers to give the corresponding Ziegler-Nicols values for the parameters K, T_I, and T_D for the three cases of a P, PI, and PID controller.

(g) Using Proposition 3.40, determine the closed-loop step response.

(h) Comment on the performance of the three controllers.

E13.3 Exercises on controlling state examples using pole placement on output.

E13.4 Let R_P be a proper plant and let $R \in \mathscr{I}(R_P)$. Suppose that the feedback and feedforward controllers giving R as the closed-loop transfer function in Figure 13.14 are R_{fb} and R_{ff}, respectively. Let (N, D), $(N_{\text{fb}}, D_{\text{fb}})$, and $(N_{\text{ff}}, D_{\text{ff}})$ be the c.f.r.'s for R, R_{fb}, and R_{ff}, respectively.

(a) Show that all interconnections of Figure E13.1 have closed-loop transfer function R. Note that the recipe of Theorem 13.21 gives $N_{\text{ff}} = D_{\text{fb}}$, so let us assume that this is the case.

(b) Show that the interconnection (a) of Figure E13.1 will not generally be IBIBO stable.

(c) Show that the interconnection (b) of Figure E13.1 will not generally be IBIBO stable.

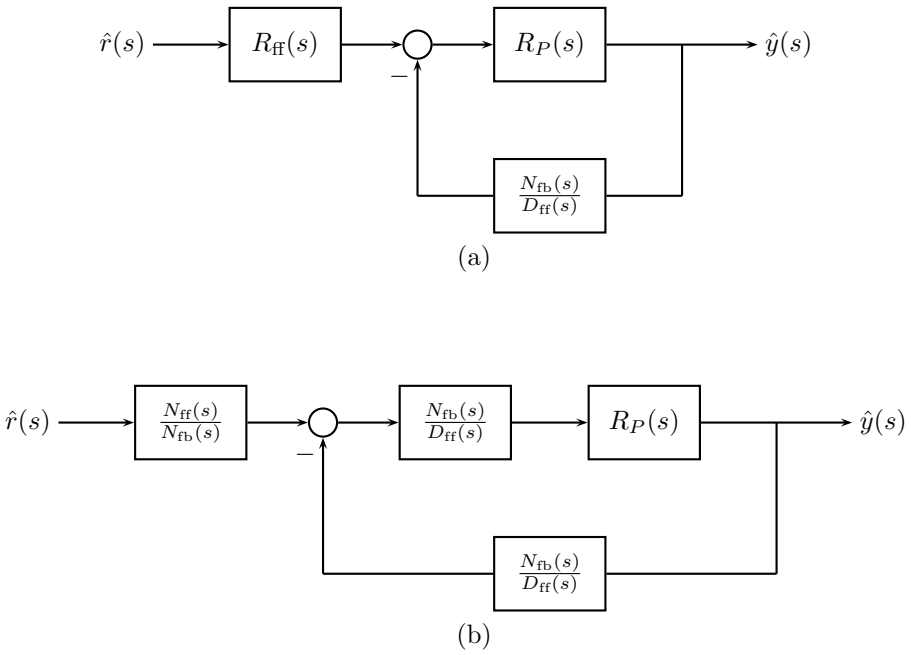


Figure E13.1 Alternate two controller configuration implementations

Chapter 14

An introduction to H_2 optimal control

The topic of optimal control is a classical one in control, and in some circles, mainly mathematical ones, "optimal control" is synonymous with "control." That we have covered as much material as we have without mention of optimal control is hopefully ample demonstration that optimal control is a subdiscipline, albeit an important one, of control. One of the useful features of optimal control is that is guides one is designing controllers *in a systematic way*. Much of the control design we have done so far has been somewhat *ad hoc*.

What we shall cover in this chapter is a very bare introduction to the subject of optimal control, or more correctly, linear quadratic regulator theory. This subject is dealt with thoroughly in a number of texts. Classical texts are [Brockett 1970] and [Bryson and Ho 1975]. A recent mathematical treatment is contained in the book [Sontag 1998]. Design issues are treated in [Goodwin, Graebe, and Salgado 2001]. Our approach closely follows that of Brockett. We also provide a treatment of optimal estimator design. Here a standard text is [Bryson and Ho 1975]. More recent treatments include [Goodwin, Graebe, and Salgado 2001] and [Davis 2002]. We have given this chapter a somewhat pretentious title involving "H_2." The reason for this is the frequency domain interpretation of our optimal control problem in Section 14.4.2.

Contents

14.1 Problems in optimal control and optimal state estimation

We begin by clearly stating the problems we consider in this chapter. The problems come in two flavours, optimal feedback and optimal estimation. While the problem formulations have a distinct flavour, we shall see that their solutions are intimately connected.

14.1.1 Optimal feedback

We shall start with a SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ which we assume to be in controller canonical form, and which we suppose to be observable. As was discussed in Section 2.5.1, following Theorem 2.37, this means that we are essentially studying the system

$$\begin{aligned} x^{(n)}(t) + p_{n-1}x^{(n-1)} + \dots p_1 x^{(1)}(t) + p_0 x(t) &= u(t) \\ y(t) = c_{n-1}x^{(n-1)}(t) + c_{n-2}x^{(n-2)}(t) + \cdots + c_1 x^{(1)}(t) + c_0 x(t), & \end{aligned} \tag{14.1}$$

where x is the first component of the state vector $\boldsymbol{x}$,

$$P_{\boldsymbol{A}}(\lambda) = \lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0$$

is the characteristic polynomial for $\boldsymbol{A}$, and $\boldsymbol{c} = (c_0, c_1, \dots, c_{n-1})$. Recall that $\mathscr{U}$ denotes the set of admissible inputs. For $u \in \mathscr{U}$ defined on an interval $I \in \mathscr{I}$ and for $\boldsymbol{x}_0 = (x_{00}, x_{01}, \dots, x_{0,n-1}) \in \mathbb{R}$, the ***solution*** for (14.1) corresponding to u and $\boldsymbol{x}_0$ is the unique map $x_{u,\boldsymbol{x}_0} \colon I \to \mathbb{R}$ that satisfies the first of equations (14.1) and for which

$$x(0) = x_{00}, \;\; x^{(1)}(0) = x_{01}, \;\; , \dots, x^{(n-1)}(0) = x_{0,n-1}.$$

The corresponding output, defined by the second of equations (14.1), we denote by $y_{u,\boldsymbol{x}_0} \colon I \to \mathbb{R}$. We fix $\boldsymbol{x}_0 \in \mathbb{R}^n$ and define $\mathscr{U}_{\boldsymbol{x}_0}$ to be the set of admissible inputs $u \in \mathscr{U}$ for which

1. $u(t)$ is defined on either
   (a) $[0, T]$ for some $T > 0$ or
   (b) $[0, \infty)$,

   and
2. if $y_{u,\boldsymbol{x}_0}$ is the output corresponding to u and $\boldsymbol{x}_0$, then
   (a) $y_{u,\boldsymbol{x}_0}(T) = 0$ if u is defined on $[0, T]$ or
   (b) $\lim_{t\to\infty} y_{u,\boldsymbol{x}_0}(t) = 0$ if u is defined on $[0, \infty)$.

Thus $\mathscr{U}_{\boldsymbol{x}_0}$ are those controls that, in possibly infinite time, drive the output from $y = y(0)$ to $y = 0$. For $u \in \mathscr{U}_{\boldsymbol{x}_0}$ define the cost function

$$J_{\boldsymbol{x}_0}(u) = \int_0^\infty \left(y^2_{u,\boldsymbol{x}_0}(t) + u^2(t)\right) \mathrm{d}t.$$

Thus we assign cost on the basis of equally penalising large inputs and large outputs, but we do not penalise the state, except as it is related to the output.

With this in hand we can state the following optimal control problem.

14.1 Optimal control problem in terms of output Seek a control $u \in \mathscr{U}_{\boldsymbol{x}_0}$ that has the property that $J_{\boldsymbol{x}_0}(u) \leq J_{\boldsymbol{x}_0}(\tilde{u})$ for any $\tilde{u} \in \mathscr{U}_{\boldsymbol{x}_0}$. Such a control u is called an ***optimal control law***. •

14.2 Remark Note that the output going to zero does not imply that the state also goes to zero, unless we impose additional assumptions of $(\boldsymbol{A}, \boldsymbol{c})$. For example, if $(\boldsymbol{A}, \boldsymbol{c})$ is detectable, then by driving the output to zero as $t \to \infty$, we can also ensure that the state is driven to zero as $t \to \infty$. •

Now let us formulate another optimal control problem that is clearly related to Problem 14.1, but is not obviously the same. In Section 14.3 we shall see that the two problems are, in actuality, the same. To state this problem, we proceed as above, except that we seek our control law in the form of state feedback. Thus we now work with the state equations

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t),$$

and we no longer make the assumption that $(\boldsymbol{A}, \boldsymbol{b})$ is in controller canonical form, but we do ask that Σ be complete. We recall that $\mathscr{S}_{\mathrm{s}}(\Sigma) \subset \mathbb{R}^n$ is the subset of vectors $\boldsymbol{f}$ for which $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ is Hurwitz. If $\boldsymbol{x}_0 \in \mathbb{R}^n$, we let $\boldsymbol{x}_{\boldsymbol{f},\boldsymbol{x}_0}$ be the solution to the initial value problem

$$\dot{\boldsymbol{x}}(t) = (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)\boldsymbol{x}(t), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

Thus $\boldsymbol{x}_{\boldsymbol{f},\boldsymbol{x}_0}$ is the solution for the state equations under state feedback via $\boldsymbol{f}$ with initial condition $\boldsymbol{x}_0$. We define the cost function

$$J_{\boldsymbol{c},\boldsymbol{x}_0}(\boldsymbol{f}) = \int_0^\infty \left(\boldsymbol{x}_{\boldsymbol{f},\boldsymbol{x}_0}^t(t)(\boldsymbol{c}\boldsymbol{c}^t + \boldsymbol{f}\boldsymbol{f}^t)\boldsymbol{x}_{\boldsymbol{f},\boldsymbol{x}_0}(t)\right) \mathrm{d}t.$$

Note that this is the same as the cost function $J_{\boldsymbol{x}_0}(u)$ if we take $u(t) = -\boldsymbol{f}^t\boldsymbol{x}_{\boldsymbol{f},\boldsymbol{x}_0}(t)$ and $y_{u,\boldsymbol{x}_0}(t) = \boldsymbol{c}^t\boldsymbol{x}_{\boldsymbol{f},\boldsymbol{x}_0}(t)$. Thus there is clearly a strong relationship between the cost functions for the two optimal control problems we are considering.

In any case, for $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{x}_0 \in \mathbb{R}^n$, we then define the following optimal control problem.

14.3 Optimal control problem in terms of state Seek $\boldsymbol{f} \in \mathscr{S}_{\mathrm{s}}(\Sigma)$ so that $J_{\boldsymbol{c},\boldsymbol{x}_0}(\boldsymbol{f}) \leq J_{\boldsymbol{c},\boldsymbol{x}_0}(\tilde{\boldsymbol{f}})$ for every $\tilde{\boldsymbol{f}} \in \mathscr{S}_{\mathrm{s}}(\Sigma)$. Such a feedback vector $\boldsymbol{f}$ is called an ***optimal state feedback vector***. •

14.4 Remark Our formulation of the two optimal control problems is adapted to our SISO setting. More commonly, and more naturally, the problem is cast in a MIMO setting, and let us do this for the sake of context. We let $\Sigma = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{0})$ be a MIMO linear system satisfying the equations

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t)\\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{x}(t).\end{aligned}$$

The natural formulation of a cost for the system is given directly in terms of states and outputs, rather than inputs and outputs. Thus we take as cost

$$J = \int_0^\infty \left(\boldsymbol{x}^t(t)\boldsymbol{Q}\boldsymbol{x}(t) + \boldsymbol{u}^t(t)\boldsymbol{R}\boldsymbol{u}(t)\right) \mathrm{d}t,$$

where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{R} \in \mathbb{R}^{m\times m}$ are symmetric, and typically positive-definite. The positive-definiteness of $\boldsymbol{R}$ is required for the problem to be nice, but that $\boldsymbol{Q}$ can be taken to be only positive-semidefinite. The solution to this problem will be discussed, but not proved, in Section 14.3.2. •

14.1.2 Optimal state estimation

Next we consider a problem that is "dual" to that considered in the preceding section. This is the problem of constructing a Luenberger observer that is optimal in some sense. The problem we state in this section is what is known as the *deterministic* version of the problem. There is a stochastic version of the problem statement that we do not consider. However, the stochastic version is very common, and is indeed where the problem originated in the work of Kalman [1960] and Kalman and Bucy [1960]. The stochastic problem can be found, for example, in the recent text of Davis [2002]. Our deterministic approach follows roughly that of Goodwin, Graebe, and Salgado [2001]. The version of the problem we formulate for solution in Section 14.3.3 is not quite the typical formulation. We have modified the typical formulation to be compatible with the SISO version of the optimal control problem of the preceding section.

Our development benefits from an appropriate view of what a state estimator does. We let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a complete SISO linear system, and recall that the equations governing the state $\boldsymbol{x}$, the estimated $\hat{\boldsymbol{x}}$, the output y, and the estimated output $\hat{y}$ are

$$\begin{aligned}\dot{\hat{\boldsymbol{x}}}(t) &= \boldsymbol{A}\hat{\boldsymbol{x}}(t) + \boldsymbol{b}u(t) + \boldsymbol{\ell}(y(t) - \hat{y}(t))\\ \hat{y}(t) &= \boldsymbol{c}^t\hat{\boldsymbol{x}}(t)\\ \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u(t)\\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t).\end{aligned}$$

The equations for the state estimate $\hat{\boldsymbol{x}}$ and output estimate $\hat{y}$ can be rewritten as

$$\begin{aligned}\dot{\hat{\boldsymbol{x}}}(t) &= (\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t)\hat{\boldsymbol{x}}(t) + \boldsymbol{b}u(t) + \boldsymbol{\ell}y(t)\\ \hat{y}(t) &= \boldsymbol{c}^t\hat{\boldsymbol{x}}(t).\end{aligned} \tag{14.2}$$

The form of this equation is essential to our point of view, as it casts the relationship between the output y and the estimated output $\hat{y}$ as the input/output relationship for the SISO linear system $\Sigma_{\boldsymbol{\ell}} = (\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t, \boldsymbol{\ell}, \boldsymbol{c}^t, \boldsymbol{0}_1)$. The objective is to choose $\boldsymbol{\ell}$ in such a way that the state error $\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ tends to zero as $t \to \infty$. If the system is observable this is equivalent to asking that the output error also $i = y - \hat{y}$ has the property that $\lim_{t\to\infty} i(t) = 0$. Thus we wish to include as part of the "cost" of a observer gain vector $\boldsymbol{\ell}$ a measure of the output error. We do this in a rather particular way, based on the following result.

14.5 Proposition *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a complete SISO linear system. If the output error corresponding to the initial value problem*

$$\begin{aligned}\dot{\hat{\boldsymbol{x}}}(t) &= \boldsymbol{A}\hat{\boldsymbol{x}}(t) + \boldsymbol{\ell}(y(t) - \hat{y}(t)), \qquad \hat{\boldsymbol{x}}(0) = -\boldsymbol{e}_0\\ \hat{y}(t) &= \boldsymbol{c}^t\hat{\boldsymbol{x}}(t)\\ \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}\delta(t), \qquad \boldsymbol{x}(0) = \boldsymbol{0}\\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t)\end{aligned}$$

satisfies $\lim_{t\to\infty} i(t) = 0$ then the output error has this same property for any input, and for any initial condition for the state and estimated state.

Proof Since the system is observable, by Lemma 10.44 the state error estimate tends to zero as $t \to \infty$ for all inputs and all state and estimated state initial conditions if and only if $\boldsymbol{\ell} \in \mathscr{D}(\Sigma)$. Thus we only need to show that by taking $u(t) = 0$ and $y(t) = h_\Sigma(t)$ in (14.2),

the output error i will tend to zero for all error initial conditions if only if $\boldsymbol{\ell} \in \mathscr{D}(\Sigma)$. The output error for the equations given in the statement of the proposition are

$$\begin{aligned}\dot{\boldsymbol{e}}(t) &= \dot{\boldsymbol{x}}(t) - \dot{\hat{\boldsymbol{x}}}(t)\\ &= \boldsymbol{A}\boldsymbol{x}(t) - \boldsymbol{A}\hat{\boldsymbol{x}}(t) + \boldsymbol{\ell}\boldsymbol{c}^t\hat{\boldsymbol{x}}(t) + \boldsymbol{\ell}h_\Sigma(t)\\ &= (\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t)\boldsymbol{e}(t),\end{aligned}$$

for $t > 0$, where we have used the fact that $\boldsymbol{x}(t) = h_\Sigma(t) = \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{b}$. From this the proposition immediately follows. ■

What this result allows us to do is specialise to the case when we consider the particular output that is the impulse response. That is, one portion of the cost of the state estimator will be the penalise the difference between the input to the system (14.2) and its output, when the input is the impulse response for Σ. We also wish to include in our cost a penalty for the "size" of the estimator. This penalty we take to be the L_2-norm of the impulse response for the system $\Sigma_{\boldsymbol{\ell}}$. This gives the total cost for the observer gain vector $\boldsymbol{\ell}$ to be

$$J(\boldsymbol{\ell}) = \int_0^\infty \Big(h_\Sigma(t) - \int_0^t h_{\Sigma_{\boldsymbol{\ell}}}(t-\tau)h_\Sigma(\tau)\,\mathrm{d}\tau\Big)^2\,\mathrm{d}t + \int_0^\infty h_{\Sigma_{\boldsymbol{\ell}}}(t)^2\,\mathrm{d}t. \tag{14.3}$$

With the above as motivation, we pose the following problem.

14.6 Optimal estimator problem Seek $\boldsymbol{\ell} \in \mathbb{R}^n$ so that $J(\boldsymbol{\ell}) \leq J(\tilde{\boldsymbol{\ell}})$ for every $\tilde{\boldsymbol{\ell}} \in \mathscr{D}(\Sigma)$. Such an observer gain vector $\boldsymbol{\ell}$ is called an ***optimal observer gain vector***. •

14.7 Remark The cost function for the optimal control problems of the preceding section seem somehow more natural than the cost (14.3) for the state estimator $\boldsymbol{\ell}$. This can be explained by the fact that we are formulating a problem for an estimator in terms of what is really a filter. We do not explain here the distinction between these concepts, as these are best explained in the context of stochastic control. For details we refer to [Kamen and Su 1999]. We merely notice that when one penalises a filter as if it were an estimator, one deals with quantities like impulse response, rather than with more tangible things like inputs and outputs that we see in the optimal control problems. One can also formulate more palatable versions cost for an estimator that make more intuitive sense than (14.3). However, the result will not then be the famous Kalman-Bucy filter that we come up with in Section 14.3.3. For a description of natural estimator costs, we refer to the "deterministic Kalman filter" described by Sontag [1998]. We also mention that the problem looks more natural, not in the time-domain, but in the frequency domain, where it relates to the model matching problem that we will discuss in Section 14.2.6. •

14.8 Remark Our deterministic development of the cost function for an optimal state estimator differs enough from that one normally sees that it is worth quickly presenting the normal formulation for completeness. There are two essential restrictions we made in our above derivation that need not be made in the general context. These are (1) the use of output rather than state error and (2) the use of the impulse response for Σ as the input to the estimator (14.2). The most natural setting for the general problem, as was the case for the optimal control setting of Remark 14.4, in MIMO. Thus we consider a MIMO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{0})$. The state can then be estimated, just as in the SISO case, via a Luenberger observer defined by an observer gain vector $\boldsymbol{L} \in \mathbb{R}^{m\times r}$. To define the analogue

of the cost (14.3) in the MIMO setting we need some notation. Let $\boldsymbol{S} \in \mathbb{R}^{\rho\times\rho}$ be symmetric and positive-definite. The ***S*-Frobenius norm** of a matrix $\boldsymbol{M} \in \mathbb{C}^{\sigma\times\rho}$ is given by

$$\|\boldsymbol{M}\|_{\boldsymbol{S}} = \big(\mathrm{tr}(\boldsymbol{M}^*\boldsymbol{S}\boldsymbol{M})\big)^{1/2}.$$

With this notation, the cost associated to an observer gain vector $\boldsymbol{L}$ is given by

$$J(\boldsymbol{L}) = \int_0^\infty \left\|e^{\boldsymbol{A}^t t} - e^{\boldsymbol{A}^t t}\boldsymbol{C}^t\boldsymbol{L}^t e^{(\boldsymbol{A}^t-\boldsymbol{C}^t\boldsymbol{L}^t)t}\right\|_{\boldsymbol{\Psi}} \mathrm{d}t + \int_0^\infty \left\|L^t e^{(\boldsymbol{A}^t-\boldsymbol{C}^t\boldsymbol{L}^t)t}\right\|_{\boldsymbol{\Phi}} \mathrm{d}t, \tag{14.4}$$

for symmetric, positive-definite matrices $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{r\times r}$. While we won't discuss this in detail, one can easily see that (14.4) is the natural analogue of (14.3) after one removes the restrictions we made to render the problem compatible with our SISO setting. We shall present, but not prove, the solution of this problem in Section 14.3.3. •

14.2 Tools for H_2 optimisation

Before we can solve the optimal feedback and the optimal state estimation problems posed in the preceding section, we must collect some tools for the job. The first few sections deal with factorisation of rational functions. Factorisation plays an important rôle in control synthesis for linear systems. While in the SISO context of this book these topics are rather elementary, things are not so straightforward in the MIMO setting; indeed, significant effort has been expended in this direction. The recent paper of Oară and Varga [2000] indicates "tens of papers" have been dedicated to this since the original work of Youla [1961]. An approach for controller synthesis based upon factorisation is given by Vidyasagar [1987]. In this book, factorisation will appear in the current chapter in our discussion of optimal control, as well as in Chapter 15 when we talk about synthesising controllers that solve the robust performance problem. Indeed, only the results of Section 14.2.3 will be used in this chapter. The remaining result will have to wait until Chapter 15 for their application. However, it is perhaps in the best interests of organisation to have all the factorisation results in one place. In Section 14.2.5 we consider a class of path independent integrals. These will allow us to simplify the optimisation problem, and make its solution "obvious." Finally, in Section 14.2.6 we describe "H_2 model matching." The subject of model matching will be revisited in Chapter 15 in the H_∞ context.

For the most part, this background is peripheral to applying the results of Sections 14.3. Thus the reader looking for the shortest route to "the point" can skip ahead to Section 14.3 after understanding how to compute the spectral factorisation of a polynomial.

14.2.1 An additive factorisation for rational functions

Our first factorisation is a simple one for rational functions. The reader will wish to recall some of our notation for system norms from Section 5.3.2. In that section, RH_2^+ and RH_∞^+ denoted the strictly proper and proper, respectively, rational functions with no poles in $\overline{\mathbb{C}}_+$. Also recall that RL_∞ denotes the proper rational functions with no poles on the imaginary axis. Given a rational function $R \in \mathbb{R}(s)$ let us denote $R^* \in \mathbb{R}(s)$ as the rational function $R^*(s) = R(-s)$. With this notation, let us additionally define

$$\mathrm{RH}_2^- = \{R \in \mathbb{R}(s) \mid R^* \in \mathrm{RH}_2^+\}$$
$$\mathrm{RH}_\infty^- = \{R \in \mathbb{R}(s) \mid R^* \in \mathrm{RH}_\infty^+\}.$$

Thus RH_2^- and RH_∞^- denote the strictly proper and proper, respectively, rational functions with no poles in $\overline{\mathbb{C}}_-$. Now we make a decomposition using this notation.

14.9 Proposition *If $R \in \mathrm{RL}_\infty$ then there exists unique rational functions $R_1, R_2 \in \mathbb{R}(s)$ so that*

(i) $R = R_1 + R_2$,

(ii) $R_1 \in \mathrm{RH}_2^-$, *and*

(iii) $R_2 \in \mathrm{RH}_\infty^+$.

Proof For existence, suppose that we have produced the partial fraction expansion of R. Since R has no poles on $i\mathbb{R}$, by Theorem C.6 it follows that the partial fraction expansion will have the form

$$R = \tilde{R}_1 + \tilde{R}_2 + C$$

where $\tilde{R}_1 \in \mathrm{RH}_2^-$, $\tilde{R}_2 \in \mathrm{RH}_2^+$, and $C \in \mathbb{R}$. Defining $R_1 = \tilde{R}_1$ and $R_2 = \tilde{R}_2 + C$ gives existence of the stated factorisation. For uniqueness, suppose that $R = \bar{R}_1 + \bar{R}_2$ for $\bar{R}_1 \in \mathrm{RH}_2^-$ and $\bar{R}_2 \in \mathrm{RH}_\infty^+$. Then, uniqueness of the partial fraction expansion guarantees that $\bar{R}_1 = \tilde{R}_1$ and $\bar{R}_2 = \tilde{R}_2 + C$. ■

Note that the proof of the result is constructive; it merely asks that we produce the partial fraction expansion.

14.2.2 The inner-outer factorisation of a rational function

A rational function $R \in \mathbb{R}(s)$ is ***inner*** if $R \in \mathrm{RH}_\infty^+$ and if $RR^* = 1$, and is ***outer*** if $R \in \mathrm{RH}_\infty^+$ and if R^{-1} is analytic in $\mathbb{C}_+$. Note that outer rational functions are exactly those that are proper, BIBO stable, and minimum phase. The following simple result tells the story about inner and outer rational functions as they concern us.

14.10 Proposition *If $R \in \mathrm{RH}_\infty^+$ then there exists unique $R_{\mathrm{in}}, R_{\mathrm{out}} \in \mathbb{R}(s)$ so that*

(i) $R = R_{\mathrm{in}} R_{\mathrm{out}}$,

(ii) R_{in} *is inner and* R_{out} *is outer, and*

(iii) $R_{\mathrm{in}}(0) = (-1)^{\ell_0}$, *where* $\ell_0 \geq 0$ *is the multiplicity of the root* $s = 0$ *for* R.

Proof Let $z_0, \dots, z_k \in \mathbb{C}_+$ be the collection of distinct zeros of R in $\mathbb{C}_+$ with each having multiplicity ℓ_j, $j = 0, \dots, k$. Suppose that the zeros are ordered so that

$$z_j = \begin{cases} 0, & j = 0 \\ \bar{z}_{j+\frac{1}{2}(n+\ell_0)} = z_{j+\ell_0}, & j = 1, \dots, \frac{1}{2}(n - \ell_0). \end{cases}$$

Thus the last $\frac{1}{2}(n - \ell_0)$ zeros are the complex conjugates of the $\frac{1}{2}(n - \ell_0)$ preceding zeros. For the remainder of the proof, let us denote $m = \frac{1}{2}(n - \ell_0)$. If we then define

$$R_{\mathrm{in}} = \prod_{j=1}^{k} \frac{(s - z_j)^{\ell_j}}{(s + z_j)^{\ell_j}}, \quad R_{\mathrm{out}} = \frac{R}{R_{\mathrm{in}}}, \tag{14.5}$$

then clearly R_{in} and R_{out} satisfy (i), (ii), and (iii). To see that these are unique, suppose that $\tilde{R}_{\mathrm{in}}$ and $\tilde{R}_{\mathrm{out}}$ are two inner and outer functions having the properties that $R = \tilde{R}_{\mathrm{in}} \tilde{R}_{\mathrm{out}}$ and $\tilde{R}_{\mathrm{in}}(0) = 1$. Since $\tilde{R}_{\mathrm{out}}^{-1}$ is analytic in $\mathbb{C}_+$, if $z \in \mathbb{C}_+$ is a zero for R it must be a zero for $\tilde{R}_{\mathrm{in}}$. Thus

$$\tilde{R}_{\mathrm{in}} = \prod_{j=1}^{k} (s - z_j)^{\ell_j} T(s)$$

for some $T \in \mathbb{R}(s)$. Because $\tilde{R}_{\text{in}}$ is inner we have

$$\begin{aligned} 1 &= \prod_{j=1}^{k} (s - z_j)^{\ell_j} T(s) \prod_{j=1}^{k} (-1)^{\ell_j} (s + z_j)^{\ell_j} T(-s) \\ &= T(s)T(-s)(-1)^{\ell_0} \prod_{j=1}^{k} (s - z_j)^{\ell_j} \prod_{j=1}^{k} (s + z_j)^{\ell_j} \end{aligned}$$

which gives

$$T(s)T(-s) = (-1)^{\ell_0} \frac{1}{\prod_{j=1}^{k} (s - z_j)^{\ell_j} \prod_{j=1}^{k} (s + z_j)^{\ell_j}}.$$

From this we conclude that either

$$T(s) = \pm \frac{1}{\prod_{j=1}^{k} (s - z_j)^{\ell_j}} \quad \text{or} \quad T(s) = \pm \frac{1}{\prod_{j=1}^{k} (s + z_j)^{\ell_j}}.$$

The facts that $\tilde{R}_{\text{in}} \in \text{RH}_\infty^+$ and $\tilde{R}_{\text{in}}(0) = (-1)^{\ell_0}$ ensures that

$$T(s) = \frac{1}{\prod_{j=1}^{k} (s + z_j)^{\ell_j}},$$

and from this the follows uniqueness. ■

Note that the proof is constructive; indeed, the inner and outer factor whose existence is declared are given explicitly in (14.5).

14.2.3 Spectral factorisation for polynomials

Spectral factorisation, while still quite simple in principle, requires a little more attention. Let us first look at the factorisation of a polynomial, as this is easily carried out.

14.11 Definition A polynomial $P \in \mathbb{R}[s]$ ***admits a spectral factorisation*** if there exists a polynomial $Q \in \mathbb{R}[s]$ with the properties

(i) $P = QQ^*$ and

(ii) all roots of Q lie in $\overline{\mathbb{C}}_-$.

If P satisfies these two conditions, then P ***admits a spectral factorisation by Q***. •

Let us now classify those polynomials that admit a spectral factorisation. A polynomial $P \in \mathbb{R}[s]$ is ***even*** if $P(-s) = P(s)$. Thus an even polynomial will have nonzero coefficients only for even powers of s, and in consequence the polynomial will be $\mathbb{R}$-valued on $i\mathbb{R}$.

14.12 Proposition *Let $P \in \mathbb{R}[s]$ have a zero at $s = 0$ of multiplicity $\tilde{k}_0 \geq 0$. Then P admits a spectral factorisation if and only if*

(i) P is even and

(ii) if $\frac{\tilde{k}_0}{2}$ is odd then $P(i\omega) \leq 0$ for $\omega \in \mathbb{R}$, and if $\frac{\tilde{k}_0}{2}$ is even then $P(i\omega) \geq 0$ for $\omega \in \mathbb{R}$.

Furthermore, if P admits a spectral factorisation by Q then Q is uniquely defined by requiring that the coefficient of the highest power of s in Q be positive.

Proof First suppose that P satisfies (i) and (ii). If z is a root for P, then so will $-z$ be a root since P is even. Since P is real, $\bar{z}$ and hence $-\bar{z}$ are also roots. Thus, if we factor P it will have the factored form

$$P(s) = A_0 s^{2k_0} \prod_{j_1=1}^{k_1} (a_{j_1}^2 - s^2) \prod_{j_2=1}^{k_2} (s^2 + b_{j_2}^2) \prod_{j_3=1}^{k_3} \big((s^2 - \sigma_{j_3}^2)^2 + 2\omega_{j_3}^2(s^2 + \sigma_{j_3}^2) + \omega_{j_3}^4\big). \quad (14.6)$$

where $2k_0 + 2k_1 + 2k_2 + 4k_3 = 2n = \deg(P)$. Here $a_{j_1} > 0$, $j_1 = 1, \ldots, k_1$, $b_{j_2} > 0$, $j_2 = 1, \ldots, k_2$, $\sigma_{j_3} \geq 0$, $j_3 = 1, \ldots, k_3$, and $\omega_{j_3} > 0$, $j_3 = 1, \ldots, k_3$. One may check that condition (ii) implies that $A_0 > 0$. By (ii), P must not change sign on the imaginary axis. This implies that all nonzero imaginary roots must have even multiplicity, since otherwise, terms like $(s^2 + b_{j_2}^2)$ will change sign on the imaginary axis. Therefore we may suppose that $k_2 = 0$ as the purely imaginary roots are then captured by the third product in (14.6). Now we can see that taking

$$Q(s) = \sqrt{A_0} s^{k_0} \prod_{j_1=1}^{k_1} (s + a_{j_1}) \prod_{j_3=1}^{k_3} \big((s + \sigma_{j_3})^2 + \omega_{j_3}^2\big)$$

satisfies the conditions of the definition of a spectral factorisation.

Now suppose that P admits a spectral factorisation by $Q \in \mathbb{R}[s]$. Since Q must have all roots in $\overline{\mathbb{C}}_-$, we may write

$$Q(s) = B_0 s^{k_0} \prod_{j_1=1}^{k_1} (s + a_{j_1}) \prod_{j_2=1}^{k_2} (s^2 + b_{j_2}^2) \prod_{j_3=1}^{k_3} ((s + \sigma_{j_3})^2 + \omega_{j_3}^2),$$

where $a_{j_1} > 0$, $j_1 = 1, \ldots, k_1$, $b_{j_2} > 0$, $j_2 = 1, \ldots, k_2$, $\sigma_{j_3}, \omega_{j_3} > 0$, $j_3 = 1, \ldots, k_3$. Computing QQ^* gives exactly the expression (14.6) with $A_0 = B_0^2$, thus showing that P satisfies (i) and (ii).

Uniqueness up to sign of the spectral factorisation follows directly from the above computations. ■

We denote the polynomial specified by the theorem by $[P]^+$ which we call the ***left half-plane spectral factor*** for P. The polynomial $[P]^- = Q^*$ we call the ***right half-plane spectral factor*** of P.

The following result gives a common example of when a spectral factorisation arises.

14.13 Corollary *If $P \in \mathbb{R}[s]$ then PP^* admits a spectral factorisation. Furthermore, if P has no roots on the imaginary axis, then the left half-plane spectral factor will be Hurwitz.*

Proof Suppose that

$$P(s) = p_n s^n + \cdots + p_1 s + p_0.$$

Since PP^* is even, if z is a root, so is $-z$. Thus if z is a root, so is $\bar{z}$, and therefore $-\bar{z}$. Arguing as in the proof of Proposition 14.12 this gives

$$P(s)P(-s) = p_n^2 s^{2k_0} \prod_{j_1=1}^{k_1} (a_{j_1}^2 - s^2) \prod_{j_2=1}^{k_s} \big((s^2 - \sigma_{j_2}^2)^2 + 2\omega_{j_2}^2(s^2 + \sigma_{j_2}^2) + \omega_{j_2}^4\big),$$

where $2k_0 + 2k_1 + 2k_2 + 4k_3 = 2\deg(P)$. Here $a_{j_1} > 0$, $j_1 = 1, \ldots, k_1$, $\sigma_{j_2} \geq 0$, $j_2 = 1, \ldots, k_2$, and $\omega_{j_2} > 0$, $j_2 = 1, \ldots, k_3$. This form for PP^* is a consequence of all imaginary axis roots

necessarily having even multiplicity. It is now evident that

$$Q(s) = |p_n| s^{k_0} \prod_{j_1=1}^{k_1} (s + a_{j_1}) \prod_{j_2=1}^{k_2} \big((s + \sigma_{j_2})^2 + \omega_{j_2}^2\big)$$

is a spectral factor. The final assertion of the corollary is clear. ■

Let us extract the essence of Proposition 14.12.

14.14 Remark The proof of Proposition 14.12 is constructive. To determine the spectral factorisation, one proceeds as follows.

1. Let P be an even polynomial for which $P(i\omega)$ does not change sign for $\omega \in \mathbb{R}$. Let A_0 be the coefficient of the highest power of s. Divide P by $|A_0|$ so that the resulting polynomial is monic, or -1 times a monic polynomial. Rename the resulting polynomial $\tilde{P}$.
2. Compute the roots for $\tilde{P}$ that will then fall into one of the following categories:
   (a) k_0 roots at $s = 0$;
   (b) k_1 roots $s = a_{j_1}$ for a_{j_1} real and positive (and the associated k_1 roots $s = -a_{j_1}$);
   (c) k_2 roots $s = \sigma_{j_2} + i\omega_{j_2}$ for σ_{j_2} and ω_{j_2} real and nonnegative (and the associated $3k_1$ roots $s = \sigma_{j_2} - i\omega_{j_2}$, $s = -\sigma_{j_2} + i\omega_{j_2}$, and $s = -\sigma_{j_2} - i\omega_{j_2}$).
3. A left half-plane spectral factor for $\tilde{P}$ will then be
$$\tilde{Q}(s) = \prod_{j_1=1}^{k_1} (s + a_{j_1}) \prod_{j_2=1}^{k_2} \big((s + \sigma_{j_2})^2 + \omega_{j_2}^2\big).$$
4. A left half-plane spectral factor for P is then $Q = \sqrt{|A_0|}\tilde{Q}$.

Let's compute the spectral factors for a few even polynomials.

14.15 Examples

1. Let us look at the easy case where $P(s) = s^2 + 1$. This polynomial is certainly even. However, note that it changes sign on the imaginary axis. Indeed, $P(\sqrt{2}i) = -1$ and $P(0) = 1$. This demonstrates why nonzero roots along the imaginary axis should appear with even multiplicity if a spectral factorisation is to be admitted.
2. The previous example can be "fixed" by considering instead $P(s) = s^4 + 2s^2 + 1 = (s^2 + 1)^2$. This polynomial has roots $\{i, i, -i, -i\}$. The left half-plane spectral factor is then $[P(s)]^+ = s^2 + 1$, and this is also the right half-plane spectral factor in this case.
3. Let $P(s) = -s^2 + 1$. Clearly P is even and is nonnegative on the imaginary axis. The roots of P are $\{1, -1\}$. Thus the left half-plane spectral factor is $[P(s)]^+ = s + 1$.
4. Let $P(s) = s^4 - s^2 + 1$. Obviously P is even and nonnegative on $i\mathbb{R}$. This polynomial has roots $\big\{\frac{\sqrt{3}}{2} + i\frac{1}{2}, \frac{\sqrt{3}}{2} - i\frac{1}{2}, -\frac{\sqrt{3}}{2} + i\frac{1}{2}, -\frac{\sqrt{3}}{2} - i\frac{1}{2}\big\}$. Thus we have
$$[P(s)]^+ = \big((s + \tfrac{\sqrt{3}}{2})^2 + \tfrac{1}{4}\big).$$ •

14.2.4 Spectral factorisation for rational functions

One can readily extend the notion of spectral factorisation to rational functions. The notion for such a factorisation is defined as follows.

14.16 Definition A rational function $R \in \mathbb{R}(s)$ ***admits a spectral factorisation*** if there exists a rational function $\rho \in \mathbb{R}(s)$ with the properties

(i) $R = \rho\rho^*$,

(ii) all zeros and poles of ρ lie in $\overline{\mathbb{C}}_-$.

If ρ satisfies these two conditions, then R ***admits a spectral factorisation by*** ρ. •

Let us now classify those rational functions that admit a spectral factorisation.

14.17 Proposition *Let $R \in \mathbb{R}(s)$ be a rational function with (N, D) its c.f.r. R admits a spectral factorisation if and only if both N and D admits a spectral factorisation.*

Proof Suppose that N and D admit a spectral factorisation by $[N]^+$ and $[D]^+$, respectively. Then we have

$$R = \frac{[N]^+[N]^-}{[D]^+[D]^-}.$$

Clearly $\frac{[N]^+}{[D]^+}$ is a spectral factorisation of R.

Conversely, suppose that R admits a spectral factorisation by ρ and let (N_ρ, D_ρ) be the c.f.r. of ρ. Note that all roots of N_ρ and D_ρ lie in $\overline{\mathbb{C}}_-$. We then have

$$R = \rho\rho^* = \frac{N_\rho N_\rho^*}{D_\rho D_\rho^*}.$$

Since N_ρ and D_ρ are coprime, so are $N_\rho N_\rho^*$ and $D_\rho D_\rho^*$. Since $D_\rho D_\rho^*$ is also monic, it follows that the c.f.r. of R is $(N_\rho N_\rho^*, D_\rho D_\rho^*)$. Thus the c.f.r.'s of R admits spectral factorisations. ■

Following our notation for the spectral factor for polynomials, let us denote by $[R]^+$ the rational function guaranteed by Proposition 14.17. As with polynomial spectral factorisation, there is a common type of rational function for which one wishes to obtain a spectral factorisation.

14.18 Corollary *If $R \in \mathbb{R}(s)$ then RR^* admits a spectral factorisation. Furthermore, if $R \in \mathrm{RL}_\infty$ then $[R]^+ \in \mathrm{RH}_\infty^+$.*

Proof Follows directly from Proposition 14.17 and Corollary 14.13. ■

14.2.5 A class of path independent integrals

In the proof of Theorem 14.25, it will be convenient to have at hand a condition that tells us when the integral of a quadratic function of t and its derivatives only depends on the data evaluated at the endpoints. To this end, we say that a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{(n+1)\times(n+1)}$ is ***integrable*** if for any interval $[a, b] \subset \mathbb{R}$ there exists a map $F\colon \mathbb{R}^{2n} \to \mathbb{R}$ so that for any n times continuously differentiable function $\phi\colon [a, b] \to \mathbb{R}$, we have

$$\int_a^b \sum_{i,j=0}^{n} M_{ij}\phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t = F\big(\phi(a), \phi(b), \phi^{(1)}(a), \phi^{(1)}(b), \dots, \phi^{(n-1)}(a), \phi^{(n-1)}(b)\big), \quad (14.7)$$

where M_{ij}, $i, j = 0, \dots, n$, are the components of the matrix $\boldsymbol{M}$. Thus the matrix $\boldsymbol{M}$ is integrable when the integral in (14.7) depends only on the values of ϕ and its derivatives at the endpoints, and not on the values of ϕ on the interval (a, b). As an example, consider the two symmetric 2×2 matrices

$$\boldsymbol{M}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{M}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

The first isintegrable on $[a, b]$ since

$$\int_a^b \sum_{i,j=0}^{2} M_{1,ij}\phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t = \int_a^b 2\phi(t)\dot{\phi}(t)\,\mathrm{d}t = \phi^2(b) - \phi^2(b).$$

$\boldsymbol{M}_2$ is not integrable. Indeed we have

$$\int_a^b \sum_{i,j=0}^{2} M_{2,ij}\phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t = \int_a^b \phi^2(t)\,\mathrm{d}t,$$

and this integral will depend not only on the value of ϕ at the endpoints of the interval, but on its value between the endpoints. For example, the two functions $\phi_1(t) = t$ and $\phi_2(t) = \frac{t^3}{3}$ have the same value at the endpoints of the interval $[-1, 1]$, and their first derivatives also have the same value at the endpoints, but

$$\int_{-1}^{1} \phi_1^2(t)\,\mathrm{d}t = \frac{2}{3}, \quad \int_{-1}^{1} \phi_2^2(t)\,\mathrm{d}t = \frac{2}{63}.$$

The following result gives necessary and sufficient conditions on the coefficients M_{ij}, $i, j = 1, \ldots, n$, for a symmetric matrix $\boldsymbol{M}$ to be integrable.

14.19 Proposition *A symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{(n+1)\times(n+1)}$ is integrable if and only if the polynomial*

$$P(s) = \sum_{i,j=0}^{n} M_{ij}\big(s^i(-s)^j + (-s)^i s^j\big)$$

is the zero polynomial.

Proof First let us consider the terms

$$\int_a^b \phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t$$

when $i + j$ is odd. Suppose without loss of generality that $i > j$. If $i = j + 1$ then the integral

$$\int_a^b \phi^{(j+1)}(t)\phi^{(j)}(t)\,\mathrm{d}t = \frac{\phi^{(j)}(t)}{2}\Big|_{t=a}^{t=b}$$

If $i > j + 1$ then a single integration by parts yields

$$\int_a^b \phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t = \phi^{(i-1)}(t)\phi^{(j)}(t)\Big|_{t=a}^{t=b} - \int_a^b \phi^{(i-1)}(t)\phi^{(j+1)}(t)\,\mathrm{d}t.$$

One can carry on in this manner $\frac{1}{2}(i - j - 1)$ times until one arrives at an expression that is a sum of evaluations at the endpoints, and an integral of the form

$$\int_a^b \phi^{(k+1)}(t)\phi^{(k)}(t)\,\mathrm{d}t$$

which is then evaluated to

$$\frac{\phi^{(k)}(t)}{2}\Big|_{t=a}^{t=b}.$$

Thus all terms where $i+j$ is odd lead to expressions that are only endpoint dependent.

Now we look at the terms of the form

$$\int_a^b \phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t$$

when $i+j$ is even. We integrate by parts as above $\frac{1}{2}(i-j)$ times to get an expression that is a sum of terms that are evaluations at the endpoints, and an integral of the form

$$\frac{1}{2}\big((-1)^i+(-1)^j\big)\int_a^b \big(\phi^{((i+j)/2)}(t)\big)^2\,dt.$$

Thus we will have

$$\int_a^b \sum_{i,j=0}^n M_{ij}\phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t = I_0 + \frac{1}{2}\sum_{i,j=0}^n \int_a^b M_{ij}\big((-1)^i+(-1)^j\big)\big(\phi^{((i+j)/2)}(t)\big)^2\,dt, \quad (14.8)$$

where I_0 is a sum of terms that are evaluations at the endpoints. In order for the expression (14.8) to involve evaluations at the endpoints for every function ϕ, it must be the case that

$$\sum_{i,j=0}^n M_{ij}\big((-1)^i+(-1)^j\big)\big(\phi^{((i+j)/2)}(t)\big)^2 = 0$$

for all $t\in[a,b]$. Now we compute directly that

$$\sum_{i,j=0}^n M_{ij}\big((-1)^i+(-1)^j\big)s^{i+j} = \sum_{i,j=0}^n M_{ij}\big(s^i(-s)^j+(-s)^i s^j\big). \quad (14.9)$$

From this the result follows. ■

Note that the equation (14.9) tells us that $\boldsymbol{M}$ is integrable if and only if for $k=0,\dots,2n$ we have

$$\sum_{i+j=2k} M_{ij} = 0.$$

The result yields the following corollary that makes contact with the Riccati equation method of Section 14.3.2.

14.20 Corollary *A symmetric matrix $\boldsymbol{M}\in\mathbb{R}^{(n+1)\times(n+1)}$ is integrable if and only if there exists a symmetric matrix $\boldsymbol{P}\in\mathbb{R}^{n\times n}$ so that*

$$\int_a^b \sum_{i,j=0}^n M_{ij}\phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t = \boldsymbol{x}^t(b)\boldsymbol{P}\boldsymbol{x}(b) - \boldsymbol{x}^t(a)\boldsymbol{P}\boldsymbol{x}(a),$$

where $\boldsymbol{x}(t) = \big(\phi(t),\phi^{(1)}(t),\dots,\phi^{(n-1)}(t)\big)$.

Proof From Proposition 14.19 we see that if $\boldsymbol{M}$ is integrable, then the only contributions to the integral come from terms of the form

$$\int_a^b M_{ij}\phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t,$$

where $i+j$ is odd. As we saw in the proof of Proposition 14.19, the resulting integrated expressions are sums of pairwise products of the form

$$\phi^{(k)}(t)\phi^{(\ell)}(t)\Big|_{t=a}^{t=b}$$

where $k, \ell \in \{0, 1, \dots, n-1\}$. This shows that there is a matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ so that

$$\int_a^b \sum_{i,j=0}^{n} M_{ij}\phi^{(i)}(t)\phi^{(j)}(t)\,\mathrm{d}t = \boldsymbol{x}^t(b)\boldsymbol{P}\boldsymbol{x}(b) - \boldsymbol{x}^t(a)\boldsymbol{P}\boldsymbol{x}(a).$$

That $\boldsymbol{P}$ may be taken to be symmetric is a consequence of the fact that the expressions $\boldsymbol{x}^t(b)\boldsymbol{P}\boldsymbol{x}(b)$ and $\boldsymbol{x}^t(a)\boldsymbol{P}\boldsymbol{x}(a)$ depend only on the symmetric part of $\boldsymbol{P}$ (see Section A.6). ■

It is in the formulation of the corollary that the notion of path independence in the title of this section is perhaps best understood, as here we can interpret the integral (14.7) as a path integral in $\mathbb{R}^{n+1}$ where the coordinate axes are the values of ϕ and its first n derivatives.

The following result is key, and combines our discussion in this section with the spectral factorisation of the previous section.

14.21 Proposition *If $P, Q \in \mathbb{R}[s]$ are coprime then the polynomial $PP^* + QQ^*$ admits a spectral factorisation. Let F be the left half-plane spectral factor for this polynomial. Then for any n times continuously differentiable function $\phi\colon [a,b] \to \mathbb{R}$, the integral*

$$\int_a^b \Big(\big(P\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)\phi(t)\big)^2 + \big(Q\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)\phi(t)\big)^2 - \big(F\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)\phi(t)\big)^2 \Big)\,\mathrm{d}t$$

is expressible in terms of the value of the function ϕ and its derivatives at the endpoints of the interval $[a,b]$.

Proof For the first assertion, we refer to Exercise E14.2. Let us suppose that $n = \deg(P) \geq \deg(Q)$ so that $\deg(F) = \deg(P)$. Let us also write

$$\begin{aligned} P(s) &= p_n s^n + \cdots + p_1 s + p_0 \\ Q(s) &= q_n s^n + \cdots + q_1 s + q_0 \\ F(s) &= f_n s^n + \cdots + f_1 s + f_0. \end{aligned}$$

According to Proposition 14.19, we should show that the coefficients of the even powers of s in the polynomial $P^2 + Q^2 - F^2$ vanish. By definition of F we have

$$PP^* + QQ^* = FF^*,$$

or

$$\Big(\sum_{i=0}^{n} p_i s^i\Big)\Big(\sum_{j=0}^{n} (-1)^j p_j s^j\Big) + \Big(\sum_{i=0}^{n} q_i s^i\Big)\Big(\sum_{j=0}^{n} (-1)^j q_j s^j\Big) = \Big(\sum_{i=0}^{n} f_i s^i\Big)\Big(\sum_{j=0}^{n} (-1)^j f_j s^j\Big)$$

$$\implies \quad \sum_{k=0}^{n}\sum_{i+j=2k} (-1)^i p_i p_j s^{2k} + \sum_{k=0}^{n}\sum_{i+j=2k} (-1)^i q_i q_j s^{2k} = \sum_{k=0}^{n}\sum_{i+j=2k} (-1)^i f_i f_j s^{2k}.$$

In particular, it follows that for $k = 0, \dots, n$ we have

$$\sum_{i+j=2k} (p_i p_j + q_i q_j) - \sum_{i+j=2k} f_i f_j = 0.$$

However, these are exactly the coefficients of the even powers of s in the polynomial $P^2 + Q^2 - F^2$, and thus our result follows. ■

14.2.6 H_2 model matching

In this section we state a so-called "model matching problem" that on the surface has nothing to do with the optimal control problems of Section 14.1. However, we shall directly use the solution to this model matching problem to solve Problem 14.6, and in Section 14.4.2 we shall see that the famous LQG control scheme is the solution of a certain model matching problem.

The problem in this section is precisely stated as follows.

14.22 H_2 model matching problem Given $T_1, T_2 \in \mathrm{RH}_2^+$, find $\theta \in \mathrm{RH}_2^+$ so that $\|T_1 - \theta T_2\|_2^2 + \|\theta\|_2^2$ is minimised. •

The norm $\|\cdot\|_2$ is the H_2-norm defined in Section 5.3.2:

$$\|f\|_2^2 = \int_{-\infty}^{\infty} |f(i\omega)|^2 \,\mathrm{d}\omega.$$

The idea is that given T_1 and T_2, we find θ so that the cost

$$J_{T_1,T_2}(\theta) = \int_{-\infty}^{\infty} |T_1(i\omega) - \theta(i\omega)T_2(i\omega)|^2 \,\mathrm{d}\omega + \int_{-\infty}^{\infty} |\theta(i\omega)|^2 \,\mathrm{d}\omega$$

is minimised. The problem may be thought of as follows. Think of T_1 as given. We wish to see how close we can get to T_1 with a multiple of T_2, with closeness being measured by the H_2-norm. The cost should be thought of as being the cost of the difference of T_1 from θT_2, along with a penalty for the size of the matching parameter θ.

To solve the H_2 model matching problem we turn the problem into a problem much like that posed in Section 14.1. To do this, we let $\Sigma_j = (\boldsymbol{A}_j, \boldsymbol{b}_j, \boldsymbol{c}_j^t, \boldsymbol{0}_1)$, $j \in \{1, 2\}$, be the canonical minimal realisations of T_1 and T_2. The following lemma gives a time-domain characterisation of the model matching cost $J_{T_1,T_2}(\theta)$.

14.23 Lemma *If u_θ is the inverse Laplace transform of $\theta \in \mathrm{RH}_2^+$ then*

$$J_{T_1,T_2}(\theta) = \frac{1}{2\pi} \int_0^{\infty} \big(y(t)^2 + u_\theta(t)^2\big) \,\mathrm{d}t,$$

where y satisfies

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}u_\theta(t) \\ y(t) &= \boldsymbol{c}^t\boldsymbol{x}(t), \end{aligned} \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0,$$

with

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{b}_2 \end{bmatrix}, \quad \boldsymbol{c}^t = \begin{bmatrix} \boldsymbol{c}_1^t & -\boldsymbol{c}_2^t \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0} \end{bmatrix}.$$

Proof By Parseval's theorem we immediately have

$$\int_{-\infty}^{\infty} |\theta(i\omega)|^2 \,\mathrm{d}\omega = \frac{1}{2\pi} \int_0^{\infty} |u_\theta(t)|^2 \,\mathrm{d}t,$$

since $\theta \in \mathrm{RH}_2^+$. We also have

$$T_1(s) = \boldsymbol{c}_1^t(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1}\boldsymbol{b}_1, \quad T_2(s)\theta(s) = \boldsymbol{c}_2^t(s\boldsymbol{I}_{n_2} - \boldsymbol{A}_2)^{-1}\boldsymbol{b}_2\theta(s),$$

giving

$$\mathscr{L}^{-1}(T_1(s))(t) = \boldsymbol{c}_1^t e^{\boldsymbol{A}_1 t}\boldsymbol{b}, \quad \mathscr{L}^{-1}(T_2(s)\theta(s))(t) = \int_0^t \boldsymbol{c}_2^2 e^{\boldsymbol{A}_2(t-\tau)}\boldsymbol{b}_2 u_\theta(\tau)\,\mathrm{d}\tau.$$

In other words, the inverse Laplace transform of $T_1(s) - \theta(s)T_2(s)$ is exactly y, if y is specified as in the statement of the lemma. The result then follows by Parseval's theorem since $T_1(s) - \theta(s)T_2(s) \in \mathrm{RH}_2^+$. ■

The punchline is that we now know how to minimise $J_{T_1,T_2}(\theta)$ by finding u_θ as per the methods of Section 14.3. The following result contains the upshot of translating our problem here to the previous framework.

14.24 Proposition *Let $\Sigma_j = (\boldsymbol{A}_j, \boldsymbol{b}_j, \boldsymbol{c}_j^t, \boldsymbol{0}_1)$ be the canonical minimal realisation of $T_j \in \mathrm{RH}_2^+$, $j \in \{1,2\}$, let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be as defined in Lemma 14.23, and let $\boldsymbol{f} = \begin{bmatrix}\boldsymbol{f}_1^t & \boldsymbol{f}_2^t\end{bmatrix}$ be the solution to Problem 14.3 for Σ. If we denote $\tilde{\Sigma}_1 = (\boldsymbol{A}_1, \boldsymbol{b}_1, \boldsymbol{f}_1^t, \boldsymbol{0}_1)$ and $\tilde{\Sigma}_2 = (\boldsymbol{A}_2 - \boldsymbol{b}_2\boldsymbol{f}_2, \boldsymbol{b}_2, \boldsymbol{f}_2^2, \boldsymbol{0}_1)$, then the solution to the model matching problem Problem 14.22 is*

$$\theta = (-1 + T_{\tilde{\Sigma}_2})T_{\tilde{\Sigma}_1}.$$

Proof By Theorem 14.25, Corollary 14.26, and Lemma 14.23 we know that $u_\theta(t) = -\boldsymbol{f}\boldsymbol{x}(t)$ where $\boldsymbol{x}$ satisfies the initial value problem

$$\dot{\boldsymbol{x}}(t) = (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)\boldsymbol{x}(t), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

Then we have

$$\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t = \begin{bmatrix}\boldsymbol{A}_1 & \boldsymbol{0}\\ -\boldsymbol{b}_2\boldsymbol{f}_1^t & \boldsymbol{A}_2 - \boldsymbol{b}_2\boldsymbol{f}_2^t\end{bmatrix}.$$

Therefore

$$(s\boldsymbol{I}_n - (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)^{-1} = \begin{bmatrix}(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1} & \boldsymbol{0}\\ -(s\boldsymbol{I}_{n_2} - (\boldsymbol{A}_2 - \boldsymbol{b}_2\boldsymbol{f}_2^t))^{-1}\boldsymbol{b}_2\boldsymbol{f}_1^t(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1} & (s\boldsymbol{I}_{n_2} - (\boldsymbol{A}_2 - \boldsymbol{b}_2\boldsymbol{f}_2^t))^{-1}\end{bmatrix},$$

giving

$$(s\boldsymbol{I}_n - (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)^{-1}\boldsymbol{x}_0 = \begin{bmatrix}(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1}\boldsymbol{b}_1\\ -(s\boldsymbol{I}_{n_2} - (\boldsymbol{A}_2 - \boldsymbol{b}_2\boldsymbol{f}_2^t))^{-1}\boldsymbol{b}_2\boldsymbol{f}_1^t(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1}\boldsymbol{b}_1\end{bmatrix}.$$

This finally gives

$$\theta(s) = -\boldsymbol{f}_1^t(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1}\boldsymbol{b}_1 + \boldsymbol{f}_2^t(s\boldsymbol{I}_{n_2} - (\boldsymbol{A}_2 - \boldsymbol{b}_2\boldsymbol{f}_2^t))^{-1}\boldsymbol{b}_2\boldsymbol{f}_1^t(s\boldsymbol{I}_{n_1} - \boldsymbol{A}_1)^{-1}\boldsymbol{b}_1,$$

as given in the statement of the result. ■

14.3 Solutions of optimal control and state estimation problems

With the developments of the previous section under our belts, we are ready to state our main results. The essential results are those for the optimal control problems, Problems 14.1 and 14.3, and these are given in Section 14.3.1. By going through the model matching problem of Section 14.2.6, the optimal state estimation problem, Problem 14.6, is seen to follow

from the optimal control results. The way of approaching the optimal control problem we undertake is not the usual route, and is not easily adaptable to MIMO control problems. In this latter case, the typical development uses the so-called "Riccati equation." In Section 14.3.2 we indicate how the connection between our approach and the Riccati equation approach is made. Here the developments of Section 5.4, that seemed a bit unmotivated at first glance, become useful.

14.3.1 Optimal control results

The main optimal control result is the following.

14.25 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be an observable SISO linear system in controller canonical form with (N, D) the c.f.r. for T_Σ. Let $F \in \mathbb{R}[s]$ be defined by*

$$F = -[DD^* + NN^*]^+ + D.$$

For $\boldsymbol{x}_0 \in \mathbb{R}^n$, a control $u \in \mathscr{U}_{\boldsymbol{x}_0}$ solves Problem 14.1 if and only if it satisfies

$$u(t) = F\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t),$$

and is defined on $[0, \infty)$.

Proof First note that observability of Σ implies that the polynomial $DD^* + NN^*$ is even and positive on the imaginary axis (see Exercise E14.2), so the left half-plane spectral factor $[DD^* + NN^*]^+$ does actually exist. The system equations are given by (14.1), which we write as

$$D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t) = u(t), \quad N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t) = y(t).$$

Let $u \in \mathscr{U}_{\boldsymbol{x}_0}$. If u is defined on $[0, T]$ then we may extend u to $[0, \infty)$ by asking that $u(t) = 0$ for $t > T$. Note that if we do this, the value of $J_{\boldsymbol{x}_0}(u)$ is unchanged. Therefore, let us suppose that u is defined on $[0, \infty)$. Now let us write

$$\begin{aligned} J_{\boldsymbol{x}_0}(u) &= \int_0^\infty \big(y^2_{u,\boldsymbol{x}_0}(t) + u^2(t)\big)\,\mathrm{d}t \\ &= \int_0^\infty \Big(\big(N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t)\big)^2 + \big(D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t)\big)^2\Big)\,\mathrm{d}t. \end{aligned}$$

Now define $\tilde{F} = [DD^* + NN^*]^+$ and write

$$J_{\boldsymbol{x}_0}(u) = \int_0^\infty \Big(\big(N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t)\big)^2 + \big(D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t)\big)^2 - \big(\tilde{F}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t)\big)^2\Big)\,\mathrm{d}t + \int_a^b \big(\tilde{F}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t)\big)^2\,\mathrm{d}t.$$

By Proposition 14.21 the first integral depends only on the value of $x_{u,\boldsymbol{x}_0}(t)$ and its derivatives at its initial point and terminal point. Thus it cannot be changed by changing the control. This means the best we can hope to achieve will be by choosing u so that

$$\tilde{F}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t) = 0.$$

Note that if $x_{u,\boldsymbol{x}_0}$ does satisfy this condition, then it and its derivatives do indeed go to zero since the zeros of $\tilde{F}$ are in $\mathbb{C}_-$ (see Exercise E14.2). Clearly, since $D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t) = u$, this can be done if and only if $u(t)$ satisfies

$$u(t) = -\tilde{F}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t) + D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)x_{u,\boldsymbol{x}_0}(t),$$

as specified in the theorem statement. ■

Note that, interestingly, the control law is independent of the initial condition $\boldsymbol{x}_0$ for the state. Also, motivated by Corollary 14.20, let $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ be the symmetric matrix with the property that

$$\int_a^b \Big(\big(N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)\phi(t)\big)^2 + \big(D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)\phi(t)\big)^2 - \big(\tilde{F}\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)\phi(t)\big)^2 \Big)\,\mathrm{d}t = \boldsymbol{x}(b)\boldsymbol{P}\boldsymbol{x}(b) - \boldsymbol{x}(a)\boldsymbol{P}\boldsymbol{x}(a),$$

where $\boldsymbol{x}(t) = (\phi(t), \phi^{(1)}(t), \dots, \phi^{(n-1)}(t))$. We then see that the cost of the optimal control with initial state condition $\boldsymbol{x}_0$ is

$$J_{\boldsymbol{x}_0}(u) = \boldsymbol{x}_0^t \boldsymbol{P} \boldsymbol{x}_0.$$

Let us compare Theorem 14.25 to something we are already familiar with, namely the method of finding a state feedback vector $\boldsymbol{f}$ to stabilise a system.

14.26 Corollary *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, 0_1)$ be a controllable and detectable SISO linear system, and let $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ be an invertible matrix with the property that $(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}\boldsymbol{b})$ is in controller canonical form. If $\tilde{\boldsymbol{f}} = (\tilde{f}_0, \tilde{f}_1, \dots, \tilde{f}_{n-1}) \in \mathbb{R}^n$ is defined by requiring that*

$$\tilde{f}_{n-1}s^{n-1} + \cdots + \tilde{f}_1 s + \tilde{f}_0 = [D(s)D(-s) + N(s)N(-s)]^+ - D(s),$$

where (N, D) is the c.f.r. for T_Σ, then $\boldsymbol{f} = \boldsymbol{T}^t\tilde{\boldsymbol{f}}$ is a solution of Problem 14.3.

Proof Let us denote by

$$\tilde{\Sigma} = (\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{c}}^t, 0_1) = (\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}\boldsymbol{b}, \boldsymbol{T}^{-t}\boldsymbol{c}^t, 0_1)$$

the system in coordinates where $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}})$ is in controller canonical form. We make the observation that the cost function is independent of state coordinates. That is to say, if $\tilde{\boldsymbol{x}}_0 = \boldsymbol{T}\boldsymbol{x}_0$ then $J_{\tilde{\boldsymbol{c}},\tilde{\boldsymbol{x}}_0}(\tilde{f}) = J_{\boldsymbol{c},\boldsymbol{x}_0}(\boldsymbol{f})$, where $\tilde{\boldsymbol{f}} = \boldsymbol{T}^{-t}\boldsymbol{f}$. Thus it suffices to prove the corollary for the system in controller canonical form.

Thus we proceed with the proof under this assumption. Theorem 14.25 gives the form of the control law u in terms of $x_{u,\boldsymbol{x}_0}$ that must be satisfied for optimality. It remains to show that the given u is actually in state feedback form. However, since the system is in controller canonical form we have $\boldsymbol{x}_{u,\boldsymbol{x}_0}(t) = (x(t), x^{(1)}(t), \dots, x^{(n-1)}(t))$, and so it does indeed follow that

$$\Big([D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)D\big(-\tfrac{\mathrm{d}}{\mathrm{d}t}\big) + N\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big)N\big(-\tfrac{\mathrm{d}}{\mathrm{d}t}\big)]^+ - D\big(\tfrac{\mathrm{d}}{\mathrm{d}t}\big) \Big) x_{u,\boldsymbol{x}_0}(t) = -\boldsymbol{f}^t \boldsymbol{x}_{u,\boldsymbol{x}_0}(t),$$

with $\boldsymbol{f}$ as defined in the statement of the corollary. The corollary now follows since in this case we have $J_{\boldsymbol{c},\boldsymbol{x}_0}(\boldsymbol{f}) = J_{\boldsymbol{x}_0}(u)$. ■

14.27 Remarks

1. The corollary gives, perhaps, the easiest way of seeing what is going on with Theorem 14.25 in that it indicates that the control law that solves Problem 14.1 is actually a state feedback control law. This is not obvious from the statement of the problem, but is a consequence of Theorem 14.25.
2. The assumption of controllability in the corollary may be weakened to stabilisability. In this case, to construct the optimal state feedback vector, one would put the system into the canonical form of Theorem 2.39, with $(\boldsymbol{A}_1, \boldsymbol{b}_1)$ in controller canonical form. One then constructs $\boldsymbol{f}^t = [\,\boldsymbol{f}_1^t\ 0^t\,]$, where $\boldsymbol{f}_1$ is defined by applying the corollary to $\Sigma_1 = (\boldsymbol{A}_1, \boldsymbol{b}_1, \boldsymbol{c}_1^t, 0_1)$. •

Let us look at an example of an application of Theorem 14.25, or more properly, of Corollary 14.26.

14.28 Example **(Example 6.50 cont'd)** The system we look at in this example had

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

In order to specify an optimal control problem, we also need an output vector $\boldsymbol{c}$ to specify an output cost in Problem 14.3. In actuality, one may wish to modify the output vector to obtain suitable controller performance. Let us look at this a little here by choosing two output vectors

$$\boldsymbol{c}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{c}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

To simplify things, note that $(\boldsymbol{A}, \boldsymbol{b})$ is in controller canonical form. Thus Problems 14.1 and 14.3 are related in a trivial manner. If $\Sigma_1 = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}_1^t, \boldsymbol{0}_1)$ and $\Sigma_2 = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}_2^t, \boldsymbol{0}_1)$, then one readily computes

$$T_{\Sigma_1}(s) = \frac{1}{s^2+1}, \quad T_{\Sigma_2}(s) = \frac{s}{s^2+1}.$$

Let us denote $(N_1(s), D_1(s)) = (1, s^2+1)$ and $(N_2(s), D_2(s)) = (s, s^2+1)$. We then compute

$$\begin{aligned} D_1(s)D_1(-s) + N_1(s)N_1(-s) &= s^4 + 2s^2 + 2 \\ D_2(s)D_2(-s) + N_2(s)N_2(-s) &= s^4 + s^2 + 1. \end{aligned}$$

Following the recipe of Remark 14.14, one determines that

$$\begin{aligned} [D_1(s)D_1(-s) + N_1(s)N_1(-s)]^+ &= s^2 + 2\sqrt[4]{2}\sin\tfrac{\pi}{8}s + \sqrt{2} \\ [D_2(s)D_2(-s) + N_2(s)N_2(-s)]^+ &= s^2 + s + 1. \end{aligned}$$

Then one follows the recipe of Corollary 14.26 and computes

$$\begin{aligned} [D_1(s)D_1(-s) + N_1(s)N_1(-s)]^+ - D_1(s) &= 2\sqrt[4]{2}\sin\tfrac{\pi}{8}s + \sqrt{2} - 1 \\ [D_2(s)D_2(-s) + N_2(s)N_2(-s)]^+ - D_2(s) &= s. \end{aligned}$$

Thus the two optimal state feedback vectors are

$$\boldsymbol{f}_1 = \begin{bmatrix} \sqrt{2}-1 \\ 2\sqrt[4]{2}\sin\frac{\pi}{8} \end{bmatrix}, \quad \boldsymbol{f}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The eigenvalues of the closed-loop system matrices

$$\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}_1^t, \quad \boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}_2^t$$

are $\{\sqrt[4]{2}(-\sin\frac{\pi}{8} \pm i\sqrt{1-\sin^2\frac{\pi}{8}})\} \approx \{-0.46 \pm i1.10\}$ and $\{-\frac{1}{2} \pm i\frac{\sqrt{3}}{2}\}$. In Figure 14.1 are plotted the trajectories for the closed-loop system in the (x_1, x_2)-plane for the initial condition $(1, 1)$. One can see a slight difference in that in the optimal control law for $\boldsymbol{c}_2$ the x_2-component of the solution is tending to zero somewhat more sharply. In practice, one uses ideas such as this to refine a controller based upon the principles we have outlined here. •

14.3.2 Relationship with the Riccati equation

In classical linear quadratic regulator theory, one does not normally deal with polynomials in deriving the optimal control law. Normally, one solves a quadratic matrix equation called the "algebraic Riccati equation."[1] In this section, that can be regarded as optional, we make this link explicit by proving the following theorem.

[1] After Jacopo Francesco Riccati (1676–1754) who made original contributions to the theory of differential equations. Jacopo also had a son, Vincenzo Riccati (1707–1775), who was a mathematician of some note.

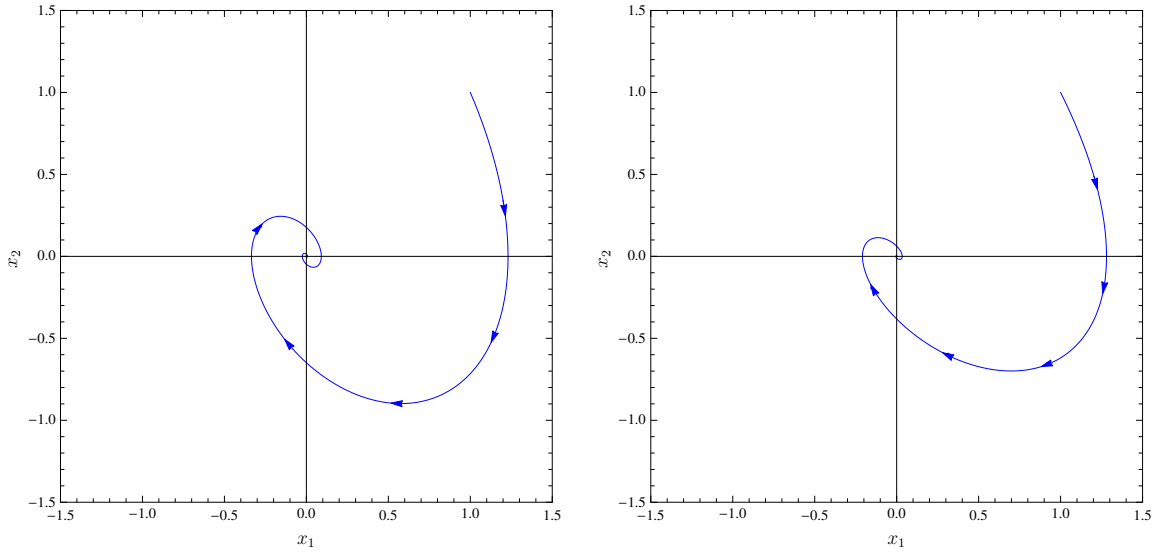


Figure 14.1 Optimal trajectories for output vector $\boldsymbol{c}_1$ (left) and $\boldsymbol{c}_2$ (right)

14.29 Theorem *If $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \mathbf{0}_1)$ is a controllable SISO linear system, then there exists at least one symmetric solution $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ of the equation*

$$\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} - \boldsymbol{P}\boldsymbol{b}\boldsymbol{b}^t\boldsymbol{P} = -\boldsymbol{c}\boldsymbol{c}^t, \tag{14.10}$$

and, furthermore, exactly one of these solutions is positive-definite. The equation (14.10) is called the ***algebraic Riccati equation****.*

Proof Let

$$\tilde{\Sigma} = (\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{c}}^t, \mathbf{0}_1) = (\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}\boldsymbol{b}, \boldsymbol{T}^{-t}\boldsymbol{c}^t, \mathbf{0}_1)$$

be the system in state coordinates where $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}})$ is in controller canonical form. Note that if $\tilde{\boldsymbol{P}}$ is a solution for the equation

$$\tilde{\boldsymbol{A}}^t\tilde{\boldsymbol{P}} + \tilde{\boldsymbol{P}}\tilde{\boldsymbol{A}} - \tilde{\boldsymbol{P}}\tilde{\boldsymbol{b}}\tilde{\boldsymbol{b}}^t\tilde{\boldsymbol{P}} = -\tilde{\boldsymbol{c}}\tilde{\boldsymbol{c}}^t,$$

then $\boldsymbol{P} = \boldsymbol{T}\tilde{\boldsymbol{P}}\boldsymbol{T}^{-1}$ is a solution to (14.10). Therefore, without loss of generality we assume that $(\boldsymbol{A}, \boldsymbol{b})$ is in controller canonical form. Let $T_\Sigma(s) = \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}$ be the transfer function for Σ with (N, D) its c.f.r. Then we have

$$T_\Sigma^t(-s)T_\Sigma(s) + 1 = \boldsymbol{b}^t(-s\boldsymbol{I}_n - \boldsymbol{A})\boldsymbol{c}\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} + 1,$$

which is nonnegative for $s = i\omega$. Let us denote this rational function by R. Defining $\tilde{F} = [RDD^*]^+$ we see that

$$T_\Sigma^t(-s)T_\Sigma(s) + 1 = \frac{\tilde{F}(s)\tilde{F}(-s)}{D(s)D(-s)}$$

(here we have used the fact that $\boldsymbol{A}$ and $\boldsymbol{A}^t$ have the same eigenvalues). Note that $\tilde{F}$ is monic and of degree n. Because $\boldsymbol{A}$ is in controller canonical form, a simple computation along the lines of that in the proof of Theorem 6.49 shows that if $\boldsymbol{f} \in \mathbb{R}^n$ is defined by

$$\boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = \frac{\tilde{F}(s) - D(s)}{D(s)} \tag{14.11}$$

then $\det(s\boldsymbol{I}_n - (\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t)) = \tilde{F}(s)$. By the spectral factorisation theorem, $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ is thus Hurwitz. By Theorem 5.32(i) there exists a unique positive-definite solution to the equation

$$(\boldsymbol{A}^t - \boldsymbol{f}\boldsymbol{b}^t)\boldsymbol{P} + \boldsymbol{P}(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t) = -\boldsymbol{c}\boldsymbol{c}^t - \boldsymbol{f}\boldsymbol{f}^t. \tag{14.12}$$

Let us denote this solution by $\boldsymbol{P}_1$. A straightforward computation shows that the previous equation is equivalent to

$$\boldsymbol{A}^t\boldsymbol{P}_1 + \boldsymbol{P}_1\boldsymbol{A} - \boldsymbol{P}_1\boldsymbol{b}\boldsymbol{b}^t\boldsymbol{P}_1 = -\boldsymbol{c}\boldsymbol{c}^t - (\boldsymbol{P}_1\boldsymbol{b} - \boldsymbol{f})(\boldsymbol{P}_1\boldsymbol{b} - \boldsymbol{f})^t. \tag{14.13}$$

Thus the existence part of the theorem will follow if we can show that $\boldsymbol{P}_1\boldsymbol{b} = \boldsymbol{f}$. Let us show this.

The argument is outlined as follows:

1. multiply (14.13) by -1, add and subtract $\boldsymbol{P}_1 s$, and multiply on the left and right by $\boldsymbol{b}^t(-s\boldsymbol{I}_n - \boldsymbol{A})^{-1}$ and $(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}$, respectively, to get an equation $(*)$;
2. let $Q_1 \in \mathbb{R}[s]$ be defined by
$$\boldsymbol{b}^t\boldsymbol{P}_1(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = \frac{Q_1(s)}{D(s)};$$
3. let $P = Q_1 - \tilde{F} + D$;
4. substitute the above definitions into the equation $(*)$ and the resulting expression turns out to be
$$\frac{P\tilde{F}^*}{DD^*} + \frac{P^*\tilde{F}}{DD^*} = 0;$$
5. as D and $\tilde{F}$ are monic, this reduces to
$$\frac{P}{\tilde{F}} + \frac{P^*}{\tilde{F}^*} = 0; \tag{14.14}$$
6. since $\tilde{F}$ is analytic in $\mathbb{C}_+$, the rational functions $\frac{P}{\tilde{F}}$ and $\frac{P^*}{\tilde{F}^*}$ have no common poles so (14.14) holds if and only if $P = P^* = 0$;
7. by the definition of P, $P(s) = 0$ implies that
$$(\boldsymbol{b}^t\boldsymbol{P}_1 - \boldsymbol{f}^t)(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = 0; \tag{14.15}$$
8. this implies that the components of the vector $\boldsymbol{b}^t\boldsymbol{P}_1 - \boldsymbol{f}^t$ are zero, or that $\boldsymbol{P}_1\boldsymbol{b} = \boldsymbol{f}$, as desired.

Thus we have shown that (14.10) has a solution, and the solution $\boldsymbol{P}_1$ we found was positive-definite. Let us show that this is the only positive-definite solution. Let $\boldsymbol{P}_2$ be positive-definite and suppose that

$$\boldsymbol{A}^t\boldsymbol{P}_2 + \boldsymbol{P}_2\boldsymbol{A} - \boldsymbol{P}_2\boldsymbol{b}\boldsymbol{b}^t\boldsymbol{P}_2 = -\boldsymbol{c}\boldsymbol{c}^t. \tag{14.16}$$

Now argue as follows:

1. multiply (14.16) by -1, add and subtract $\boldsymbol{P}_2 s$, and multiply on the left and right by $\boldsymbol{b}^t(-s\boldsymbol{I}_n - \boldsymbol{A})^{-1}$ and $(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}$, respectively, to get an expression $(**)$;

2. Define $Q_2 \in \mathbb{R}[s]$ by requiring that

$$\boldsymbol{b}^t \boldsymbol{P}_2 (s\boldsymbol{I}_n - \boldsymbol{A})^{-1} \boldsymbol{b} = \frac{Q_2(s)}{D(s)};$$

3. following the arguments in the existence part of the proof, show that $\boldsymbol{P}_2 \boldsymbol{b} = \boldsymbol{f}$;
4. this implies that (14.16) is equivalent to

$$(\boldsymbol{A}^t - \boldsymbol{f}\boldsymbol{b}^t)\boldsymbol{P}_2 + \boldsymbol{P}_2(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t) = -\boldsymbol{c}\boldsymbol{c}^t - \boldsymbol{f}\boldsymbol{f}^t;$$

5. by Theorem 5.32(i), $\boldsymbol{P}_2 = \boldsymbol{P}_1$.

This completes the proof. ■

During the course of the proof of the theorem, we arrived at the relationship between the solution to the algebraic Riccati equation and the optimal state feedback vector $\boldsymbol{f}$. Let us record this.

14.30 Corollary *Consider a controllable SISO linear system $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$. If $\boldsymbol{f}$ is the optimal state feedback vector of Corollary 14.26 and $\boldsymbol{P}$ is the unique positive-definite solution to the algebraic Riccati equation (14.10), then $\boldsymbol{f} = \boldsymbol{P}\boldsymbol{b}$.*

Proof By Corollary 14.26 the vector $\boldsymbol{f} = (f_0, f_1, \ldots, f_{n-1})$ is defined by

$$f_{n-1}s^{n-1} + \cdots + f_1 s + f_0 = [D(s)D(-s) + N(s)N(-s)]^+ - D(s).$$

However, this is exactly the relation (14.11) in the proof of Theorem 14.29, since in the statement of the theorem we had assumed $(\boldsymbol{A}, \boldsymbol{b})$ to be in controller canonical form. During the course of the same proof, the relation $\boldsymbol{f} = \boldsymbol{P}\boldsymbol{b}$ was also shown to be true when $(\boldsymbol{A}, \boldsymbol{b})$ is in controller canonical form. The result now follows from the transformation property for control systems under state similarity stated in Proposition 2.36. ■

14.31 Example **(Example 14.28 cont'd)** We resume with the situation where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

and where we used the two output vectors

$$\boldsymbol{c}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{c}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The two feedback vectors were computed to be

$$\boldsymbol{f}_1 = \begin{bmatrix} \sqrt{2} - 1 \\ 2\sqrt[4]{2} \sin \frac{\pi}{8} \end{bmatrix}, \quad \boldsymbol{f}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

respectively. Let us denote by $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ the two corresponding positive-definite matrices guaranteed by Theorem 14.29. Referring to the equation (14.12) in the proof of Theorem 14.29, we see that the matrices $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ satisfy

$$(\boldsymbol{A}^t - \boldsymbol{f}_j \boldsymbol{b}^t)\boldsymbol{P}_j + \boldsymbol{P}_j(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}_j^t) = -\boldsymbol{c}_j \boldsymbol{c}_j^t - \boldsymbol{f}_j \boldsymbol{f}_j^t, \quad j = 1, 2.$$

Now we refer to the proof of part (i) of Theorem 5.32 to see that

$$\boldsymbol{P}_j = \int_0^\infty e^{\boldsymbol{A}_j^t t} \boldsymbol{Q}_j e^{\boldsymbol{A}_j t} \, \mathrm{d}t, \quad j = 1, 2,$$

where

$$\boldsymbol{A}_j = \boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}_j^t, \quad \boldsymbol{Q}_j = \boldsymbol{c}_j \boldsymbol{c}_j^t + \boldsymbol{f}_j \boldsymbol{f}_j^t, \quad j = 1, 2.$$

One may do the integration (I used Mathematica®, of course) to obtain

$$\boldsymbol{P}_1 = \begin{bmatrix} 2\sqrt[4]{8} \sin \frac{\pi}{8} & \sqrt{2} - 1 \\ \sqrt{2} - 1 & 2\sqrt[4]{2} \sin \frac{\pi}{8} \end{bmatrix}, \quad \boldsymbol{P}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In each case we readily verify that $\boldsymbol{f}_j = \boldsymbol{P}_j \boldsymbol{b}$, $j = 1, 2$, just as predicted by Corollary 14.30. •

14.32 Remarks

1. The algebraic Riccati equation must generally be solved numerically. Indeed, note that Theorem 14.25 provides essentially the same information as the algebraic Riccati equation (as made precise in Corollary 14.30). In the former case, we must find the left half-plane spectral factor of an even polynomial, and this involves finding the roots of this polynomial. This itself is something that typically must be done numerically.
2. Note that the matrix on the right-hand side of the algebraic Riccati equation is the matrix that determines the penalty given to states (as opposed to control) in the cost function of Problems 14.1 and 14.3. One can easily imagine using more general symmetric matrices to define this cost, thus looking at cost functions of the form

$$J_{\boldsymbol{x}_0}(u) = \int_0^\infty \left(\boldsymbol{x}_{u,\boldsymbol{x}_0}^t(t) \boldsymbol{Q} \boldsymbol{x}_{u,\boldsymbol{x}_0}(t) + Ru^2 \right) \mathrm{d}t.$$

   where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is symmetric and positive-semidefinite, and $R > 0$. This can also be seen to be generalisable to multiple inputs by making the cost associated to the input be of the form $\boldsymbol{u}^t(t) \boldsymbol{R} \boldsymbol{u}(t)$ for an $m \times m$ symmetric matrix $\boldsymbol{Q}$. Making the natural extrapolation gives the analogue of equation (14.10) to be

$$\boldsymbol{A}^t \boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} - \boldsymbol{P}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^t\boldsymbol{P} = -\boldsymbol{Q}.$$

   This is indeed the form of the algebraic Riccati equation that gets used in MIMO generalisations of our Theorem 14.25.
3. The optimal feedback vector $\boldsymbol{f}$ determined in this section is often referred to as the ***linear quadratic regulator*** (***LQR***). •

14.3.3 Optimal state estimation results

With the optimal control results of the preceding two sections, and with the (now known) solution of the model matching problem of Section 14.2.6, we can prove the following result which characterises the solution to the optimal state estimation problem.

14.33 Theorem *For $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{D})$ a complete SISO linear system, the following statements regarding $\boldsymbol{\ell} \in \mathbb{R}^n$ are equivalent:*

(i) *$\boldsymbol{\ell}$ is a solution for Problem 14.6;*

(ii) *if (N, D) is the c.f.r. of T_Σ, if $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ is the invertible matrix with the property that $(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}, \boldsymbol{T}^{-t}\boldsymbol{c})$ is in observer canonical form, and if $\tilde{\boldsymbol{\ell}} = (\tilde{\ell}_0, \tilde{\ell}_1, \dots, \tilde{\ell}_{n-1})$ is defined by requiring that*

$$\tilde{\ell}_{n-1}s^{n-1} + \cdots + \tilde{\ell}_1 s + \tilde{\ell}_0 = [D(s)D(-s) + N(s)N(-s)]^+ - D(s),$$

then $\boldsymbol{\ell} = \boldsymbol{T}^{-1}\tilde{\boldsymbol{\ell}}$;

(iii) *$\boldsymbol{\ell} = \boldsymbol{P}\boldsymbol{c}$ where $\boldsymbol{P}$ is the unique positive-definite solution to the algebraic Riccati equation*

$$\boldsymbol{A}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A}^t - \boldsymbol{P}\boldsymbol{c}\boldsymbol{c}^t\boldsymbol{P} = -\boldsymbol{b}\boldsymbol{b}^t.$$

Proof First let us demonstrate the equivalence of (ii) and (iii). Let us define $\tilde{\Sigma} = (\boldsymbol{A}^t, \boldsymbol{c}, \boldsymbol{b}^t, \boldsymbol{D})$, and let $\tilde{\boldsymbol{T}} \in \mathbb{R}^{n\times n}$ be the invertible matrix which puts $(\boldsymbol{A}^t, \boldsymbol{c})$ into controller canonical form. From Corollaries 14.26 and 14.30 we infer the following. If $(\tilde{N}, \tilde{D})$ is the c.f.r. for $T_{\tilde{\Sigma}}$ then if $\tilde{\boldsymbol{\ell}}$ is defined by

$$\tilde{\ell}_{n-1}s^{n-1} + \cdots + \tilde{\ell}_1 s + \tilde{\ell}_0 = [\tilde{D}(s)\tilde{D}(-s) + \tilde{N}(s)\tilde{N}(-s)]^+ - \tilde{D}(s),$$

we have $\boldsymbol{\ell} = \tilde{\boldsymbol{T}}^t\tilde{\boldsymbol{\ell}}$. Now we note that $(\tilde{\boldsymbol{T}}\boldsymbol{A}^t\tilde{\boldsymbol{T}}^{-1}, \tilde{\boldsymbol{T}}\boldsymbol{c})$ is in controller canonical form if and only if $(\tilde{\boldsymbol{T}}^{-t}\boldsymbol{A}\tilde{\boldsymbol{T}}^t, \tilde{\boldsymbol{T}}\boldsymbol{c})$ is in observer canonical form. From this we infer that $\tilde{\boldsymbol{T}} = \boldsymbol{T}^{-t}$. The equivalence of parts (ii) and (iii) now follow from the fact that $(\tilde{N}, \tilde{D}) = (N, D)$.

We now show the equivalence of parts (i) and (iii). First we note that by Parseval's Theorem we have

$$J(\boldsymbol{\ell}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|T_\Sigma(i\omega) - T_{\Sigma_{\boldsymbol{\ell}}}(i\omega)T_\Sigma(i\omega)\right|^2 \mathrm{d}\omega + \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|T_{\Sigma_{\boldsymbol{\ell}}}(i\omega)\right|^2 \mathrm{d}\omega,$$

where $\Sigma_{\boldsymbol{\ell}} = (\boldsymbol{A} - \boldsymbol{\ell}\boldsymbol{c}^t, \boldsymbol{\ell}, \boldsymbol{c}^t, \boldsymbol{0}_1)$. Since the transfer functions are scalar, they may be transposed without changing anything. That is to say, we have

$$\begin{aligned} T_\Sigma(s) &= \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} = \boldsymbol{b}^t(s\boldsymbol{I}_n - \boldsymbol{A}^t)^{-1}\boldsymbol{c}^t \\ T_{\Sigma_{\boldsymbol{\ell}}}(s) &= \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}\boldsymbol{\ell} = \boldsymbol{\ell}^t(s\boldsymbol{I}_n - \boldsymbol{A}^t + \boldsymbol{c}\boldsymbol{\ell}^t)^{-1}\boldsymbol{c}. \end{aligned}$$

We shall think of these transfer functions as being in this form for the moment. Now, thinking of $T_{\Sigma_{\boldsymbol{\ell}}}$ as the unknown, we wish to choose this unknown to minimise $J(\boldsymbol{\ell})$. This, however, is exactly the model matching problem posed in Section 14.2.6. The solution to the model matching problem is given, if we know the solution to the optimal control problem Problem 14.3. But since we now know this, we do indeed know the solution to the model matching problem, and let us translate the solution into our current notation. The transfer function $T_{\Sigma_{\boldsymbol{\ell}}}$ minimising $J(\boldsymbol{\ell})$ is given by $T_{\Sigma_{\boldsymbol{\ell}}} = (-1 + T_{\tilde{\Sigma}_2})T_{\tilde{\Sigma}_1}$ where $\tilde{\Sigma}_1 = (\boldsymbol{A}^t, \boldsymbol{c}, \boldsymbol{\ell}_1^t, \boldsymbol{0}_1)$ and $\tilde{\Sigma}_2 = (\boldsymbol{A}^t - \boldsymbol{c}\boldsymbol{\ell}_2^t, \boldsymbol{c}, \boldsymbol{\ell}_2^t, \boldsymbol{0}_1)$, and where $\boldsymbol{\ell}^t = [\,\ell_1^t\ \ell_1^t\,]$ is given by $\boldsymbol{\ell}^t = \tilde{\boldsymbol{b}}^t\tilde{\boldsymbol{P}}$, where $\tilde{\boldsymbol{P}}$ is the unique positive-definite solution to the algebraic Riccati equation

$$\tilde{\boldsymbol{A}}^t\tilde{\boldsymbol{P}} + \tilde{\boldsymbol{P}}\tilde{\boldsymbol{A}} - \tilde{\boldsymbol{P}}\tilde{\boldsymbol{b}}\tilde{\boldsymbol{b}}^t\tilde{\boldsymbol{P}} = -\tilde{\boldsymbol{c}}\tilde{\boldsymbol{c}}^t, \tag{14.17}$$

with

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A}^t & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}^t \end{bmatrix}, \quad \tilde{\boldsymbol{b}} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{c} \end{bmatrix}, \quad \tilde{\boldsymbol{c}}^t = \begin{bmatrix} \boldsymbol{b}^t & -\boldsymbol{b}^t \end{bmatrix}.$$

Writing

$$\tilde{\boldsymbol{P}} = \begin{bmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{12} & \boldsymbol{P}_{22} \end{bmatrix}$$

and expanding (14.17), we arrive at the following four equations:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{P}_{11} + \boldsymbol{P}_{11}\boldsymbol{A}^t - \boldsymbol{P}_{12}\boldsymbol{c}\boldsymbol{c}^t\boldsymbol{P}_{12} &= -\boldsymbol{b}\boldsymbol{b}^t \\ \boldsymbol{A}\boldsymbol{P}_{12} + \boldsymbol{P}_{12}\boldsymbol{A}^t - \boldsymbol{P}_{12}\boldsymbol{c}\boldsymbol{c}^t\boldsymbol{P}_{22} &= \boldsymbol{b}\boldsymbol{b}^t \\ \boldsymbol{A}\boldsymbol{P}_{12} + \boldsymbol{P}_{12}\boldsymbol{A}^t - \boldsymbol{P}_{22}\boldsymbol{c}\boldsymbol{c}^t\boldsymbol{P}_{12} &= \boldsymbol{b}\boldsymbol{b}^t \\ \boldsymbol{A}\boldsymbol{P}_{22} + \boldsymbol{P}_{22}\boldsymbol{A}^t - \boldsymbol{P}_{22}\boldsymbol{c}\boldsymbol{c}^t\boldsymbol{P}_{22} &= -\boldsymbol{b}\boldsymbol{b}^t. \end{aligned} \tag{14.18}$$

Since $\boldsymbol{P}$ is symmetric and positive-definite, so too is $\boldsymbol{P}_{22}$ (since $\boldsymbol{x}^t\boldsymbol{P}\boldsymbol{x}^t > 0$ for all $\boldsymbol{x}$ of the form $\boldsymbol{x}^t = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{x}_2^t \end{bmatrix}$). The last of equations (14.18) then uniquely prescribes $\boldsymbol{P}_{22}$ as being the matrix prescribed in part (iii) of the theorem. Adding the last two of equations (14.18) gives

$$\boldsymbol{A}(\boldsymbol{P}_{12} + \boldsymbol{P}_{22}) + (\boldsymbol{P}_{12} + \boldsymbol{P}_{22})\boldsymbol{A}^t - \boldsymbol{P}_{22}\boldsymbol{c}\boldsymbol{c}^t(\boldsymbol{P}_{12} + \boldsymbol{P}_{22}) = \boldsymbol{0}.$$

This gives $\boldsymbol{P}_{12} + \boldsymbol{P}_{22} = \boldsymbol{0}$ by . We then have

$$\begin{bmatrix} \boldsymbol{\ell}_1^t & \boldsymbol{\ell}_2^t \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{c}^t \end{bmatrix} \begin{bmatrix} \boldsymbol{P}_{11} & -\boldsymbol{P}_{22} \\ -\boldsymbol{P}_{22} & \boldsymbol{P}_{22} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{c}^t\boldsymbol{P}_{22} & \boldsymbol{c}^t\boldsymbol{P}_{22} \end{bmatrix},$$

and we take $\boldsymbol{\ell} = \boldsymbol{\ell}_2$. Now we obtain

$$\begin{aligned} T_{\tilde{\Sigma}_1}(s) &= -\boldsymbol{\ell}^t(s\boldsymbol{I}_n - \boldsymbol{A}^t)^{-1}\boldsymbol{c} = -\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{\ell} \\ T_{\tilde{\Sigma}_2}(s) &= \boldsymbol{\ell}^t(s\boldsymbol{I}_n - \boldsymbol{A}^t + \boldsymbol{c}\boldsymbol{\ell}^t)^{-1}\boldsymbol{c} = \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}\boldsymbol{\ell}, \end{aligned}$$

thus giving

$$\begin{aligned} (-1 + T_{\tilde{\Sigma}_2}(s))T_{\tilde{\Sigma}_1}(s) &= (1 - \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1}\boldsymbol{\ell})\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{\ell} \\ &= \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}(\boldsymbol{I}_n - \boldsymbol{\ell}\boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)^{-1})\boldsymbol{\ell} \\ &= \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t - \boldsymbol{\ell}\boldsymbol{c}^t)(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)\boldsymbol{\ell} \\ &= \boldsymbol{c}^t(s\boldsymbol{I}_n - \boldsymbol{A} + \boldsymbol{\ell}\boldsymbol{c}^t)\boldsymbol{\ell}, \end{aligned}$$

as desired. ■

14.4 The linear quadratic Gaussian controller

In Section 10.5.3 we showed how one could combine state estimation via a Luenberger observer with static state feedback to obtain a controller that worked by using the estimated states in the state feedback law.

14.4.1 LQR and pole placement

We know that if the state feedback vector $\boldsymbol{f}$ is chosen according to Corollary 14.26, then the matrix $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ will be Hurwitz. However, we have said nothing about the exact nature of the eigenvalues of this matrix. In this section we address this issue.

14.4.2 Frequency domain interpretations

Our above presentations for the linear quadratic regulator and the optimal state estimator are presented in the time-domain. While we present their solutions in terms of spectral factorisation of polynomials, we also see that these solutions are obtainable using the algebraic Riccati equation, and this is how these problems are typically solved in the MIMO case. However, it is also possible to give frequency response formulations and solutions to these problems, and in doing so we make a connection to the H_2 model matching problem of Section 14.2.6.

14.4.3 H_2 model matching and LQG

In this section we show that LQG control may be posed as an H_2 model matching problem. This will provide us with a natural segue to the next chapter where we discuss a more difficult model matching problem, that of H_∞ model matching. By representing the somewhat easily understood LQG control in the context of model matching, we hope to motivate the less easily understood material in the next chapter.

14.5 Stability margins for optimal feedback

In this section we wish to investigate some properties of our optimal feedback law. We shall work in the setting of Corollary 14.26. It is our desire to talk about the gain and phase margin for the closed-loop system in this case. However, it is not quite clear that it makes sense to do this as gain and phase margins are defined in the context of unity gain feedback loops, not in the context of static state feedback. Therefore, the first thing we shall is recall from Exercise E7.11 the connection between static state feedback and unity gain feedback loop concepts. In particular, recall that $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{f}^t$ is Hurwitz if and only if

$$\frac{\boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}}{1 + \boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}} \in \mathrm{RH}_\infty^+.$$

This result tells us that closed-loop stability of the closed-loop system $\Sigma_{\boldsymbol{f}}$ is equivalent to IBIBO stability of a unity gain feedback loop with loop gain $R_L(s) = \boldsymbol{f}^t(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}$. Note that this enables us to employ our machinery for these interconnections, thinking of $\boldsymbol{f}(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}$ as being the loop gain.

14.5.1 Stability margins for LQR

In particular, as is made clear in Exercise E7.11, we may employ the Nyquist criterion to determine closed-loop stability under static state feedback. Recall that rather than the poles of the loop gain in $\mathbb{C}_+$, one uses the eigenvalues of $\boldsymbol{A}$ in $\mathbb{C}_+$ to compare with the encirclements of $-1 + i0$. Let us illustrate the Nyquist criterion on an unstable system.

14.34 Example We take

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & -2 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We ascertain that $\boldsymbol{A}$ has a repeated eigenvalue of $+1$, thus $n_p = 2$. Exercise E7.11(b) indicates that any stabilising state feedback vector $\boldsymbol{f}$ will have the property that the Nyquist plot for the loop gain $R_{\boldsymbol{f}}(s) = \boldsymbol{f}^t(s\boldsymbol{I}_s - \boldsymbol{A})^{-1}\boldsymbol{b}$ will encircle the origin twice in the clockwise direction. Let us test this, not for just any stabilising state feedback vector, but for the

optimal one. Without going through the details as we have done this for one example already, the optimal feedback state vector is

$$\boldsymbol{f} = \begin{bmatrix} 1+\sqrt{2} \\ \sqrt{\frac{7}{2} - \frac{\sqrt{41}}{2}} + \sqrt{\frac{7}{2} + \frac{\sqrt{41}}{2}} - 2 \end{bmatrix}.$$

In Figure 14.2 is shown the Nyquist plot for the loop gain $R_{\boldsymbol{f}}(s) = \boldsymbol{f}^t(s\boldsymbol{I}_2 - \boldsymbol{A})^{-1}\boldsymbol{b}$. One

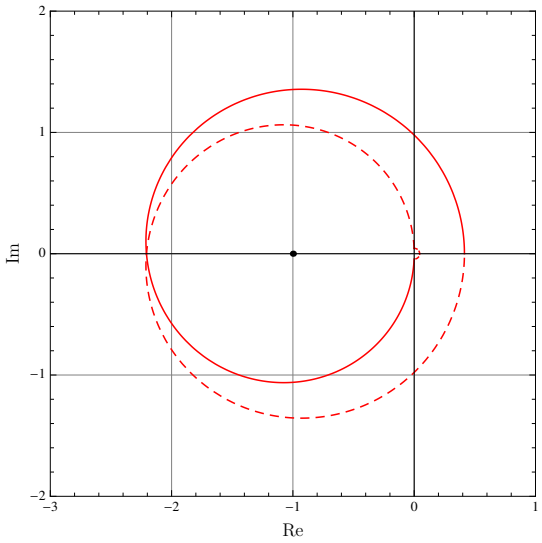


Figure 14.2 Nyquist plot for optimal state feedback with two unstable eigenvalues

can see that it encircles the origin twice in the clockwise direction, as predicted. •

The discussion to this point has largely been with respect to general feedback vectors. However, notice that the Nyquist plot of Example 14.34, done for the optimal state feedback vector of Corollary 14.26, has an interesting feature: the Nyquist contour remains well clear of the critical point $-1 + i0$. The following result tells us that we can generally expect this to happen when we use the optimal state feedback vector.

14.35 Theorem *Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be a controllable SISO linear system and let $\boldsymbol{f} \in \mathbb{R}^n$ be the optimal state feedback vector of Corollary 14.26. We have*

$$|-1 - \boldsymbol{f}(i\omega\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}| \geq 1$$

for every $\omega \in \mathbb{R}$. In other words, the point $R_{\boldsymbol{f}}(i\omega)$ is at least distance 1 away from the point $-1 + i0$.

Proof We assume without loss of generality that $(\boldsymbol{A}, \boldsymbol{b})$ is in controller canonical form. We let $\boldsymbol{P}$ be the unique positive-definite matrix guaranteed by Theorem 14.29. Thus

$$\boldsymbol{A}^t\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} - \boldsymbol{P}\boldsymbol{b}\boldsymbol{b}^t\boldsymbol{P} + \boldsymbol{c}\boldsymbol{c}^t = \boldsymbol{0}_{n\times n}. \tag{14.19}$$

Now perform the following computations:

1. to (14.19), add and subtract $i\omega\boldsymbol{P}$;

2. multiply the resulting equation on the left by $\boldsymbol{b}^t(-i\omega\boldsymbol{I}_n - \boldsymbol{A}^t)^{-1}$ and on the right by $(i\omega\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}$;
3. use the identity $\boldsymbol{f} = \boldsymbol{P}\boldsymbol{b}$ from Corollary 14.30.

The result of these computations is

$$|1 + \boldsymbol{f}^t(i\omega\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}|^2 = 1 + \boldsymbol{b}^t(-i\omega\boldsymbol{I}_n - \boldsymbol{A}^t)^{-1}\boldsymbol{c}\boldsymbol{c}^t(i\omega\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}.$$

Since the matrix $\boldsymbol{c}\boldsymbol{c}^t$ is positive-semidefinite, we have

$$\boldsymbol{b}^t(-i\omega\boldsymbol{I}_n - \boldsymbol{A}^t)^{-1}\boldsymbol{c}\boldsymbol{c}^t(i\omega\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} \geq 0.$$

The result follows. ■

This interesting result tells us that the Nyquist contour remains outside the circle of radius 1 in the complex plane with centre $-1 + i0$. Thus, it remains well clear of the critical point.

14.36 Example **(Example 14.28 cont'd)** We resume looking at the system with

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and where we'd considered the two output vectors

$$\boldsymbol{c}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{c}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

leading to the optimal state feedback vectors

$$\boldsymbol{f}_1 = \begin{bmatrix} \sqrt{2} - 1 \\ 2\sqrt[4]{2}\sin\frac{\pi}{8} \end{bmatrix}, \quad \boldsymbol{f}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

respectively. In Figure 14.3 we give the Nyquist plots for both loop gains $R_{\boldsymbol{f}_1}$ and $R_{\boldsymbol{f}_2}$. As predicted by Theorem 14.35, both Nyquist contours remain outside the circle of radius 1 centred at $-1 + i0$. •

14.5.2 Stability margins for LQG

Results of the nature of Theorem 14.35 *demand* full knowledge of the state. Doyle [1978] shows that as soon as one tries to estimate the state from the output using a Kalman filter, the stability margin of Theorem 14.35 disappears. Although state estimation, and in particular Kalman filtering, is something we do not cover in this book, the reader should be aware of these lurking dangers.

14.6 Summary

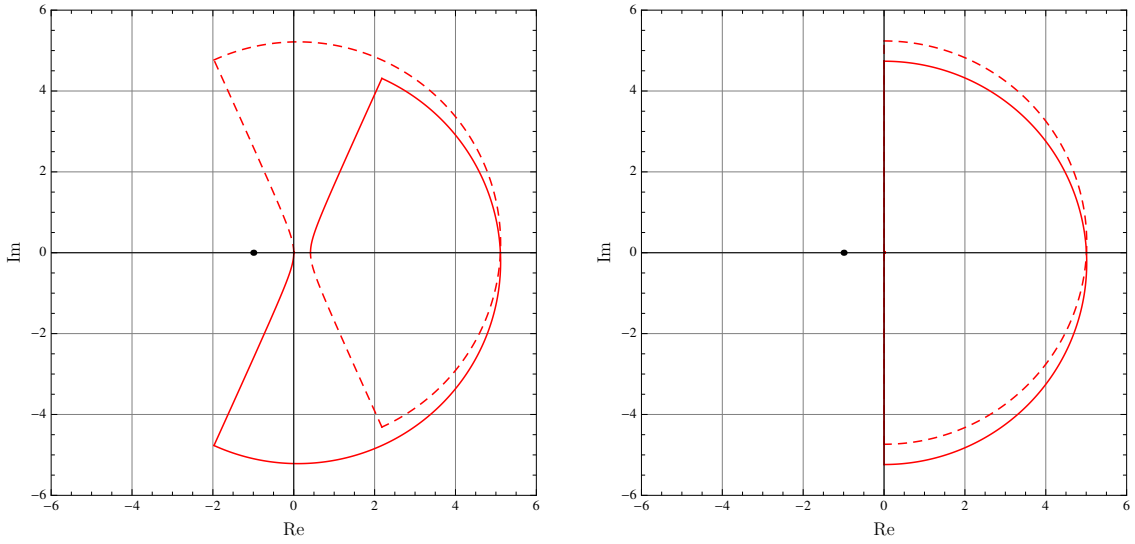


Figure 14.3 Nyquist plots for the optimal state feedback vectors of Example 14.36

Exercises

E14.1 Show that an inner function has a magnitude Bode plot that is a constant 0dB.

E14.2 Let $P, Q \in \mathbb{R}[s]$ have no common roots on the imaginary axis.

(a) Show that the polynomial $PP^* + QQ^*$ is even and positive when evaluated on $i\mathbb{R}$.

(b) Conclude that $PP^* + QQ^*$ admits a spectral factorisation.

(c) Show that $[PP^* + QQ^*]^+$ is Hurwitz (i.e., show that $[PP^* + QQ^*]^+$ has no roots on the imaginary axis).

E14.3 Show that if polynomials $R_1, R_2 \in \mathbb{R}(s)$ admit a spectral factorisation, then so does $R_1 R_2$.

E14.4 Show that if $R \in \mathrm{RH}_\infty^+$, then its outer factor is also a spectral factor.

E14.5 In Theorem 14.25 the optimal control law drives the output to zero, but it is not said what happens to the state. In this exercise, you will redress this problem.

(a) Show that not only does $\lim_{t\to\infty} y(t) = 0$, but that $\lim_{t\to\infty} y^{(k)}(t) = 0$ for all $k > 0$.

(b) Now use the fact that in the statement of Theorem 14.25, Σ is said to be observable to show that $\lim_{t\to\infty} \boldsymbol{x}(t) = \mathbf{0}$.

In Exercise E10.5 you provided a characterisation of a stabilising state feedback vector using a linear matrix inequality (LMI). In the following exercise, you will further characterise the *optimal* state feedback vector using LMI's. The characterisation uses the algebraic Riccati equation of Theorem 14.29.

E14.6

E14.7 For the pendulum on a cart of Exercises E1.5 and E2.4, choose parameter values for the mass of the cart, the mass of the pendulum, the gravitational constant, and the length of the pendulum arm to be $M = 1\frac{1}{2}$, $m = 1$, $g = 9.81$, and $\ell = \frac{1}{2}$. For each of the following linearisations:

(a) the equilibrium point $(0, \pi)$ with cart position as output;
(b) the equilibrium point $(0, \pi)$ with cart velocity as output;
(c) the equilibrium point $(0, \pi)$ with pendulum angle as output;
(d) the equilibrium point $(0, \pi)$ with pendulum angular velocity as output,

do the following:

1. construct the optimal state feedback vector of Problem 14.3;
2. compute the closed-loop eigenvalues;
3. plot a few trajectories of the full system with the control determined by the state feedback vector of part (1).

E14.8 For the double pendulum system of Exercises E1.6 and E2.5, choose parameter values for the first link mass, the second link mass, the first link length, and the second link length to be $m_1 = 1$, $m_2 = 2$, $\ell_1 = \frac{1}{2}$, and $\ell_2 = \frac{1}{3}$. For each of the following linearisations:

(a) the equilibrium point $(0, \pi, 0, 0)$ with the pendubot input;
(b) the equilibrium point $(\pi, 0, 0, 0)$ with the pendubot input;
(c) the equilibrium point $(\pi, \pi, 0, 0)$ with the pendubot input;
(d) the equilibrium point $(0, \pi, 0, 0)$ with the acrobot input;
(e) the equilibrium point $(\pi, 0, 0, 0)$ with the acrobot input;
(f) the equilibrium point $(\pi, \pi, 0, 0)$ with the acrobot input,

do the following:

1. construct the optimal state feedback vector of Problem 14.3;
2. compute the closed-loop eigenvalues;
3. plot a few trajectories of the full system with the control determined by the state feedback vector of part (1).

E14.9 Consider the coupled tank system of Exercises E1.11, E2.6, and E3.17. Choose the parameters $\alpha = \frac{1}{3}$, $\delta_1 = 1$, $A_1 = 1$, $A_2 = \frac{1}{2}$, $a_1 = \frac{1}{10}$, $a_2 = \frac{1}{20}$, and $g = 9.81$. For the linearisations in the following cases:

(a) the output is the level in tank 1;
(b) the output is the level in tank 2;
(c) the output is the difference in the levels,

do the following:

1. construct the optimal state feedback vector of Problem 14.3;
2. compute the closed-loop eigenvalues;
3. plot a few trajectories of the full system with the control determined by the state feedback vector of part (1).

E14.10 Let $\Sigma = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be controllable and suppose that $\boldsymbol{A}$ is Hurwitz. Let $\boldsymbol{f}$ be the optimal state feedback vector of Corollary 14.26 with $R_{\boldsymbol{f}}(s) = \boldsymbol{f}^t(s\boldsymbol{A}_n - \boldsymbol{A})^{-1}\boldsymbol{b}$ the corresponding loop gain. Show that the gain margin for the resulting Nyquist plot is infinite, and that the phase margin exceeds $60°$.

Chapter 15

An introduction to H_∞ control theory

The task of designing a controller which accomplishes a specified task is a challenging chore. When one wishes to add "robustness" to the mix, things become even more challenging. By robust design, what is meant is that one should design a controller which works not only for a given model, but for plants which are close to that model. In this way, one can have some degree of certainty that even if the model is imperfect, the controller will behave in a satisfactory manner. The development of systematic design procedures for robust control can be seen to have been initiated with the important paper of Francis and Zames [1984]. Since this time, there have been many developments and generalisations. The understanding of so-called H_∞ methods has progressed to the point that a somewhat elementary treatment is possible. We shall essentially follow the approach of Doyle, Francis, and Tannenbaum [1990]. For a recent account of MIMO developments, see [Dullerud and Paganini 1999]. The book by Francis [1987] is also a useful reference.

Although all of the material in this chapter can be followed by any student who has come to grips with the more basic material in this book, much of what we do here is a significant diversion from control theory, per se, and is really a development of the necessary mathematical tools. Since a complete understanding of the tools is not necessary in order to apply them, it is perhaps worthwhile to outline the bare bones guide to getting through this chapter. This might be as follows.

1. One should first make sure that one understands the problem being solved: the robust performance problem. The first thing done is to modify this problem so that it is tractable; this is the content of Problem 15.2.
2. The modified robust performance problem is first converted to a model matching problem, which is stated in generality in Problem 15.3. The content of this conversion is contained in Algorithm 15.5.
3. The model matching problem is solved in this book in two ways. The first method involves Nevanlinna-Pick interpolation, and to apply this method, follow the steps outlined in Algorithm 15.18. It is entirely possible that you will have to make some modifications to the algorithm to ensure that you arrive at a proper controller. The necessary machinations are discussed following the algorithm.
4. The second method for solving the model matching problem involves approximation of unstable rational functions by stable ones via Nehari's Theorem. To apply this method, follow the steps in Algorithm 15.29. As with Nevanlinna-Pick interpolation, one should be prepared to make some modifications to the algorithm to ensure that things work out. The manner in which to carry out these modifications is discussed following the algorithm.
5. Other problems that can be solved in this manner are discussed in Section 15.5.

Readers wishing only to be able to apply the methods in this chapter should go through the chapter with the above skeleton as a guide. However, those wishing to see the details will be happy to see very little omitted. It should also be emphasised that Mathematica® and Maple® packages for solving these problems may be found at the URL `http://mast.queensu.ca/~math332/`.[1]

Contents

15.1 Reduction of robust performance problem to model matching problem

For SISO systems, the problem of designing for robust performance turns out to be reducible in a certain sense to two problems investigated by mathematicians coming under the broad heading of complex function approximation, or, more descriptively, model matching. In this section we shall discuss how this reduction takes place, as in itself it is not an entirely obvious step.

15.1.1 A modified robust performance problem

First recall the robust performance problem, Problem 9.23. Given a proper nominal plant $\bar{R}_P$, a function $W_u \in \mathrm{RH}_\infty^+$ given an uncertainty set $\mathscr{P}_\times(\bar{R}_P, W_u)$ or $\mathscr{P}_+(\bar{R}_P, W_u)$, and a performance weight $W_p \in \mathbb{R}(s)$, we seek a controller R_C that stabilises the nominal plant and satisfies either

$$\big\||W_u\bar{T}_L| + |W_p\bar{S}_L|\big\|_\infty < 1 \quad \text{or} \quad \big\||W_uR_C\bar{S}_L| + |W_p\bar{S}_L|\big\|_\infty < 1,$$

depending on whether one is using multiplicative or additive uncertainty. The robust performance problem, it turns out, is quite difficult. For instance, useful necessary and sufficient conditions for there to exist a solution to the problem are not known. Thus our first step in this section is to come up with a simpler problem that is easier to solve. The simpler problem is based upon the following result.

[1] These are not currently implemented. Hopefully they will be in the near future.

15.1 Lemma *Let $R_1, R_2 \in \mathbb{R}(s)$ and denote by $|R_1| + |R_2|$ the $\mathbb{R}$-valued function $s \mapsto (|R_1(s)| + |R_2(s)|)$ and by $|R_1|^2 + |R_2|^2$ the $\mathbb{R}$-valued function $s \mapsto (|R_1(s)|^2 + |R_2(s)|^2)$. If*

$$\big\| |R_1|^2 + |R_2|^2 \big\|_\infty < \frac{1}{2}$$

then

$$\big\| |R_1| + |R_2| \big\|_\infty < 1.$$

Proof Define

$$S_1 = \{(x, y) \in \mathbb{R}^2 \mid \ x, y > 0, \ |x + y| < 1\},$$
$$S_2 = \big\{(x, y) \in \mathbb{R}^2 \ \big| \ x, y > 0, \ |x^2 + y^2| < \tfrac{1}{2}\big\}.$$

The result will follow if we can show that $S_1 \subset S_2$. However, we note that $S_1 = [0, 1] \times [0, 1]$ and S_2 is the circle of radius $\frac{1}{\sqrt{2}}$ centred at the origin. Clearly $S_1 \subset S_2$ (see Figure 15.1). ■

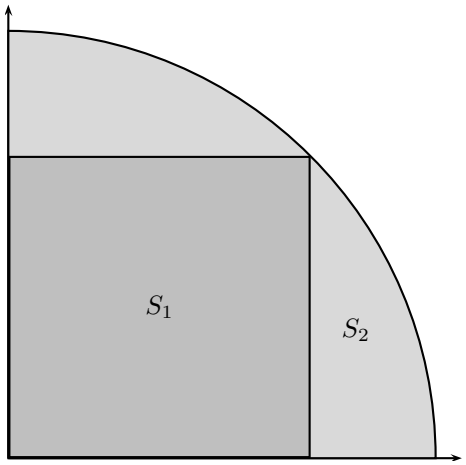


Figure 15.1 Interpretation of modified robust performance condition

This leads to a modification of the robust performance problem, and we state this formally since it is this problem to which we devote the majority of our effort in this chapter. Note that we make a few additional assumptions in the statement of the problem that are not present in the statement of Problem 9.23. Namely, we assume now the following.

1. $W_p \in \mathrm{RH}_\infty^+$: Thus we add the assumption that W_p be proper since, without loss of generality as we are only interested on the value of W_p on $i\mathbb{R}$, we can suppose that all poles of W_p lie in $\mathbb{C}_-$.
2. W_u and W_p have no common imaginary axis zeros: This is an assumption that, if not satisfied, can be satisfied with minor tweaking of W_u and W_p.
3. R_C is proper: We make this assumption mostly as a matter of convention. If we arrive at a controller that is improper but solves the problem, then it is often possible to modify the controller so that it is proper and still solves the problem.

Thus our problem becomes.

15.2 Modified robust performance problem Given

(i) a nominal proper plant $\bar{R}_P$,
(ii) a function $W_u \in \mathrm{RH}_\infty^+$,
(iii) an uncertainty model $\mathscr{P}_\times(\bar{R}_P, W_u)$ or $\mathscr{P}_+(\bar{R}_P, W_u)$, and
(iv) a performance weight $W_p \in \mathrm{RH}_\infty^+$,

so that W_u and W_p have no common imaginary axis zeros, find a proper controller R_C that

(v) stabilises the nominal system and
(vi) satisfies either $\left\| |W_u \bar{T}_L|^2 + |W_p \bar{S}_L|^2 \right\|_\infty < 1$ or $\left\| |W_u R_C \bar{S}_L|^2 + |W_p \bar{S}_L|^2 \right\|_\infty < 1$, depending on whether one is using multiplicative or additive uncertainty. •

As should be clear from Figure 15.1, it is possible that for a given $\bar{R}_P$, W_u, and W_p it will not be possible to solve the modified robust performance problem even though a solution may exist to the robust performance problem. Thus we are sacrificing something in so modifying the problem, but what we gain is a simplified problem that can be solved.

15.1.2 Algorithm for reduction to model matching problem

The objective of this section is to convert the modified robust performance problem into the model matching problem. We shall concentrate in this section on multiplicative uncertainty, with the reader filling in the details for additive uncertainty in Exercise E15.3.

First let us state the model matching problem.

15.3 A model matching problem Let $T_1, T_2 \in \mathrm{RH}_\infty^+$. Find $\theta \in \mathrm{RH}_\infty^+(s)$ so that $\|T_1 - \theta T_2\|_\infty$ is minimised. •

The model matching problem may not have a solution. In fact, it will often be the case in applications that it does not have a solution. However, as we shall see as we get into our development, even when the problem has no solution, it can be used as a guide to solve the problem that is actually of interest to us, namely the modified robust performance problem, Problem 15.2. Some issues concerning existence of solutions to the model matching problem are the topic of Exercise E15.2

15.4 Remark Note that since $T_1, T_2 \in \mathrm{RH}_\infty^+$, if θ is to be a solution of the model matching problem, then it can have no imaginary axis poles. Also, since the model matching problem only cares about the value of θ on the imaginary axis, we can without loss of generality (by multiplying a solution θ to the model matching problem by an inner function that cancels all poles in θ in $\mathbb{C}_+$) suppose that θ has no poles in $\overline{\mathbb{C}}_+$. •

Let us outline the steps in performing the reduction of Problem 15.2 to Problem 15.3. After we have said how to perform the reduction, we will actually prove that everything works. The reader will wish to recall the notion of spectral factorisation for rational functions (Proposition 14.17) and the notion of a coprime factorisation for a pair of rational functions (Theorem 10.33).

15.5 Algorithm for obtaining model matching problem for multiplicative uncertainty Given $\bar{R}_P$, W_u, and W_p as in Problem 15.2.

1. Define
$$U_3 = \frac{W_p W_p^* W_u W_u^*}{W_p W_p^* + W_u W_u^*}.$$
2. If $\|U_3\|_\infty \geq \frac{1}{2}$, then Problem 15.2 has no solution.

3. Let (P_1, P_2) be a coprime fractional representative for $\bar{R}_P$.
4. Let (ρ_1, ρ_2) be a coprime factorisation for P_1 and P_2:
$$\rho_1 P_1 + \rho_2 P_2 = 1.$$
5. Define
$$\begin{aligned} R_1 &= W_p \rho_2 P_2, & S_1 &= W_u \rho_1 P_1, \\ R_2 &= W_p P_1 P_2, & S_2 &= -W_u P_1 P_2. \end{aligned}$$
6. Define $Q = [R_2 R_2^* + S_2 S_2^*]^+$.
7. Let V be an inner function with the property that
$$\frac{R_1 R_2^* + S_1 S_2^*}{Q^*} V$$
has no poles in $\overline{\mathbb{C}}_+$.
8. Define
$$U_1 = \frac{R_1 R_2^* + S_1 S_2^*}{Q^*} V, \qquad U_2 = QV.$$
9. Define $U_4 = [\frac{1}{2} - U_3]^+$.
10. Define
$$T_1 = \frac{U_1}{U_4}, \quad T_2 = \frac{U_2}{U_4}.$$
11. Let θ be a solution to Problem 15.3, and by Remark 15.4 suppose that it has no poles in $\overline{\mathbb{C}}_+$.
12. If $\|T_1 - \theta T_2\|_\infty \geq 1$ then Problem 15.2 has no solution.
13. The controller
$$R_C = \frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1},$$
is a solution to Problem 15.2. •

The above procedure provides a way to produce a controller satisfying the modified robust performance problem, provided one can find θ in Step 11. That is to say, we have reduced the finding of a solution to the modified robust performance problem to that of solving the model matching problem. It remains to show that all constructions made in Algorithm 15.5 are sensible, and that all claims made are true. In the next section we will do this formally. However, before we get into all the details, it is helpful to give a glimpse into how Algorithm 15.5 comes about.

As we are working with multiplicative uncertainty (see Exercise E15.3 for additive uncertainty), the problem we start out with, of course, is to find a proper $R_C \in \mathbb{R}(s)$ that satisfies
$$\left\| |W_u \bar{T}_L|^2 + |W_p \bar{S}_L|^2 \right\|_\infty < \frac{1}{2}$$
for a given nominal proper plant $\bar{R}_P$, an uncertainty model $W_u \in \mathrm{RH}_\infty^+$, and a performance weight $W_p \in \mathrm{RH}_\infty^+$. We choose a coprime fractional representative (P_1, P_2) for $\bar{R}_P$ and an associated coprime factorisation (ρ_1, ρ_2). By Theorem 10.37, any proper stabilising controller is then of the form
$$R_C = \frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1}.$$

A simple computation then gives

$$\bar{T}_L = P_1P_2\theta + \rho_1P_1, \quad \bar{S}_L = -(P_1P_2\theta - \rho_2P_2).$$

Defining

$$\begin{aligned} R_1 &= W_p\rho_2P_2, & S_1 &= W_u\rho_1P_1, \\ R_2 &= W_pP_1P_2, & S_2 &= -W_uP_1P_2, \end{aligned}$$

we then obtain

$$\big\||W_u\bar{T}_L|^2 + |W_p\bar{S}_L|^2\big\|_\infty = \big\||R_1 - \theta R_2|^2 + |S_1 - \theta S_2|^2\big\|_\infty.$$

Up to this point, everything is simple enough. Now we claim that there exists functions $U_1, U_2 \in \mathrm{RH}_\infty^+$ and $U_3 \in \mathbb{R}(s)$, defined in terms of R_1, R_2, S_1, and S_2, and having the property that

$$\big\||R_1 - \theta R_2|^2 + |S_1 - \theta S_2|^2\big\|_\infty = \big\||U_1 - \theta U_2|^2 + U_3\big\|_\infty. \tag{15.1}$$

That these functions exist, and are as stated in Steps 1 and 8 of Algorithm 15.5, will be proved in the subsequent section. Finally, with U_4 as defined in Step 9, in the next section we shall show that

$$\big\||U_1 - \theta U_2|^2 + U_3\big\|_\infty < \frac{1}{2} \quad \Longleftrightarrow \quad \|U_4^{-1}U_1 - \theta U_4^{-1}U_2\|_\infty < 1.$$

With this rough justification behind us, let us turn to formal proofs of the validity of Algorithm 15.5. Readers not interested in this sort of detail can actually skip to Section 15.4.

15.1.3 Proof that reduction procedure works

Throughout this section, we let $\bar{R}_P$, W_p, and W_p are as stated in Problem 15.2.

In Step 6 of Algorithm 15.5, we are asked to compute the spectral factorisation of $R_2R_2^* + S_2S_2^*$. Let us verify that this spectral factorisation exists.

15.6 Lemma *$R_2R_2^* + S_2S_2^*$ admits a spectral factorisation.*

Proof We have

$$R_2R_2^* + S_2S_2^* = P_1P_1^*P_2P_2^*(W_pW_p^* + W_uW_u^*).$$

Since $P_1, P_2 \in \mathrm{RH}_\infty^+$ we may find an inner-outer factorisation

$$P_1 = P_{1,\mathrm{in}}P_{1,\mathrm{out}}, \quad P_2 = P_{2,\mathrm{in}}P_{2,\mathrm{out}}$$

by Proposition 14.10. Therefore

$$P_1P_1^* = P_{1,\mathrm{in}}P_{1,\mathrm{out}}P_{1,\mathrm{in}}^*P_{1,\mathrm{out}}^* = P_{1,\mathrm{out}}P_{1,\mathrm{out}}^*,$$

since $P_{1,\mathrm{in}}$ is inner. Since $P_{1,\mathrm{out}}$ is outer, it follows that $P_{1,\mathrm{out}}$ is a left spectral factor for $P_1P_1^*$. Similarly $P_2P_2^*$ admits a spectral factorisation by the outer factor $P_{2,\mathrm{out}}$ of P_2. Thus $P_1P_1^*$ and $P_2P_2^*$ admit a spectral factorisation, and so too then does $P_1P_1^*P_2P_2^*$. Now let (N_p, D_p) and (N_u, D_u) be the c.f.r.'s of W_p and W_u. We then have

$$W_pW_p^* + W_uW_u^* = \frac{N_pN_p^*D_uD_u^* + N_uN_u^*D_pD_p^*}{D_pD_p^*D_uD_u^*}.$$

Since $W_p, W_u \in \mathrm{RH}_\infty^+$ by hypothesis, D_p and D_u, and therefore D_p^* and D_u^*, have no imaginary axis roots. Also by assumption, N_p and N_u have no common roots on $i\mathbb{R}$. One can then show (along the lines of Exercise E14.2) that this infers that $N_pN_p^*D_uD_u^* + N_uN_u^*D_pD_p^*$ has constant sign on $i\mathbb{R}$. Since it is clearly even, one may infer from Proposition 14.12 that $N_pN_p^*D_uD_u^* + N_uN_u^*D_pD_p^*$ admits a spectral factorisation. Therefore, by Proposition 14.17, so too does $W_pW_p^* + W_uW_u^*$. Finally, by Exercise E14.3 we conclude that $R_2R_2^* + S_2S_2^*$ admits a spectral factorisation. ■

Our next result declares that U_1, U_2, and U_3 are as they should be, meaning that they satisfy the relation (15.1).

15.7 Lemma *If U_1, U_2, and U_3 as defined in Steps 8 and 1 satisfy*

$$\big\| |R_1 - \theta R_2|^2 + |S_1 - \theta S_2|^2 \big\|_\infty = \big\| |U_1 - \theta U_2|^2 + U_3 \big\|_\infty.$$

Proof It is sufficient that the relation

$$\big(R_1 - \theta R_2\big)\big(R_1^* - \theta^* R_2^*\big) + \big(S_1 - \theta S_2\big)\big(S_1^* - \theta^* S_2^*\big) = \big(U_1 - \theta U_2\big)\big(U_1^* - \theta^* U_2^*\big) + U_3 \quad (15.2)$$

holds for all θ and for $s = i\omega$. Doing the manipulation shows that if the three relations

$$\begin{aligned} \big(R_2R_2^* + S_2S_2^*\big) &= U_2U_2^* \\ R_1R_2^* + S_1S_2^* &= U_1U_2^* \\ R_1R_1^* + S_1S_1^* &= U_1U_1^* + U_3. \end{aligned} \quad (15.3)$$

hold for $s = i\omega$, then (15.2) will indeed hold for all θ. We let $Q = [R_2R_2^* + S_2S_2^*]^+$. If V is an inner function with the property that

$$\frac{R_1R_2^* + S_1S_2^*}{Q^*}V$$

has no poles in $\overline{\mathbb{C}}_+$, then if $U_2 = QV$ we have

$$U_2U_2^* = QVQ^*V^* = QQ^* = R_2R_2^* + S_2S_2^*.$$

Thus U_2 satisfies the first of equations (15.3). Also,

$$U_1 = \frac{R_1R_2^* + S_1S_2^*}{Q^*}V$$

clearly satisfies the second of equations (15.3). To verify the last of equations (15.3), we may directly compute

$$R_1R_1^* + S_1S_1^* - U_1U_1^* = R_1R_1^* + S_1S_1^* - \frac{(R_1^*R_2 + S_1^*S_2)(R_1R_2^* + S_1S_2^*)}{R_2R_2^* + S_2S_2^*}, \quad (15.4)$$

using our solution for U_1 and the fact that V is inner. A straightforward substitution of the definitions of R_1, R_2, S_1, and S_2 now gives U_3 as in Step 1. ■

Our next lemma verifies Step 2 of Algorithm 15.5.

15.8 Lemma *If* $\|U_3\|_\infty \geq \frac{1}{2}$ *then Problem 15.2 has no solution.*

Proof One may verify by direct computation that

$$U_3 = R_1 R_1^* + S_1 S_1^* - \frac{(R_1^* R_2 + S_1^* S_2)(R_1 R_2^* + S_1 S_2^*)}{R_2 R_2^* + S_2 S_2^*}$$

(this follows from (15.4)). Now work backwards through the proof of Lemma 15.7 to see that

$$U_3 = \big(R_1 - \theta R_2\big)\big(R_1^* - \theta^* R_2^*\big) + \big(S_1 - \theta S_2\big)\big(S_1^* - \theta^* S_2^*\big) - \big(U_1 - \theta U_2\big)\big(U_1^* - \theta^* U_2^*\big)$$

for any admissible $\theta \in \mathrm{RH}_\infty^+$. Therefore, if $\|U_3\|_\infty \geq \frac{1}{2}$ then

$$\big\||R_1 - \theta R_2|^2 + |S_1 - \theta S_2|^2\big\|_\infty \geq \frac{1}{2}.$$

However, a simple working through of the definitions of R_1, R_2, S_1, and S_2 shows that this implies that for any stabilising controller R_C we must have

$$\big\||W_u \bar{T}_L|^2 + |W_p \bar{S}_L|^2\big\|_\infty \geq \frac{1}{2},$$

as desired. ■

Next we show that with T_1 and T_2 as defined in Step 10, the modified robust performance problem is indeed equivalent to the model matching problem.

15.9 Lemma *With* T_1 *and* T_2 *as defined in Step 10 we have*

$$\big\||W_u \bar{T}_L|^2 + |W_p \bar{S}_L|^2\big\|_\infty < \frac{1}{2} \quad \Longleftrightarrow \quad \|T_1 - \theta T_2\|,$$

where

$$R_C = \frac{\rho_2 + \theta P_1}{\rho_1 - \theta P_2}.$$

Proof As outlined at the end Section 15.1.2, the condition

$$\big\||W_u \bar{T}_L|^2 + |W_p \bar{S}_L|^2\big\|_\infty < \frac{1}{2}$$

is equivalent to

$$\big\||U_1 - \theta U_2|^2 + U_3\big\|_\infty < \frac{1}{2}.$$

Also note that by definition, $U_3(i\omega) \geq 0$ for all $\omega \in \mathbb{R}$, and that $U_3 = U_3^*$, the latter fact implying that U_3 is even. Therefore, provided that $\|U_3\|_\infty < \frac{1}{2}$, $\frac{1}{2} - U_3$ admits a spectral factorisation. Now we compute

$$\begin{aligned}
& \big\||U_1 - \theta U_2|^2 + U_3\big\|_\infty < \frac{1}{2} \\
\Longleftrightarrow \quad & \big||U_1(i\omega) - \theta(i\omega) U_2(i\omega)|^2 + U_3(i\omega)\big| < \frac{1}{2}, \quad \omega \in \mathbb{R} \\
\Longleftrightarrow \quad & |U_1(i\omega) - \theta(i\omega) U_2(i\omega)|^2 + U_3(i\omega) < \frac{1}{2}, \quad \omega \in \mathbb{R} \\
\Longleftrightarrow \quad & |U_1(i\omega) - \theta(i\omega) U_2(i\omega)|^2 < \frac{1}{2} - U_3(i\omega), \quad \omega \in \mathbb{R} \\
\Longleftrightarrow \quad & |U_1(i\omega) - \theta(i\omega) U_2(i\omega)|^2 < [\tfrac{1}{2} - U_3(i\omega)]^+ [\tfrac{1}{2} - U_3(-i\omega)], \quad \omega \in \mathbb{R} \\
\Longleftrightarrow \quad & |U_1(i\omega) - \theta(i\omega) U_2(i\omega)|^2 < |U_4(i\omega)|^2, \quad \omega \in \mathbb{R} \\
\Longleftrightarrow \quad & |U_4^{-1}(i\omega) U_1(i\omega) - \theta(i\omega) U_4^{-1}(i\omega) U_2(i\omega)|^2 < 1, \quad \omega \in \mathbb{R}.
\end{aligned}$$

By definition of T_1 and T_2, the lemma follows. ■

It is necessary that $T_1, T_2 \in \mathrm{RH}_\infty^+$ in order to fit them into the model matching problem. The next lemma ensures that this follows from the constructions we have made.

15.10 Lemma $T_1, T_2 \in \mathrm{RH}_\infty^+$.

Proof First note that $\|U_3\| < \frac{1}{2}$ by the time we have gotten to defining T_1 and T_2. Therefore $U_4 = [\frac{1}{2} - U_3]^+$ is strictly proper, and so is invertible in RH_∞^+. The lemma will then follow if we can show that $U_1, U_2 \in \mathrm{RH}_\infty^+$.

First let us show that $U_2 \in \mathrm{RH}_\infty^+$. By definition, U_2 is the left half-plane spectral factor of

$$P_1 P_1^* P_2 P_2^* (W_p W_p^* + W_u W_u^*).$$

As such, it is the product of the two quantities

$$[P_1 P_1^* P_2 P_2^*]^+, \quad [W_p W_p^* + W_u W_u^*]^+.$$

Since each of $P_1 P_2 \in \mathrm{RH}_\infty^+$, $[P_1 P_1^* P_2 P_2^*]^+ \in \mathrm{RH}_\infty^+$. Since

$$W_p W_p^* + W_u W_u^* = \frac{N_p N_p^* D_u D_u^* + N_u N_u^* D_p D_p^*}{D_p D_p^* D_u D_u^*},$$

we have

$$[W_p W_p^* + W_u W_u^*]^+ = \frac{[N_p N_p^* D_u D_u^* + N_u N_u^* D_p D_p^*]^+}{D_p D_u}.$$

Since $W_p, W_u \in \mathrm{RH}_\infty^+$, it follows that $[W_p W_p^* + W_u W_u^*]^+ \in \mathrm{RH}_\infty^+$. Thus $U_2 \in \mathrm{RH}_\infty^+$.

Now let us show that $U_1 \in \mathrm{RH}_\infty^+$. A computation shows that

$$U_1 = \frac{P_1^* P_2^*}{[P_1 P_2^* P_2 P_2^*]^-} \frac{\rho_2 P_2 W_p W_p^* - \rho_1 P_1 W_u W_u^*}{[W_p W_p^* + W_u W_u^*]^-} V.$$

The inner function V is designed so that this function has no poles in $\mathbb{C}_+$. We also claim that U_1 has no poles on $i\mathbb{R}$. Since $W_p, W_u \in \mathrm{RH}_\infty^+$ and since they have no common imaginary axis zeros, it follows that $[W_p W_p^* + W_u W_u^*]^-$ has no zeros on $i\mathbb{R}$. Clearly, neither $P_1^* P_2^*$ nor $\rho_2 P_2 W_p W_p^* - \rho_1 P_1 W_u W_u^*$ have poles on $i\mathbb{R}$. Thus, our claim will follow if we can show that $\frac{P_1^* P_2^*}{[P_1 P_2^* P_2 P_2^*]^-}$ has no poles on $i\mathbb{R}$. This is true since the imaginary axis zeros of $P_1^* P_2^*$ and $[P_1 P_2^* P_2 P_2^*]^-$ agree in location and multiplicity. Thus we have shown that U_1 is analytic in $\overline{\mathbb{C}}_+$. That $U_1 \in \mathrm{RH}_\infty^+$ will now follow if we can show that U_1 is proper. ■ finish

15.2 Optimal model matching I. Nevanlinna-Pick theory

The first solution we shall give to the model matching problem comes from a seemingly unrelated interpolation problem. To state the problem, we need a little notation. We let $\mathrm{RH}_\infty^{+,\mathbb{C}}$ be the collection of proper functions in $\mathbb{C}(s)$ with no poles in $\overline{\mathbb{C}}_+$. Thus $\mathrm{RH}_\infty^{+,\mathbb{C}}$ is just like RH_∞^+, except that now we allow the functions to have complex coefficients. Note that $\|\cdot\|_\infty$ still makes sense for functions in $\mathrm{RH}_\infty^{+,\mathbb{C}}$. Now the interpolation problem is as follows.

15.11 Nevanlinna-Pick interpolation problem Let $\{a_1, \dots, a_k\} \subset \mathbb{C}_+$ and let $\{b_1, \dots, b_k\} \subset \mathbb{C}$ collections of distinct points. A ***Pick pair*** is then a pair (a_j, b_j), $i \in \{1, \dots, k\}$. Suppose that if (a_k, b_k) is a Pick pair, then so is $(\bar{a}_j, \bar{b}_j)$, $j = 1, \dots, k$.

Find $R \in \mathrm{RH}_\infty^+$ so that

(i) $\|R\|_\infty \leq 1$ and
(ii) $R(a_j) = b_j$, $j = 1, \dots, n$. •

This problem was originally solved by Pick [1916], and was solved independently by Nevanlinna [1919]. Nevanlinna also gave an algorithm for finding a solution to the interpolation problem [Nevanlinna 1929]. In this section, we will state and prove Pick's necessary and sufficient condition for the solution of the interpolation problem, and also give an algorithm for determining a solution.

15.12 Remark We should say that the problem solved by Pick is somewhat different than the one we state here. The difference occurs in three ways.

1. Pick actually allowed $\|R\|_\infty = 1$. However, our purposes will require that the H_∞-norm of R be strictly less than 1.
2. Pick was not interested in making the restriction that if a every Pick pair have the property that its complex conjugate also be a Pick pair.
3. Pick was interested in the case where the points $a_1, \dots, a_n$ lie in the open disk $D(0, 1)$ of radius 1 centred at 0. This is not a genuine distinction, however, as the map $s \mapsto \frac{1-s}{1+s}$ bijectively maps $\mathbb{C}_+$ onto $D(0, 1)$, and so translates the domain of concern for Pick to our domain.
4. Finally, Pick allowed interpolating functions to be general meromorphic functions, bounded and analytic in $\overline{\mathbb{C}}_+$. For obvious reasons, our interest is in the subset of such functions that are in $\mathbb{R}(s)$, i.e., functions in RH_∞^+. •

15.2.1 Pick's theorem

Pick's conditions for the existence of a solution to Problem 15.11 are quite simple. The proof that these conditions are necessary and sufficient is not entirely straightforward. In this section we only prove necessity, as sufficiency will follow from our algorithm for solving the Nevanlinna-Pick interpolation problem in the ensuing section. Our necessity proof follows [Doyle, Francis, and Tannenbaum 1990]. For the statement of Pick's theorem, recall that $\boldsymbol{M} \in \mathbb{C}^{n\times n}$ is ***Hermitian*** if $\boldsymbol{M} = \boldsymbol{M}^* = \overline{\boldsymbol{M}}^t$. A Hermitian matrix is readily verified to have real eigenvalues. Therefore, the notions of definiteness presented in Section 5.4.1 may be applied to Hermitian matrices.

15.13 Theorem **(Pick's Theorem)** *Problem 15.11 has a solution if and only if the* ***Pick matrix****, the complex $k \times k$ symmetric matrix $\boldsymbol{M}$ with components*

$$M_{j\ell} = \frac{1 - b_j\bar{b}_\ell}{a_j + \bar{a}_\ell}, \quad j, \ell = 1, \dots, k,$$

is positive-semidefinite.

Proof of necessity Suppose that Problem 15.11 has a solution R. For $c_1, \dots, c_k \in \mathbb{C}$, not all zero, consider the complex input $u\colon (-\infty, 0] \to \mathbb{C}$ given by

$$u(t) = \sum_{j=1}^{k} c_j e^{a_j t}.$$

This can be considered as an input to the transfer function R, with the output computed by separately computing the real and imaginary parts. Moreover, if h_R denotes the inverse Laplace transform for R, the complex output will be

Check

$$\begin{aligned} y(t) &= \int_0^\infty h_R(\tau)u(t-\tau)\,\mathrm{d}\tau \\ &= \sum_{j=1}^k c_j \int_0^\infty h_R(\tau)e^{a_j(t-\tau)}\,\mathrm{d}\tau \\ &= \sum_{j=1}^k c_j e^{a_j t} \int_0^\infty h_R(\tau)e^{-a_j\tau}\,\mathrm{d}t \\ &= \sum_{j=1}^k c_j e^{a_j t} R(a_j) \\ &= \sum_{j=1}^k c_j b_j e^{a_j t}, \end{aligned}$$

where we have used the definition of the Laplace transform. By part (i) of Theorem 5.21, and since $\|R\|_\infty \leq 1$, it follows that

$$\int_{-\infty}^0 |y(t)|^2\,\mathrm{d}t \leq \int_{-\infty}^0 |u(t)|^2\,\mathrm{d}t.$$

Substituting the current definitions of u and y gives

$$\begin{aligned} & \int_{-\infty}^0 \Big|\sum_{j=1}^k c_j b_j e^{a_j t}\Big|^2\,\mathrm{d}t \leq \int_{-\infty}^0 \Big|\sum_{j=1}^k c_j e^{a_j t}\Big|^2\,\mathrm{d}t \\ \implies \quad & \int_{-\infty}^0 \sum_{j,\ell=1}^k c_j e^{a_j t}\bar{c}_\ell e^{\bar{a}_\ell t}\,\mathrm{d}t \geq \int_{-\infty}^0 \sum_{j,\ell=1}^k c_j b_j e^{a_j t}\bar{c}_\ell \bar{b}_\ell e^{\bar{a}_\ell t}\,\mathrm{d}t \\ \implies \quad & \int_{-\infty}^0 \sum_{j,\ell=1}^k c_j \bar{c}_\ell (1 - b_j\bar{b}_\ell) e^{(a_j+\bar{a}_\ell)t}\,\mathrm{d}t \geq 0. \end{aligned}$$

We now compute

$$\int_{-\infty}^0 e^{(a_j+\bar{a}_\ell)t}\,\mathrm{d}t = \frac{1}{a_j+\bar{a}_\ell},$$

thus giving

$$\sum_{j,\ell=1}^k c_j \frac{1-b_j\bar{b}_\ell}{a_j+\bar{a}_\ell}\bar{c}_\ell \geq 0,$$

which we recognise as being equivalent to the expression

$$\boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x} \geq 0,$$

where

$$\boldsymbol{x} = \begin{bmatrix} \bar{c}_1 \\ \vdots \\ \bar{c}_n \end{bmatrix}.$$

Thus we have shown that the Pick matrix is positive-definite. ■

15.2.2 An inductive algorithm for solving the interpolation problem

In this section we provide a simple algorithm for solving the Nevanlinna-Pick interpolation problem in the situation when the Pick matrix is positive-definite. In doing so, we also complete the proof of Theorem 15.13. The algorithm we present in this section follows [Marshall 1975].

Before we state the algorithm, we need to introduce some notation. First note that by the Maximum Modulus Principle, it is necessary that for the Nevanlinna-Pick interpolation problem to have a solution, each of the numbers $b_1, \ldots, b_k$ satisfy $|b_j| \leq 1$. Thus we may assume this to be the case when we seek a solution to the problem. For $b \in \mathbb{C}$ satisfying $|b| < 1$, define the ***Blaschke function*** $B_b \in \mathbb{R}(s)$ associated with b by

$$B_b(s) = \begin{cases} \dfrac{s-b}{1-\bar{b}s}, & b \in \mathbb{R} \\ \dfrac{s^2 + 2\mathrm{Re}(b)s + |b|^2}{1 + 2\mathrm{Re}(b)s + |b|^2 s^2}, & \text{otherwise.} \end{cases}$$

Some easily verified relevant properties of Blaschke functions are the subject of Exercise E15.4. Also, for $a \in \mathbb{C}$ define a function $A_a \in \mathbb{R}(s)$ by

$$A_a(s) = \begin{cases} \dfrac{s-a}{s+\bar{a}}, & a \in \mathbb{R} \\ \dfrac{s^2 - 2\mathrm{Re}(a)s + |a|^2}{s^2 + 2\mathrm{Re}(a)s + |a|^2}, & \text{otherwise.} \end{cases}$$

Again, we refer to Exercise E15.4 for some of the easily proven properties of such functions.

Let us begin by solving the Nevanlinna-Pick interpolation problem when $k = 1$.

15.14 Lemma *Let $a_1 \in \mathbb{C}_+$ and let $b_1 \in \mathbb{C}$ have the property that $|b_1| < 1$. The associated Nevanlinna-Pick interpolation problem has an infinite number of solutions if it has one solution, and the set of all solutions is given by*

$$\big\{\mathrm{Re}(R) \;\big|\; R(s) = B_{-b_1}\big(R_1(s)A_{a_1}(s)\big),\ R_1 \in \mathbb{C}(s) \text{ has no poles in } \overline{\mathbb{C}}_+, \text{ and } \|R_1\|_\infty < 1\big\}.$$

Proof First note that the Nevanlinna-Pick interpolation problem does indeed have a solution, namely the trivial solution $R_0(s) = b_1$.

Now let $R_1 \in \mathbb{C}(s)$ have no poles in $\overline{\mathbb{C}}_+$ and suppose that $\|R_1\|_\infty < 1$. If

$$R(s) = B_{-b_1}\big(R_1(s)A_{a_1}(s)\big)$$

then R is the composition of the functions

$$\begin{aligned} s &\mapsto R_1(s)A_{a_1}(s) \\ s &\mapsto M_{-b_1}(s). \end{aligned}$$

The first of these functions in analytic in $\overline{\mathbb{C}}_+$ since both R_1 and A_{a_1} are. Also, by Exercise E15.4 and since $\|R_1\| < 1$, the first of these functions maps $\overline{\mathbb{C}}_+$ onto the disk $\bar{D}(0,1)$. The second of these functions, by Exercise E15.4, is analytic in $\bar{D}(0,1)$ and maps it onto itself. Thus we can conclude that $R \in \mathbb{C}(s)$ as defined has no poles in $\overline{\mathbb{C}}_+$ and that $\|R\|_\infty < 1$. What's more, we claim that $R(a_1) = b_1$ if $R_1(a_1) = b_1$. Indeed

$$R(a_1) = B_{-b_1}\big(R_1(a_1)A_{a_1}(a_1)\big) = B_{-b_1}(0) = b_1,$$

using the definitions of A_{a_1} and B_{b_1}. It only remains to show that $\mathrm{Re}(R)$ solves Problem 15.11. Clearly, since b_1 must be real, it follows that $\mathrm{Re}(R(a_1)) = b_1$. Furthermore, since $\|\mathrm{Re}(R)\|_\infty < \|R\|_\infty$, it follows that $\|\mathrm{Re}(R)\|_\infty < 1$. Finally, $\mathrm{Re}(R)$ can have no poles in $\overline{\mathbb{C}}_+$ since R has no poles in $\overline{\mathbb{C}}_+$.

Now suppose that $R \in \mathbb{R}(s)$ solves Problem 15.11. Define $R_1 \in \mathbb{C}(s)$ by

$$R_1(s) = \frac{B_{b_1}(R(s))}{A_{a_1}(s)}.$$

The function in the numerator is analytic in $\overline{\mathbb{C}}_+$ and has a zero at $s = a_1$. Therefore, since the only zero of A_{a_1} is at zero, R_1 is analytic in $\overline{\mathbb{C}}_+$. Furthermore, the H_∞-norm of the numerator is strictly bounded by 1, and since the H_∞-norm of the denominator equals 1, we conclude that $\|R_1\|_\infty < 1$. This concludes the proof of the lemma. ■

As stated in the proof of the lemma, if $|b_1| < 1$, then the one point interpolation problem always has the trivial solution $R_0(s) = b_1$. Let us also do this in the case when $k = 2$ and we have $a_2 = \bar{a}_1 \neq a_1$ and $b_2 = \bar{b}_2 \neq b_2$.

15.15 Lemma *Let $\{a_1, a_2 = \bar{a}_1\} \subset \mathbb{C}_+$ and let $\{b_1, b_2 = \bar{b}_1\} \subset \mathbb{C}$ have the property that $|b_1| < 1$. Also suppose that $a_1 \neq a_2$ and $b_1 \neq b_2$. The associated Nevanlinna-Pick interpolation problem has an infinite number of solutions if it has one solution, and the set of all solutions is given by*

$$\big\{\mathrm{Re}(R) \;\big|\; R(s) = B_{-b_1}\big(R_1(s)A_{a_1}(s)\big),\ R_1 \in \mathbb{C}(s) \text{ has no poles in } \overline{\mathbb{C}}_+, \text{ and } \|R_1\|_\infty < 1\big\}.$$

Proof Problem 15.11 has the solution

$$R_s(s) =$$

■

The lemmas gives the form of *all* solutions to the Nevanlinna-Pick interpolation problem in the cases when $k = 1$ and $k = 2$ with the points being complex conjugates of one another. It turns out that with this case, one can construct solutions to the general problem. To do this, one makes the clever observation (this was Nevanlinna's contribution) that one can reduce a k point interpolation problem to a $k-1$ or $k-2$ point interpolation problem by properly defining the new $k-1$ or $k-2$ points. We say how this is done in a definition.

15.16 Definition For $k > 1$, let

$$\{a_1, \dots, a_k\}, \{b_1, \dots, b_k\} \subset \mathbb{C}$$

be as in Problem 15.11. Define

$$\tilde{k} = \begin{cases} k-1, & b_k \in \mathbb{R} \\ k-1, & \text{otherwise.} \end{cases}$$

The ***Nevanlinna reduction*** of the numbers $\{a_1, \dots, a_k\}$ and $\{b_1, \dots, b_k\}$ is the collection of numbers

$$\{\tilde{a}_1, \dots, \tilde{a}_{\tilde{k}}\}, \{\tilde{b}_1, \dots, \tilde{b}_{\tilde{k}}\} \subset \mathbb{C}$$

defined by •

With this in mind, we state the algorithm that forms the main result of this section.

15.17 Algorithm for solving the Nevanlinna-Pick interpolation problem Given points

$$\{a_1, \dots, a_k\}, \{b_1, \dots, b_k\} \subset \mathbb{C}$$

as in Problem 15.11, additionally assume that $|b_j| < 1$, $j = 1, \dots, k$, and that the Pick matrix is positive-definite.

1. •

15.2.3 Relationship to the model matching problem

The above discussion of the Nevanlinna-Pick interpolation problem is not obviously related to the model matching problem, Problem 15.3.

15.18 Model matching by Nevanlinna-Pick interpolation Given $T_1, T_2 \in \mathrm{RH}_\infty^+$.

15.3 Optimal model matching II. Nehari's Theorem

In this section we present another method for obtaining a solution, or an approximate solution, to the model matching problem, Problem 15.3. The strategy in this section involves significantly more development than does the Nevanlinna-Pick procedure from Section 15.2. However, the algorithm produced in actually easier to apply than is that using Nevanlinna-Pick theory. Unfortunately, the methods in this section suffer from on occasion producing a controller that is improper, and one must devise hacks to get around this, just as with Nevanlinna-Pick theory. Our presentation in this section follows that of Francis [1987].

15.3.1 Hankel operators in the frequency-domain

The key tool for the methods of this section is something new for us: a Hankel operator of a certain type. To initiate this discussion, let us note that as in Proposition 14.9, but restriction to functions in RL_2, we have a decomposition $\mathrm{RL}_2 = \mathrm{RH}_2^- \oplus \mathrm{RH}_2^+$. That is to say, any strictly proper rational function with no poles on $i\mathbb{R}$ has a unique expression as a sum of a strictly proper rational function with no poles in $\overline{\mathbb{C}}_+$ and a strictly proper rational function with no poles in $\overline{\mathbb{C}}_-$. This is no surprise as this decomposition is simply obtained by partial fraction expansion. The essential idea of this section puts this mundane idea to good use. Let us denote by $\Pi^+ \colon \mathrm{RL}_2 \to \mathrm{RH}_2^+$ and $\Pi^- \colon \mathrm{RL}_2 \to \mathrm{RH}_2^-$ the projections.

Let us list some operators that are readily verified to have the stated properties.

1. ***The Laurent operator with symbol R:*** Given $R \in \mathrm{RL}_\infty$ and $Q \in \mathrm{RL}_2$, one readily sees that $RQ \in \mathrm{RL}_2$. Thus, given $R \in \mathrm{RL}_\infty$ we have a map $\Lambda_R \colon \mathrm{RL}_2 \to \mathrm{RL}_2$ defined by $\Lambda_R(Q) = RQ$. This is the Laurent operator with symbol R.
2. ***The Toeplitz operator with symbol R:*** Clearly, if $R \in \mathrm{RL}_\infty$ and if $Q \in \mathrm{RH}_2^+$, then $RQ \in \mathrm{RL}_2$. Therefore, $\Pi^+(RQ) \in \mathrm{RH}_2^+$. Thus, for $R \in \mathrm{RL}_\infty$, the map $\Theta_R \colon \mathrm{RH}_2^+ \to \mathrm{RH}_2^+$ defined by $\Theta_R(Q) = \Pi^+(RQ)$ is well-defined, and is called the Toeplitz operator with symbol R.
3. ***The Hankel operator with symbol R:*** Here again, if $R \in \mathrm{RL}_\infty$ and if $Q \in \mathrm{RH}_2^-$, then $RQ \in \mathrm{RL}_2$. Now we map this rational function into RH_2^+ using the projection Π^-. Thus, for $R \in \mathrm{RL}_\infty$, we define a map $\Gamma_R \colon \mathrm{RH}_2^+ \to \mathrm{RH}_2^-$ by $\Gamma_R(Q) = \Pi^-(RQ)$. This is the Hankel operator with symbol R.

15.19 Remarks

1. Note that the Toeplitz and Hankel operators together specify the value of the Laurent operator when applied to functions in RH_2^+. That is to say, if $R \in \mathrm{RL}_2$ then
$$\Lambda_R(Q) = \Theta_R(Q) + \Gamma_R(Q).$$
2. It is more common to see the Laurent, Toeplitz, and Hankel operators defined for general analytic functions rather than just rational functions. However, since our interest is entirely in the rational case, it is to this is that we restrict our interest. •

The Laurent, Toeplitz, and Hankel operators are linear. Thus it makes sense to ask questions about the nature of their spectrum. However, the spaces RL_2, RH_2^-, and RH_2^+ are infinite-dimensional, so these issues are not immediately approachable as they are in finite-dimensions. The good news, however, is that these operators are "essentially" finite-dimensional. The easiest way to make sense of this is with state-space techniques, and this is done in the next section.

It also turns out that the Laurent, Toeplitz, and Hankel operators are defined on spaces with an inner product. Indeed, on RL_2 we may define an inner product by

$$\langle R_1, R_2 \rangle_2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \overline{R_1(i\omega)} R_2(i\omega) \, \mathrm{d}\omega. \tag{15.5}$$

Note that this is an inner product on a real vector space. This inner product may clearly be applied to any functions in RL_2, including those in the subspaces RH_2^- and RH_2^+. Indeed, RH_2^- and RH_2^+ are orthogonal with respect to this inner product (see Exercise E15.5). One may define the ***adjoint*** of any of our operators with respect to this inner product. The adjoint of the Laurent operator with symbol R is the map $\Lambda_R^* \colon \mathrm{RL}_2 \to \mathrm{RL}_2$ defined by the relation

$$\langle \Lambda_R(R_1), R_2 \rangle_2 = \langle R_1, \Lambda_R^*(R_2) \rangle_2$$

for $R_1, R_2 \in \mathrm{RL}_2$. In like fashion, the Toeplitz operator has an adjoint $\Theta_R^* \colon \mathrm{RH}_2^+ \to \mathrm{RH}_2^+$ defined by

$$\langle \Theta_R(R_1), R_2 \rangle_2 = \langle R_1, \Theta_R^*(R_2) \rangle_2, \quad R_1, R_2 \in \mathrm{RH}_2^+,$$

and the Hankel operator has an adjoint $\Gamma_R^* \colon \mathrm{RH}_2^- \to \mathrm{RH}_2^+$ defined by

$$\langle \Gamma_R(R_1), R_2 \rangle_2 = \langle R_1, \Gamma_R^*(R_2) \rangle_2, \quad R_1 \in \mathrm{RH}_2^+, \ R_2 \in \mathrm{RH}_2^-.$$

The following result gives explicit formulae for the adjoints.

15.20 Proposition *For $R \in \mathrm{RL}_2$ the following statements hold:*

(i) $\Lambda_R^* = \Lambda_{R^*}$;

(ii) $\Theta_R^* = \Theta_{R^*}$;

(iii) $\Gamma_R^*(Q) = \Pi^+(\Lambda_{R^*}(Q))$.

Proof (i) We compute

$$\begin{aligned}\langle \Lambda_R(R_1), R_2\rangle_2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{\Lambda_R(R_1)(i\omega)} R_2(i\omega)\,\mathrm{d}\omega\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{R(i\omega)R_1(i\omega)} R_2(i\omega)\,\mathrm{d}\omega\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{R_1(i\omega)} R(-i\omega) R_2(i\omega)\,\mathrm{d}\omega\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{R_1(i\omega)} R^*(i\omega) R_2(i\omega)\,\mathrm{d}\omega\\ &= \langle R_1, \Lambda_{R^*}(R_2)\rangle_2.\end{aligned}$$

This then gives $\Lambda_R^* = \Lambda_{R^*}$ as desired.

(ii) For $R_1, R_2 \in \mathrm{RH}_2^+$ we compute

$$\begin{aligned}\langle \Theta_R(R_1), R_2\rangle_2 &= \langle \Lambda_R(R_1), R_2\rangle_2\\ &= \langle R_1, \Lambda_R^*(R_2)\rangle_2\\ &= \langle R_1, \Lambda_{R^*}(R_2)\rangle_2\\ &= \langle R_1, \Theta_{R^*}(R_2)\rangle_2,\end{aligned}$$

and this part of the proposition follows.

(iii) For $R_1 \in \mathrm{RH}_2^+$ and $R_2 \in \mathrm{RH}_2^-$ we compute

$$\begin{aligned}\langle \Gamma_R(R_1), R_2\rangle_2 &= \langle \Lambda_R(R_1), R_2\rangle_2\\ &= \langle R_1, \Lambda_R^*(R_2)\rangle_2\\ &= \langle R_1, \Lambda_{R^*}(R_2)\rangle_2,\end{aligned}$$

and from this the result follows. ■

In the next section, we will come up with concrete realisations of the Hankel operator and its adjoint using time-domain methods.

15.3.2 Hankel operators in the time-domain

The above operators defined in the rational function domain are simple enough, but they have interesting and nontrivial counterparts in the time-domain. To simplify matters, let us denote by $\bar{\mathrm{L}}_2(-\infty,\infty)$ those functions of time that, when Laplace transformed, give functions in RL_2. As we saw in Section E.3, this consists exactly of sums of products of polynomial functions, trigonometric functions, and exponential functions of time. Let us denote by $\bar{\mathrm{L}}_2(-\infty,0]$ the subset of $\bar{\mathrm{L}}_2(-\infty,\infty)$ consisting of functions that are bounded for $t<0$, and by $\bar{\mathrm{L}}_2[0,\infty)$ the subset of $\bar{\mathrm{L}}_2(-\infty,\infty)$ consisting of functions that are bounded for $t>0$. Note that

$$\bar{\mathrm{L}}_2(-\infty,\infty) = \bar{\mathrm{L}}_2(-\infty,0] \oplus \bar{\mathrm{L}}_2[0,\infty).$$

That is, every function in $\bar{\mathrm{L}}_2(-\infty,\infty)$ can be uniquely decomposed into a sum of two functions, one that is bounded for $t<0$ and one that is bounded for $t>0$. It is clear that this decomposition corresponds exactly to the decomposition $\mathrm{RL}_2 = \mathrm{RH}_2^- \oplus \mathrm{RH}_2^+$ that uses the partial fraction expansion. Let us also define projections

$$\begin{aligned}\bar{\Pi}^+ &: \bar{\mathrm{L}}_2(-\infty,\infty) \to \bar{\mathrm{L}}_2[0,\infty)\\ \bar{\Pi}^- &: \bar{\mathrm{L}}_2(-\infty,\infty) \to \bar{\mathrm{L}}_2(-\infty,0].\end{aligned}$$

If we employ the inner product

$$\langle f_1, f_2 \rangle_2 = \int_{-\infty}^{\infty} f_1(t) f_2(t) \, \mathrm{d}t \tag{15.6}$$

on $\bar{\mathrm{L}}_2(-\infty, \infty)$, then obviously $\bar{\mathrm{L}}_2(-\infty, 0]$ and $\bar{\mathrm{L}}_2[0, \infty)$ are orthogonal. We hope that it will be clear from context what we mean when we use the symbol $\langle \cdot, \cdot \rangle_2$ in two different ways, one for the frequency-domain, and the other for the time-domain.

The following result summarises the previous discussion.

15.21 Proposition *The Laplace transform is a bijection*

(i) from $\bar{\mathrm{L}}_2(-\infty, \infty)$ *to* RL_2,

(ii) from $\bar{\mathrm{L}}_2(-\infty, 0]$ *to* RH_2^-, *and*

(iii) from $\bar{\mathrm{L}}_2[0, \infty)$ *to* RH_2^+.

Furthermore, the diagrams

$$\begin{array}{ccc} \bar{\mathrm{L}}_2(-\infty, \infty) & \xrightarrow{\bar{\Pi}^+} & \bar{\mathrm{L}}_2[0, \infty) \\ \downarrow{\scriptstyle \mathcal{L}} & & \downarrow{\scriptstyle \mathcal{L}} \\ \mathrm{RL}_2 & \xrightarrow[\Pi^+]{} & \mathrm{RL}_2^+ \end{array} \qquad \begin{array}{ccc} \bar{\mathrm{L}}_2(-\infty, \infty) & \xrightarrow{\bar{\Pi}^-} & \bar{\mathrm{L}}_2(-\infty, 0] \\ \downarrow{\scriptstyle \mathcal{L}} & & \downarrow{\scriptstyle \mathcal{L}} \\ \mathrm{RL}_2 & \xrightarrow[\Pi^+]{} & \mathrm{RL}_2^- \end{array}$$

commute.

We now turn our attention to describing how the operators of Section 15.3.1 appear in the time-domain, given the correspondence of Proposition 15.21. Our interest is particularly in the Hankel operator. Given Proposition 15.21 we expect the analogue of the frequency domain Hankel operator to map $\bar{\mathrm{L}}_2[0, \infty)$ to $\bar{\mathrm{L}}_2(-\infty, 0]$, given $R \in \mathrm{RL}_\infty$. To do this, given $R \in \mathrm{RL}_\infty$, let us write $R = R_1 + R_2$ for $R_1 \in \mathrm{RH}_2^-$ and $R_2 \in \mathrm{RH}_\infty^+$ as in Proposition 14.9. We then let $\Sigma_1 = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be the complete SISO linear system in controller canonical form with the property that $T_{\Sigma_1} = R_1$. Therefore, the inverse Laplace transform for R_1 is the impulse response for Σ_1. Note that $\sigma(\boldsymbol{A}) \subset \mathbb{C}_+$. Thus if $r_1 \in \bar{\mathrm{L}}_2(-\infty, 0]$ is the inverse Laplace transform of R_1 we have

$$r_1(t) = \begin{cases} -\boldsymbol{c}^t e^{\boldsymbol{A}t} \boldsymbol{b}, & t \leq 0 \\ 0, & t > 0, \end{cases}$$

which is the anticausal impulse response for R_1. Now, for $u \in \bar{\mathrm{L}}_2[0, \infty)$ and for $t \leq 0$, let us define

$$\bar{\Gamma}_R(u)(t) = \int_0^{\infty} r_1(t - \tau) u(\tau) \, \mathrm{d}\tau. \tag{15.7}$$

We take $\bar{\Gamma}_R(u)(t) = 0$ for $t > 0$. We claim that $\bar{\Gamma}_R$ is the time-domain version of the Hankel operator Γ_R. Let us first prove that its takes its values in the right space.

15.22 Lemma $\bar{\Gamma}_R(u) \in \bar{\mathrm{L}}_2(-\infty, 0]$.

Proof For $t \geq 0$ we have

$$\bar{\Gamma}_R(u)(t) = -\boldsymbol{c}^t e^{\boldsymbol{A}t} \int_0^{\infty} e^{-\boldsymbol{A}\tau} \boldsymbol{b} u(\tau) \, \mathrm{d}\tau.$$

Since $\boldsymbol{A}$ has all eigenvalues in $\mathbb{C}_+$, $-\boldsymbol{A}$ has all eigenvalues in $\mathbb{C}_-$, so the integral converges. Also, for the same reason, $e^{\boldsymbol{A}t}$ is bounded for $t \leq 0$. This shows that $\bar{\Gamma}_R(u) \in \bar{\mathrm{L}}_2(-\infty, 0]$, as claimed. ■

Now let us show that this is indeed "the same" as the frequency domain Hankel operator.

15.23 Proposition *Let $R \in \mathrm{RL}_\infty$ and write $R = R_1 + R_2$ with $R_1 \in \mathrm{RH}_2^-$ and $R_2 \in \mathrm{RH}_\infty^+$. Let $r_1 \in \bar{\mathrm{L}}_2(-\infty, 0]$ be the inverse Laplace transform of R_1 and define $\bar{\Gamma}_R$ as in (15.7). Then the diagram*

$$\begin{array}{ccc} \bar{\mathrm{L}}_2[0,\infty) & \xrightarrow{\bar{\Gamma}_R} & \bar{\mathrm{L}}_2(-\infty,0] \\ \downarrow{\scriptstyle \mathscr{L}} & & \downarrow{\scriptstyle \mathscr{L}} \\ \mathrm{RL}_2^+ & \xrightarrow[\Gamma_R]{} & \mathrm{RL}_2^- \end{array}$$

commutes.

Proof Let $u \in \bar{\mathrm{L}}_2[0,\infty)$ and denote $y = \bar{\Gamma}_R(u) \in \bar{\mathrm{L}}(-\infty, 0]$. Denote as usual the Laplace transforms of u and y by $\hat{u}$ and $\hat{y}$. We then have

$$\Gamma_R(\hat{u}) = \Pi^-((R_1 + R_2)\hat{u}).$$

Note that $R_2\hat{u} \in \mathrm{RH}_2^+$. Therefore, $\Pi^-((R_1 + R_2)\hat{u}) = \Pi^-(R_1\hat{u})$. Now let us compute the inverse Laplace transform of $R_1\hat{u}$. Let $\tilde{\Sigma} = (\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{c}}^t, \boldsymbol{0}_1)$ be a complete SISO linear system defined so that $T_{\tilde{\Sigma}} = \hat{u}$. Thus $\tilde{\boldsymbol{A}}$ has all eigenvalues in $\mathbb{C}_-$. Now we compute

$$\begin{aligned} \mathscr{L}^{-1}(R_1\hat{u})(t) &= \int_{-\infty}^{\infty} r_1(t-\tau)u(\tau)\,\mathrm{d}\tau \\ &= -\boldsymbol{c}^t e^{\boldsymbol{A}t} \int_{-\infty}^{\infty} 1(\tau - t)e^{-\boldsymbol{A}\tau}\boldsymbol{b}u(\tau)\,\mathrm{d}\tau \\ &= -\boldsymbol{c}^t e^{\boldsymbol{A}t} \int_{0}^{\infty} 1(\tau - t)e^{-\boldsymbol{A}\tau}\boldsymbol{b}u(\tau)\,\mathrm{d}\tau, \end{aligned}$$

since for $\tau < 0$, $u(\tau) = 0$. Now note that $\bar{\Pi}^-(\mathscr{L}^{-1}(R_1\hat{u}))$ is nonzero only for $t < 0$ so that we can write

$$(\bar{\Pi}^-(\mathscr{L}^{-1}(R_1\hat{u})))(t) = -\boldsymbol{c}^t e^{\boldsymbol{A}t} \int_0^{\infty} e^{\boldsymbol{A}\tau}\boldsymbol{b}u(\tau)\,\mathrm{d}\tau.$$

Thus $\bar{\Pi}^-(\mathscr{L}^{-1}(R_1\hat{u})) = \bar{\Gamma}_R(\hat{u})$. By Proposition 15.21 this means that

$$\mathscr{L}^{-1}(\Pi^-(R_1\hat{u})) = \mathscr{L}^{-1}(\Gamma_R(\hat{u})) = \bar{\Gamma}_R(u),$$

or, equivalently, that $\Gamma_R \circ \mathscr{L} = \mathscr{L} \circ \bar{\Gamma}_R$, as claimed. ■

Up to this point, the value of the time-domain formulation of a Hankel operator is not at all clear. The simple act of multiplication and partial fraction expansion in the frequency-domain becomes a little more abstract in the time-domain. However, the value of the time-domain formulation is in its presenting a concrete representation of the Hankel operator and its adjoint. To come up with this representation, we introduce some machinery harking back to our observability and controllability discussion in Sections 2.3.1 and 2.3.2. In particular, we begin to dig into the proof of Theorem 2.21. We resume with the situation where $R = R_1 + R_2 \in \mathrm{RL}_\infty$ with $R_1 \in \mathrm{RH}_2^-$ and $R_2 \in \mathrm{RH}_\infty^+$. As above, we let $\Sigma_1 = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be the canonical minimal realisation of R_1. With this notation, we define a map $\mathscr{C}_R\colon \bar{\mathrm{L}}_2[0,\infty) \to \mathbb{R}^n$ by

$$\mathscr{C}_R(u) = -\int_0^{\infty} e^{-\boldsymbol{A}\tau}\boldsymbol{b}u(\tau)\,\mathrm{d}\tau,$$

and we call this the ***controllability operator***. Similarly, we define $\mathscr{O}_R\colon \mathbb{R}^n \to \bar{\mathrm{L}}_2(-\infty, 0]$ by

$$(\mathscr{O}_R(\boldsymbol{x}))(t) = \begin{cases} \boldsymbol{c}^t e^{\boldsymbol{A}t}\boldsymbol{x}, & t \geq 0 \\ 0, & t < 0, \end{cases}$$

and we call this the ***observability operator***. Note that the adjoint of the controllability operator will be the map $\mathscr{C}_R^*\colon \mathbb{R}^n \to \bar{\mathrm{L}}_2[0, \infty)$ satisfying

$$\langle \mathscr{C}_R(u), \boldsymbol{x}\rangle = \langle u, \mathscr{C}_R^*(\boldsymbol{x})\rangle_2, \quad u \in \bar{\mathrm{L}}_2[0, \infty),\ \boldsymbol{x} \in \mathbb{R}^n,$$

and the adjoint of the observability operator will be the map $\mathscr{O}_R^*\colon \bar{\mathrm{L}}_2(-\infty, 0] \to \mathbb{R}^n$ satisfying

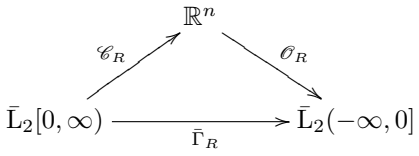
$$\langle \mathscr{O}_R(\boldsymbol{x}), y\rangle_2 = \langle \boldsymbol{x}, \mathscr{O}_R^*(y)\rangle, \quad y \in \bar{\mathrm{L}}_2(-\infty, 0],\ \boldsymbol{x} \in \mathbb{R}^n.$$

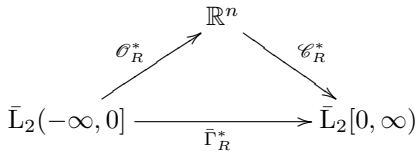
In each case, the inner product of equation (15.6) is being used on $\bar{\mathrm{L}}_2(-\infty, 0]$ and $\bar{\mathrm{L}}_2[0, \infty)$.

The following result summarises the value of introducing this notation.

15.24 Proposition *With the above notation, the following statements hold:*

(i) the diagrams





commute;

(ii) $(\mathscr{C}_R^*(\boldsymbol{x}))(\tau) = \begin{cases} -\boldsymbol{b}^t e^{-\boldsymbol{A}^t\tau}\boldsymbol{x}, & t \leq 0 \\ 0, & t > 0; \end{cases}$

(iii) $\mathscr{O}_R^*(y) = \displaystyle\int_{-\infty}^{0} e^{\boldsymbol{A}^t t}\boldsymbol{c}y(t)\,\mathrm{d}t;$

(iv) $(\bar{\Gamma}_R^*(y))(\tau) = \begin{cases} -\int_0^\infty \boldsymbol{b}^t e^{\boldsymbol{A}^t(t-\tau)}\boldsymbol{c}y(t)\,\mathrm{d}t, & \tau \leq 0 \\ 0, & \tau > 0; \end{cases}$

(v) $\mathscr{C}_R^*$ *is injective;*

Proof (i) It suffices to show that the left diagram commutes, since if it does, the right diagram will also commute by the definition of the adjoint. However, the left diagram may be easily seen to commute by virtue of the very definitions of $\mathscr{C}_R$, $\mathscr{O}_R$, and $\bar{\Gamma}_R$.

(ii) This follows from the definition of $\mathscr{C}_R$ and the inner product in equation (15.6).

(iii) This follows from the definition of $\mathscr{O}_R$ and the inner product in equation (15.6).

(iv) This follows from the right diagram in part (i), along with parts (ii) and (ii).

(v) It suffices to show that $\mathscr{C}_R^*(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x} = 0$. If $(\mathscr{C}_R^*(\boldsymbol{x}))(\tau) = -\boldsymbol{b}^t e^{-\boldsymbol{A}^t\tau}\boldsymbol{x} = 0$ for all τ then successive differentiation with respect to τ and evaluation at $\tau = 0$ gives

$$-\boldsymbol{b}^t\boldsymbol{x} = 0, \quad \boldsymbol{b}^t\boldsymbol{A}^t\boldsymbol{x} = 0, \ldots, (-1)^n\boldsymbol{b}^t(\boldsymbol{A}^t)^{n-1}\boldsymbol{x} = 0.$$

Since $(\boldsymbol{A}, \boldsymbol{b})$ is controllable, this implies that $\boldsymbol{x} = \boldsymbol{0}$. ■

Thus the above result gives a simple way of relating a Hankel operator and its adjoint to operators with either a domain or a range that is finite-dimensional. In the next section, we shall put this to good use.

15.3.3 Hankel singular values and Schmidt pairs

Recall that a ***singular value*** for a linear map A between inner product spaces is, by definition, an eigenvalue of A^*A. One readily verifies that singular values are real and nonnegative. Our objective in this section is to find the singular values of a Hankel operator Γ_R using our representation $\bar{\Gamma}_R$ in the time-domain.

As in the previous section, for $R \in \mathrm{RL}_\infty$ we write $R = R_1 + R_2$ with $R_1 \in \mathrm{RH}_2^+$ and $R_2 \in \mathrm{RH}_\infty^-$. We let $\Sigma_1 = (\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}^t, \boldsymbol{0}_1)$ be the complete SISO linear system in controller canonical form so that $T_{\Sigma_1} = R_1$. We next introduce the ***controllability Gramian***

$$\boldsymbol{C}_R = \int_0^\infty e^{-\boldsymbol{A}t}\boldsymbol{b}\boldsymbol{b}^t e^{-\boldsymbol{A}^t t}\,\mathrm{d}t,$$

and the ***observability Gramian***

$$\boldsymbol{O}_R = \int_0^\infty e^{-\boldsymbol{A}^t t}\boldsymbol{c}\boldsymbol{c}^t e^{-\boldsymbol{A}t}\,\mathrm{d}t.$$

These are each elements of $\mathbb{R}^{n\times n}$. We have previously encountered the controllability Gramian in the proof of Theorem 2.21, and the observability Gramian may be used in a similar manner. However, here we are interested in their relationship with the Hankel operator Γ_R. The following result gives this relationship, as well as providing a characterisation of $\boldsymbol{C}_R$ and $\boldsymbol{O}_R$ in terms of the Liapunov ideas of Section 5.4.

15.25 Proposition *With the above notation, the following statements hold:*

(i) $(\boldsymbol{A}^t, \boldsymbol{C}_R, -\boldsymbol{b}\boldsymbol{b}^t)$ *is a Liapunov triple;*

(ii) $(\boldsymbol{A}, \boldsymbol{O}_R, -\boldsymbol{c}\boldsymbol{c}^t)$ *is a Liapunov triple;*

(iii) $\mathscr{C}_R \circ \mathscr{C}_R^* = \boldsymbol{C}_R$*;*

(iv) $\mathscr{O}_R^* \circ \mathscr{O}_R = \boldsymbol{O}_R$*;*

(v) $\boldsymbol{O}_R$ *is invertible.*

Proof (i) Since $-\boldsymbol{A}$ is Hurwitz, by part (i) of Theorem 5.32 there is a unique symmetric matrix $\boldsymbol{P}$ so that $(-\boldsymbol{A}^t, \boldsymbol{P}, \boldsymbol{b}\boldsymbol{b}^t)$ is a Liapunov triple. What's more, the proof of Theorem 5.32 gives $\boldsymbol{P}$ explicitly as

$$\boldsymbol{P} = \int_0^\infty e^{-\boldsymbol{A}t}\boldsymbol{b}\boldsymbol{b}^t e^{-\boldsymbol{A}^t t}\,\mathrm{d}t.$$

Now one sees trivially that $(\boldsymbol{A}, -\boldsymbol{P}, -\boldsymbol{b}, \boldsymbol{b}^t)$ is also a Liapunov triple. This part of the proposition now follows because $\boldsymbol{C}_R = -\boldsymbol{P}$.

(ii) The proof here is exactly as for part (i).

(iii) This follows from the characterisations of $\mathscr{C}_R$ and $\mathscr{C}_R^*$ given in Proposition 15.24.

(iv) This follows from the characterisations of $\mathscr{O}_R$ and $\mathscr{O}_R^*$ given in Proposition 15.24.

(v) Since $\boldsymbol{O}_R$ is square, injectivity is equivalent to invertibility. Suppose that $\boldsymbol{O}_R$ is not invertible. Then, since $\boldsymbol{O}_R$ is positive-semidefinite, there exists $\boldsymbol{x} \in \mathbb{R}^n$ so that $\boldsymbol{x}^t\boldsymbol{O}_R\boldsymbol{x} = 0$, or so that

$$\int_0^\infty \boldsymbol{x}^t e^{-\boldsymbol{A}^t t}\boldsymbol{c}\boldsymbol{c}^t e^{-\boldsymbol{A}t}\boldsymbol{x}\,\mathrm{d}t.$$

This means that $\boldsymbol{c}^t e^{-\boldsymbol{A}t}\boldsymbol{x} = 0$ for all $t \in [0, \infty)$. Differentiating successively with respect to t at $t = 0$ gives

$$\boldsymbol{c}^t\boldsymbol{x} = 0, \quad -\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{x} = 0, \ldots, (-1)^{n-1}\boldsymbol{c}^t\boldsymbol{A}^{n-1}\boldsymbol{x} = 0.$$

This implies that $(\boldsymbol{A}, \boldsymbol{c})$ is not observable. It therefore follows that $\boldsymbol{O}_R$ is indeed injective. ■

This, then, is interesting as it affords us the possibility of characterising the singular values of the Hankel operator in terms of the eigenvalues of an $n \times n$ matrix. This is summarised in the following result.

15.26 Theorem *The nonzero eigenvalues of the following three operators,*

(i) $\Gamma_R^* \circ \Gamma_R$,

(ii) $\bar{\Gamma}_R^* \circ \bar{\Gamma}_R$, *and*

(iii) $\boldsymbol{C}_R\boldsymbol{O}_R$,

agree.

Proof That the eigenvalues for $\Gamma_R^*\Gamma_R$ and $\bar{\Gamma}_R^*\bar{\Gamma}_R$ agree is a simple consequence of Proposition 15.23: the Laplace transform or its inverse will deliver eigenvalues and eigenvectors for either of $\Gamma_R^*\Gamma_R$ or $\bar{\Gamma}_R^*\bar{\Gamma}_R$ given eigenvalues and eigenvectors for the other.

Now let $\sigma^2 > 0$ be an eigenvalue for $\bar{\Gamma}_R^*\bar{\Gamma}_R$ with eigenvector $u \in \bar{\mathrm{L}}_2(-\infty, 0]$. By part (i) of Proposition 15.24 this means that

$$\begin{aligned} & \mathscr{C}_R^*\mathscr{O}_R^*\mathscr{O}_R\mathscr{C}_R(u) = \sigma^2 u \\ \Longrightarrow \quad & \mathscr{C}_R\mathscr{C}_R^*\mathscr{O}_R^*\mathscr{O}_R\mathscr{C}_R(u) = \sigma^2\mathscr{C}_R(u). \end{aligned}$$

If $\boldsymbol{x} = \mathscr{C}_R(u)$ then $\boldsymbol{x} \neq \boldsymbol{0}$ since otherwise it would follow that $\sigma^2 = 0$. This shows that σ^2 is an eigenvalue of $\boldsymbol{C}_R\boldsymbol{O}_R$ with eigenvector $\boldsymbol{x}$.

Now suppose that $\sigma^2 \neq 0$ is an eigenvalue for $\boldsymbol{C}_R\boldsymbol{O}_R$ with eigenvector $\boldsymbol{x}$. Thus

$$\begin{aligned} & \boldsymbol{C}_R\boldsymbol{O}_R\boldsymbol{x} = \sigma^2\boldsymbol{x} \\ \Longrightarrow \quad & \mathscr{C}_R^*\boldsymbol{O}_R\boldsymbol{C}_R\boldsymbol{O}_R = \sigma^2\mathscr{C}_R^*\boldsymbol{O}_R\boldsymbol{x}. \end{aligned}$$

If $u = \mathscr{C}_R^*\boldsymbol{O}_R\boldsymbol{x}$ then we claim that $u \neq 0$. Indeed, from part (v) of Proposition 15.24, $\mathscr{C}_R^*$ is injective, and from part (v) of Proposition 15.25, $\boldsymbol{O}_R$ is injective. Thus $u = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$. Thus we see that σ^2 is an eigenvalue for $\bar{\Gamma}_R^*\bar{\Gamma}_R$ with eigenvalue u. ■

Thus we have a characterisation of all nonzero singular values of the Hankel operator as eigenvalues of an $n \times n$ matrix. This is something of a coup. We shall suppose that the nonzero singular values are arranged in descending order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$, so that σ_1 denotes the largest of the singular values. We shall call $\sigma_1, \ldots, \sigma_k$ the ***Hankel singular values*** for the Hankel operator Γ_R.

Now we wish to talk about the "size" of a Hankel operator Γ_R. Since Γ_R is a linear map between two inner product spaces—from RH_2^+ to RH_2^-—we may simply define its norm in the same manner in which we defined the induced signal norms in Definition 5.19. Thus we define

$$\|\Gamma_R\| = \sup_{\substack{Q \in \mathrm{RH}_2^+ \\ Q \text{ not zero}}} \frac{\|\Gamma_R(Q)\|_2}{\|Q\|_2}.$$

This is called the ***Hankel norm*** of the Hankel operator Γ_R. The following result follows easily from Theorem 15.26 if one knows just a little more operator theory than is really within the confines of this course. However, it is an essential result for us.

15.27 Corollary *If σ_1 is the largest Hankel singular value, then $\|\Gamma_R\| = \sigma_1$.*

Proof We can prove this using matrix norms. ■ finish

Next, let us look a little more closely at eigenvectors induced by singular values. Thus we let σ^2 be a nonzero singular value for $\bar{\Gamma}_R^*\bar{\Gamma}_R$ with eigenvector $u_1 \in \bar{L}_2[0,\infty)$. Now define $u_2 \in \bar{L}_2(-\infty, 0]$ by $u_2 = \frac{1}{\sigma}\bar{\Gamma}_R(u_1)$. Then one readily computes

$$\begin{aligned}\bar{\Gamma}_R(u_1) &= \sigma u_2\\ \bar{\Gamma}_R^*(u_2) &= \sigma u_1.\end{aligned}$$

When an operator and its adjoint possess the same eigenvalue in this manner, the resulting eigenvectors (u_1, u_2) are called a ***σ-Schmidt pair*** for the operator. Of course, if R_j is the Laplace transform of u_j, $j = 1, 2$, then we have

$$\begin{aligned}\Gamma_R(R_1) &= \sigma R_2\\ \Gamma_R^*(R_2) &= \sigma R_1,\end{aligned}$$

so that $(R_1, R_2) \in \mathrm{RH}_2^+ \times \mathrm{RH}_2^-$ are a σ-Schmidt pair for Γ_R. The matter of finding Schmidt pairs for Hankel operators is a simple enough proposition as one may use Theorem 15.26. Indeed, suppose that $\sigma^2 > 0$ is an eigenvalue for $\boldsymbol{C}_R\boldsymbol{O}_R$ with eigenvector $\boldsymbol{x}$. Then a σ-Schmidt pair for $\bar{\Gamma}_R$ is readily verified to be given by (u_1, u_2) where

$$\begin{aligned}u_1 &= \frac{1}{\sigma}\mathscr{O}_R^*(\boldsymbol{O}_R\boldsymbol{x})\\ u_2 &= \mathscr{O}_R(\boldsymbol{x}).\end{aligned}$$

15.3.4 Nehari's Theorem

In this section we state and prove a famous theorem of Nehari [1957]. This theorem is just one in a sweeping research effort in "Hankel norm approximation," with key contributions being made in a sequence of papers by Adamjan, Arov, and Krein (1968, 1968, 1971). Our interest in this section is in a special version of this rather general work, as we are only interested in rational functions, whereas Nehari was interested in general H_∞ functions.

15.28 Theorem *Let $R_0 \in \mathrm{RH}_\infty^-$, let $\sigma_1 > 0$ be the largest Hankel singular value for R_0, and let $(R_1, R_2) \in \mathrm{RH}_2^+ \times \mathrm{RH}_2^-$ be a σ_1-Schmidt pair. Then*

$$\inf_{R \in \mathrm{RH}_\infty^+} \|R_0 - R\|_\infty = \sigma_1,$$

and if $R \in \mathrm{RH}_\infty^+$ satisfies $R_1(R_0 - R) = \sigma_1 R_2$ then $\|R_0 - R\|_\infty = \sigma_1$.

Proof First let us show that σ_1 is a lower bound for $\|R_0 - R\|_\infty$. For any $R \in \mathrm{RH}_2^+$ we compute, using part (i) of Theorem 5.21,

$$\begin{aligned}\|R_0 - R\|_\infty &= \sup_{\substack{Q \in \mathrm{RH}_2^+\\ Q \text{ not zero}}} \frac{\|(R_0 - R)Q\|_2}{\|Q\|_2}\\ &\geq \sup_{\substack{Q \in \mathrm{RH}_2^+\\ Q \text{ not zero}}} \frac{\|\Pi^-(R_0 - R)Q\|_2}{\|Q\|_2}\\ &= \sup_{\substack{Q \in \mathrm{RH}_2^+\\ Q \text{ not zero}}} \frac{\|\Pi^-(R_0)Q\|_2}{\|Q\|_2}\\ &= \|\Gamma_R\|.\end{aligned}$$

Now let (R_1, R_2) be a σ_1-Schmidt pair and write $Q = (R_0 - R)R_1$ for $R \in \mathrm{RH}_\infty^+$. Since $R_1 \in \mathrm{RH}_2^+$, $\Gamma_{R_0}(R_1) \in \mathrm{RH}_2^+$. Since $RR_1 \in \mathrm{RH}_2^+$, $\Pi^-(Q) = \Pi^-(R_0R_1) = \Gamma_{R_0}(R_1)$. Therefore, we compute

$$\begin{aligned}
0 &\leq \|Q - \Gamma_{R_0}(R_1)\|_2^2 \\
&= \|Q\|_2^2 + \langle \Gamma_{R_0}(R_1), \Gamma_{R_0}(R_1)\rangle_2 - 2\langle Q, \Gamma_{R_0}(R_1)\rangle_2 \\
&= \|Q\|_2^2 + \langle \Gamma_{R_0}(R_1), \Gamma_{R_0}(R_1)\rangle_2 - 2\langle \Pi^-(Q), \Gamma_{R_0}(R_1)\rangle_2 \\
&= \|Q\|_2^2 - \langle \Gamma_{R_0}(R_1), \Gamma_{R_0}(R_1)\rangle_2 \\
&= \|Q\|_2^2 - \langle R_1, \Gamma_{R_0}^* \Gamma_{R_0}(R_1)\rangle_2 \\
&= \|Q\|_2^2 - \sigma_1^2 \langle R_1, R_1\rangle_2 \\
&= \|Q\|_2^2 - \sigma_1^2 \|R_1\|_2^2 \\
&\leq \|R_0 - R\|_\infty^2 \|R_1\|_2^2 - \sigma_1^2 \|R_1\|_2^2 \\
&= \left(\|R_0 - R\|_\infty^2 - \sigma_1^2\right) \|R_1\|_2^2 \\
&\geq 0.
\end{aligned}$$

This shows that $Q = \Gamma_{R_0}(R_1)$, or, equivalently,

$$(R_0 - R)R_1 = \Gamma_{R_0}(R_1) = \sigma_1 R_2,$$

as claimed. ■

15.3.5 Relationship to the model matching problem

The previous buildup has been significant, and it is perhaps not transparent how Hankel operators and Nehari's Theorem relate in any way to the model matching problem. The relationship is, in fact, quite simple, and in this section we give a simple algorithm for obtaining a solution to the model matching problem using the tools of this section. However, as with Nevanlinna-Pick theory, there is a drawback in that on occasion a hack will have to be employed. Nonetheless, the process is systematic enough.

Let us come right out and state the algorithm.

15.29 Model matching by Hankel norm approximation Given $T_1, T_2 \in \mathrm{RH}_\infty^+$.

15.4 A robust performance example

15.5 Other problems involving H_∞ methods

It turns out that the robust performance problem is only one of a number of problems falling under the umbrella of H_∞ control. In this section we briefly indicate some other problems whose solution can be reduced to a model matching problem, and thus whose solution can be obtained by the methods in this chapter.

Exercises

E15.1 Exercise in the ∞-norm giving a Banach algebra.

E15.2 Exercise on existence of solutions to the model matching problem.

E15.3 Verify that the following algorithm for reducing the modified robust performance problem for additive uncertainty actually works.

15.30 Algorithm for obtaining model matching problem for additive uncertainty Given $\bar{R}_P$, W_u, and W_p as in Problem 15.2.

1. Define
$$U_3 = \frac{W_p W_p^* W_u W_u^*}{W_p W_p^* + W_u W_u^*}.$$
2. If $\|U_3\|_\infty \geq \frac{1}{2}$, then Problem 15.2 has no solution.
3. Let (P_1, P_2) be a coprime fractional representative for $\bar{R}_P$.
4. Let (ρ_1, ρ_2) be a coprime factorisation for P_1 and P_2:
$$\rho_1 P_1 + \rho_2 P_2 = 1.$$
5. Define
$$\begin{aligned} R_1 &= W_p \rho_2 P_2, & S_1 &= W_u \rho_1 P_1, \\ R_2 &= W_p P_1 P_2, & S_2 &= -W_u P_1 P_2. \end{aligned}$$
6. Define $Q = [R_2 R_2^* + S_2 S_2^*]^+$.
7. Let V be an inner function with the property that
$$\frac{R_1 R_2^* + S_1 S_2^*}{Q^*} V$$
has no poles in $\overline{\mathbb{C}}_+$.
8. Define
$$U_1 = \frac{R_1 R_2^* + S_1 S_2^*}{Q^*} V, \qquad U_2 = QV.$$
9. Define $U_4 = [\frac{1}{2} - U_3]^+$.
10. Define
$$T_1 = \frac{U_1}{U_4}, \quad T_2 = \frac{U_2}{U_4}.$$
11. Let θ be a solution to Problem 15.3.
12. If $\|T_1 - \theta T_2\|_\infty \geq 1$ then Problem 15.2 has no solution.
13. The controller
$$R_C = \frac{\rho_1 + \theta P_2}{\rho_2 - \theta P_1},$$
is a solution to Problem 15.2. •

Finish

E15.4 Möbius functions

E15.5 Show that RH_2^- and RH_2^+ are orthogonal with respect to the inner product on RL_2 defined in equation (15.5).

Part IV

Background material

Appendix A

Linear algebra

Formulation of the time-domain setting for linear systems requires fluency with linear algebra. Those of you taking this course are expected to be *very* familiar with essentials of linear algebra. In this appendix we will review some of those essentials, mainly to introduce the notation we use. The presentation is distinguished by a chain of sometimes not obvious statements made in sequence. That is, nothing is proved in this appendix.

Contents

A.1 Vector spaces and subspaces

We will work with the ***real numbers***, denoted $\mathbb{R}$, and with the ***complex numbers***, denoted $\mathbb{C}$. We follow the convention of mathematicians rather than of electrical engineers and denote $i = \sqrt{-1}$. If we are in a situation where we wish to refer to either $\mathbb{R}$ or $\mathbb{C}$, we will write $\mathbb{F}$. $\mathbb{R}_+$ denotes the set of positive real numbers.

A ***vector space*** over $\mathbb{R}$ is a set V with two operations: (1) vector addition, denoted $v_1 + v_2 \in V$ for $v_1, v_2 \in V$, and (2) scalar multiplication, denoted $a\,v \in V$ for $a \in \mathbb{R}$ and $v \in V$. Vector addition must satisfy the rules

1. $v_1 + v_2 = v_2 + v_1$ (***commutativity***);
2. $v_1 + (v_2 + v_3) = (v_1 + v_2) + v_3$ (***associativity***);
3. there exists a unique vector $0 \in V$ with the property that $v + 0 = v$ for every $v \in V$ (***zero vector***), and
4. for every $v \in V$ there exists a unique vector $-v \in V$ such that $v + (-v) = 0$ (***negative vector***),

and scalar multiplication must satisfy the rules

5. $a(b\,v) = (ab)v$ (***associativity***);

6. $1\,v = v$;
7. $a(v_1 + v_2) = a\,v_1 + a\,v_2$ (***distributivity***);
8. $(a_1 + a_2)v = a_1\,v + a_2\,v$ (***distributivity*** again).

One can also consider vector spaces over $\mathbb{C}$, and on occasion we will do so, but not frequently. Thus we shall simply call a vector space over $\mathbb{R}$ a "vector space" when no confusion will arise from our doing so.

The vector space of primary importance to us is the collection $\mathbb{R}^n$ of n-tuples of real numbers. Thus an element of $\mathbb{R}^n$ is written $(x_1, \dots, x_n)$. We will write this also as a column vector:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Vector addition and scalar multiplication in $\mathbb{R}^n$ are done component-wise:

$$(x_1, \dots, x_n) + (y_1, \dots, y_n) = (x_1 + y_1, \dots, x_n + y_n), \quad a(x_1, \dots, x_n) = (ax_1, \dots, ax_n).$$

We shall sometimes wish to work with $\mathbb{C}^n$, the set of n-tuples of complex numbers. This can be thought of as a vector space over either $\mathbb{R}$ or $\mathbb{C}$ with the component-wise operations. We denote by $\boldsymbol{x}$ the vector with components $(x_1, \dots, x_n)$. Thus vectors are generally denoted with lowercase bold letters. We simply denote the zero vector by $\mathbf{0}$.

A subset U of a vector space V is a ***subspace*** if $u_1 + u_2 \in U$ for all $u_1, u_2 \in U$ and if $a\,u \in U$ for all $a \in \mathbb{R}$ and all $u \in U$. Note that a subspace of a vector space is itself a vector space.

A.2 Linear independence and bases

Let V be a vector space. A collection of vectors $\{v_1, \dots, v_n\} \subset V$ is ***linearly independent*** if the equality

$$c_1 v_1 + \cdots + c_n v_n = 0$$

holds only for $c_1 = \cdots = c_n = 0$. A collection of vectors $\{v_1, \dots, v_n\}$ ***spans*** V if for every vector $v \in V$ there exists constants $c_1, \dots, c_n \in \mathbb{R}$ so that

$$v = c_1 v_1 + \cdots + c_n v_n.$$

The subset $\{v_1, \dots, v_n\}$ is a ***basis*** for V if it is linearly independent and spans V. Note that the number of basis vectors for a vector space V is a constant independent of the choice of basis. This constant is the ***dimension*** of V, denoted $\dim(V)$. It is possible that a vector space will not have a finite basis. In such cases it is said to be ***infinite-dimensional***. We may talk in particular about bases for $\mathbb{R}^n$. The ***standard basis*** for $\mathbb{R}^n$ is given by the n vectors $\{\boldsymbol{e}_1 = (1, 0, \dots, 0), \boldsymbol{e}_2 = (0, 1, \dots, 0), \dots, \boldsymbol{e}_n = (0, 0, \dots, 1)\}$.

Given a basis $\{v_1, \dots, v_n\}$ for V and a vector $v \in V$, there is a unique collection of constants $c_1, \dots, c_n \in \mathbb{R}$ so that

$$v = c_1 v_1 + \cdots + c_n v_n.$$

These are the ***components*** of v in the given basis.

For vectors $v_1, \ldots, v_k \in V$ we define the ***span*** of these vectors to be the subspace

$$\operatorname{span}(v_1, \ldots, v_k) = \{c_1 v_1 + \cdots + c_k v_k \mid c_1, \ldots, c_k \in \mathbb{R}\}.$$

Note that this does not require that the vectors be linearly independent. However, if the vectors $\{v_1, \ldots, v_k\}$ are linearly independent, then they form a basis for $\operatorname{span}(v_1, \ldots, v_k)$.

A.3 Matrices and linear maps

There are close relationships between matrices and linear maps. However, let us break up our discussion to cover some matrix-specific topics.

A.3.1 Matrices

The set of $n \times m$ matrices we denote by $\mathbb{R}^{n\times m}$. We denote by $\boldsymbol{A}$ the matrix with components

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm} \end{bmatrix}.$$

Thus matrices are generally denoted with uppercase bold letters. At times we will think of vectors as $n \times 1$ matrices. We denote the $n \times m$ matrix of zeros by $\mathbf{0}_{n,m}$. If $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ then we define the ***transpose*** of $\boldsymbol{A}$, denoted $\boldsymbol{A}^t$, to be the matrix whose rows are the columns of $\boldsymbol{A}$. Thus

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \quad \Longrightarrow \quad \boldsymbol{A}^t = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}.$$

If $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times m}$ then we may multiply these to get $\boldsymbol{AB} \in \mathbb{R}^{n\times m}$. The (i, j)th element of $\boldsymbol{AB}$ is

$$\sum_{k=1}^{p} a_{ik} b_{kj}.$$

Of special interest are the $n \times n$ matrices, i.e., the "square" matrices. $\boldsymbol{I}_n$ denotes the $n \times n$ ***identity matrix***, i.e., the matrix whose entries are all zero, except for 1's on the diagonal, and $\mathbf{0}_n$ denotes the $n \times n$ matrix of zeros. The ***trace*** of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, denoted $\operatorname{tr}(\boldsymbol{A})$, as the sum of the diagonal elements of $\boldsymbol{A}$:

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}.$$

Also useful is the ***determinant*** of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, denoted $\det \boldsymbol{A}$. Let us recall the inductive definition. The determinant of a 1×1 matrix $[a]$ is simply a. The determinant of a 2×2 matrix is defined by

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Now we indicate how to compute the determinant of an $n \times n$ matrix provided one knows how to compute the determinant of an $(n-1) \times (n-1)$ matrix. One does this as follows. For a fixed $i \in \{1, \dots, n\}$, let $a_1, \dots, a_n$ be the components of the ith row of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Also let $\boldsymbol{A}(\hat{\imath}, \hat{\jmath})$ be the $(n-1) \times (n-1)$ matrix obtained by deleting the ith row and the jth column of $\boldsymbol{A}$. With this notation we define

$$\det \boldsymbol{A} = (-1)^{i+j} a_j \det \boldsymbol{A}(\hat{\imath}, \hat{\jmath}).$$

Thus, to compute the determinant of an $n \times n$ matrix, one must compute the determinant of n matrices of size $(n-1) \times (n-1)$.

There is another definition of the determinant that we will use in Section 6.1. Let S_n be the collection of permutations of $(1, \dots, n)$. We denote an element $\sigma \in S_n$ by indicating what it does to each element in the sequence $(1, \dots, n)$ like so:

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{pmatrix}.$$

A ***transposition*** is a permutation that consists of the swapping of two elements of $(1, \dots, n)$. A permutation is ***odd*** (resp. ***even***) if it is the composition of an odd (resp. even) number of transpositions. We define $\operatorname{sgn} \colon S_n \to \{-1, 1\}$ by

$$\operatorname{sgn}(\sigma) = \begin{cases} 1, & \sigma \text{ is even} \\ -1, & \sigma \text{ is odd.} \end{cases}$$

With this notation, it can be shown that

$$\det \boldsymbol{A} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n}. \tag{A.1}$$

If and only if $\det \boldsymbol{A} \neq 0$ there exists $\boldsymbol{A}^{-1} \in \mathbb{R}^{n \times n}$ so that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$. In this case we say $\boldsymbol{A}$ is ***invertible*** of ***nonsingular***. If $\boldsymbol{A}$ is not invertible, it is ***singular***.

We shall require the cofactor matrix of an $n \times n$ matrix $\boldsymbol{A}$. If $i, j \in \{1, \dots, n\}$, define the ***(i,j)th cofactor*** to be

$$C_{ij} = (-1)^{i+j} \det \boldsymbol{A}(\hat{\imath}, \hat{\jmath}).$$

The matrix $\operatorname{Cof}(\boldsymbol{A})$ is then the $n \times n$ matrix whose (i,j)th element is C_{ij}. One then verifies that

$$\boldsymbol{A}(\operatorname{Cof}(\boldsymbol{A}))^t = (\det \boldsymbol{A}) \boldsymbol{I}_n.$$

The matrix $(\operatorname{Cof}(\boldsymbol{A}))^t$ we denote $\operatorname{adj}(\boldsymbol{A})$, and we note that if $\boldsymbol{A}$ is invertible, then

$$\boldsymbol{A}^{-1} = \frac{1}{\det \boldsymbol{A}} \operatorname{adj}(\boldsymbol{A}). \tag{A.2}$$

We call $\operatorname{adj}(\boldsymbol{A})$ the ***adjugate*** of $\boldsymbol{A}$.

Let us next consider the linear equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ which we wish to solve for a given $\boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. We will only look at the easy case where $\boldsymbol{A}$ is invertible. If for $i \in \{1, \dots, n\}$, $\boldsymbol{A}(\boldsymbol{b}, i)$ denotes the matrix $\boldsymbol{A}$ but with the ith column replaced with $\boldsymbol{b}$, then ***Cramer's Rule*** states that the ith component of the solution vector $\boldsymbol{x}$ is given by

$$x_i = \frac{\det \boldsymbol{A}(\boldsymbol{b}, i)}{\det \boldsymbol{A}}.$$

More general than a cofactor of $\boldsymbol{A}$ is the notion of a minor. A ***kth-order minor*** of $\boldsymbol{A}$ is the determinant of a $k \times k$ matrix obtained from $\boldsymbol{A}$ by removing any collection of $n-k$ rows and $n-k$ columns. The ***principal minors*** of $\boldsymbol{A}$ are the n determinants of the upper left $k \times k$ blocks of $\boldsymbol{A}$ for $k \in \{1, \dots, n\}$.

A.3.2 Some useful matrix lemmas

The following results will be useful to us, and the proofs are simple enough to give here.

A.1 Lemma *If $\boldsymbol{A}$ is an invertible $k \times k$ matrix, $\boldsymbol{B}$ is a $k \times (n-k)$ matrix, $\boldsymbol{C}$ is a $(n-k) \times k$ matrix, and $\boldsymbol{D}$ is an $(n-k) \times (n-k)$ matrix, then*

$$\det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \det \boldsymbol{A} \det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}).$$

Proof We observe that

$$\begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0}_{k,(n-k)} \\ -\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{I}_{n-k} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{0}_{(n-k),k} & \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} \end{bmatrix}.$$

The determinant of the leftmost matrix is 1, and since the determinant of an upper block diagonal matrix is the determinant of the diagonal blocks, the determinant of the rightmost matrix is $\det \boldsymbol{A} \det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})$. From this the result follows. ■

A.2 Lemma *If $\boldsymbol{A}$ is an invertible $k \times k$ matrix, $\boldsymbol{B}$ is a $k \times (n-k)$ matrix, $\boldsymbol{C}$ is a $(n-k) \times k$ matrix, and $\boldsymbol{D}$ is an $(n-k) \times (n-k)$ matrix, then*

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{U} & \boldsymbol{V} \end{bmatrix},$$

where

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \\ \boldsymbol{Y} &= -\boldsymbol{A}^{-1}\boldsymbol{B}((\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ \boldsymbol{U} &= -(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \\ \boldsymbol{V} &= (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}, \end{aligned}$$

provided that the inverse exists.

Proof We perform row operations on

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{0} & \boldsymbol{I}_{n-k} \end{bmatrix}.$$

Multiply the first k rows by $\boldsymbol{A}^{-1}$ to get

$$\begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{A}^{-1}\boldsymbol{B} & \boldsymbol{A}^{-1} & \boldsymbol{0} \\ \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{0} & \boldsymbol{I}_{n-k} \end{bmatrix}.$$

Now subtract from the second $n-k$ rows the first k rows multiplied by $\boldsymbol{C}$ to get

$$\begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{A}^{-1}\boldsymbol{B} & \boldsymbol{A}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} & -\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{I}_{n-k} \end{bmatrix}.$$

Multiply the second $n-k$ rows by $(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}$ times the first k rows to get

$$\begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{A}^{-1}\boldsymbol{B} & \boldsymbol{A}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-k} & -(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \end{bmatrix}.$$

Finally, subtract from the second $n-k$ rows $\boldsymbol{A}^{-1}\boldsymbol{B}$ times the first k rows to yield the result. ■

A.3 Lemma *If* $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ *and* $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ *then* $\boldsymbol{I}_n - \boldsymbol{AB}$ *is nonsingular provided that* $\boldsymbol{I}_m - \boldsymbol{BA}$ *is nonsingular, and furthermore the relation*

$$\boldsymbol{A}(\boldsymbol{I}_m - \boldsymbol{BA})^{-1} = (\boldsymbol{I}_n - \boldsymbol{AB})^{-1}\boldsymbol{A},$$

holds in this case.

Proof We first show that $\det(\boldsymbol{I}_n - \boldsymbol{AB}) = \det(\boldsymbol{I}_m - \boldsymbol{BA})$. First we note that the matrix

$$\begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{A} \\ \boldsymbol{B} & \boldsymbol{I}_m \end{bmatrix}$$

is transformed to the matrix

$$\begin{bmatrix} \boldsymbol{B} & \boldsymbol{I}_m \\ \boldsymbol{I}_n & \boldsymbol{A} \end{bmatrix}$$

by k row switches, for some suitable k. Now this last matrix can be transformed to

$$\begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{B} \\ \boldsymbol{A} & \boldsymbol{I}_n \end{bmatrix}$$

by k column switches, for the same k as was used to make the row switches. Note that each row and column switch changes the determinant by a factor of -1. Thus we have, also employing Lemma A.1,

$$\det(\boldsymbol{I}_m - \boldsymbol{BA}) = \det \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{A} \\ \boldsymbol{B} & \boldsymbol{I}_m \end{bmatrix} = \det \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{B} \\ \boldsymbol{A} & \boldsymbol{I}_n \end{bmatrix} = \det(\boldsymbol{I}_n - \boldsymbol{AB}),$$

as desired. This shows that $\boldsymbol{I}_n - \boldsymbol{AB}$ is nonsingular if and only if $\boldsymbol{I}_m - \boldsymbol{BA}$ is nonsingular. Now we make a simple computation,

$$\begin{aligned} & \boldsymbol{A} - \boldsymbol{ABA} = \boldsymbol{A} - \boldsymbol{ABA} \\ \implies \quad & \boldsymbol{A}(\boldsymbol{I}_m - \boldsymbol{BA}) = (\boldsymbol{I}_n - \boldsymbol{AB})\boldsymbol{A} \\ \implies \quad & (\boldsymbol{I}_n - \boldsymbol{AB})^{-1}\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{I}_m - \boldsymbol{BA})^{-1}, \end{aligned}$$

as desired. ■

A.4 Lemma *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, *and* $\boldsymbol{C} \in \mathbb{R}^{m\times n}$. *If* $\boldsymbol{A}$ *and* $\boldsymbol{A} + \boldsymbol{BC}$ *are invertible then* $\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B}$ *is invertible and*

$$(\boldsymbol{A} + \boldsymbol{BC})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1}.$$

Proof We compute

$$\begin{aligned} & \big(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1}\big)(\boldsymbol{A} + \boldsymbol{BC}) \\ &= \boldsymbol{I}_n + \boldsymbol{A}^{-1}\boldsymbol{BC} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1}\boldsymbol{BC} \\ &= \boldsymbol{I}_n + \boldsymbol{A}^{-1}\boldsymbol{B}\big(\boldsymbol{I}_m - (\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1} - (\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CA}^{-1}\boldsymbol{B}\big)\boldsymbol{C} \\ &= \boldsymbol{I}_n + \boldsymbol{A}^{-1}\boldsymbol{B}\big(\boldsymbol{I}_m - (\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B})^{-1}(\boldsymbol{I}_m + \boldsymbol{CA}^{-1}\boldsymbol{B})\big)\boldsymbol{C} \\ &= \boldsymbol{I}_n + \boldsymbol{A}^{-1}\boldsymbol{B}\big(\boldsymbol{I}_m - \boldsymbol{I}_m\big)\boldsymbol{C} \\ &= \boldsymbol{I}_n. \end{aligned}$$

This gives the lemma by uniqueness of the inverse. ■

A.5 Lemma *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ and let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$. If $\boldsymbol{M}$ and $\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t$ are invertible we have*

$$\begin{aligned}\operatorname{adj}(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)\boldsymbol{u} &= (\operatorname{adj}\boldsymbol{M})\boldsymbol{u}\\ \boldsymbol{v}^t\operatorname{adj}(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t) &= \boldsymbol{v}^t(\operatorname{adj}\boldsymbol{M}).\end{aligned}$$

Proof Let us show the first equality, and the second follows in much the same manner. Using Lemma A.4 we compute

$$\begin{aligned}(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)^{-1}\boldsymbol{u} &= \frac{\operatorname{adj}(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)}{\det(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)}\boldsymbol{u}\\ &= \boldsymbol{M}^{-1}\boldsymbol{u} - \frac{\boldsymbol{M}^{-1}\boldsymbol{u}\boldsymbol{v}^t\boldsymbol{M}^{-1}\boldsymbol{u}}{1 + \boldsymbol{v}^t\boldsymbol{M}\boldsymbol{u}}\\ &= \frac{\boldsymbol{M}^{-1}\boldsymbol{u}}{1 + \boldsymbol{v}^t\boldsymbol{M}\boldsymbol{u}}.\end{aligned}$$

We also have

$$\boldsymbol{M}^{-1} = \frac{\operatorname{adj}\boldsymbol{M}}{\det \boldsymbol{M}}$$

and, by Lemma A.3.

$$\begin{aligned}1 + \boldsymbol{v}^t\boldsymbol{M}^{-1}\boldsymbol{u} &= \det(1 + \boldsymbol{v}^t\boldsymbol{M}^{-1}\boldsymbol{u})\\ &= \det(\boldsymbol{I}_n + \boldsymbol{M}^{-1}\boldsymbol{u}\boldsymbol{v}^t)\\ &= \frac{\det(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)}{\det \boldsymbol{M}}.\end{aligned}$$

This then gives

$$\frac{\boldsymbol{M}^{-1}\boldsymbol{u}}{1 + \boldsymbol{v}^t\boldsymbol{M}\boldsymbol{u}} = \frac{(\operatorname{adj}\boldsymbol{M})\boldsymbol{u}}{\det \boldsymbol{M}}\frac{\det \boldsymbol{M}}{\det(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)} = \frac{(\operatorname{adj}\boldsymbol{M})\boldsymbol{u}}{\det(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)}.$$

Thus

$$(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)^{-1}\boldsymbol{u} = \frac{(\operatorname{adj}\boldsymbol{M})\boldsymbol{u}}{\det(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)},$$

from which we deduce that $\operatorname{adj}(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^t)\boldsymbol{u} = (\operatorname{adj}\boldsymbol{M})\boldsymbol{u}$, as desired. ■

A.3.3 Linear maps

Let U and V be vector spaces. A ***linear map*** from U to V is a map $L\colon U \to V$ satisfying $L(u_1 + u_2) = L(u_1) + L(u_2)$ for all $u_1, u_2 \in U$ and $L(a\,u) = a\,L(u)$ for all $a \in \mathbb{R}$ and $u \in U$. If $U = V$ then L is a ***linear transformation***. A special type of linear map, one from the vector space $\mathbb{R}^m$ to the vector space $\mathbb{R}^n$, is defined by an $n \times m$ matrix $\boldsymbol{A}$. If $\boldsymbol{x} \in \mathbb{R}^m$ then $\boldsymbol{A}(\boldsymbol{x}) \in \mathbb{R}^n$ is defined by its ith component being

$$\sum_{j=1}^{n} A_{ij}x_j.$$

Typically we will write $\boldsymbol{A}\boldsymbol{x}$ for $\boldsymbol{A}(\boldsymbol{x})$.

Let $L\colon U \to V$ be a linear map between finite-dimensional vector spaces, and let $\{u_1, \dots, u_m\}$ be a basis for U and $\{v_1, \dots, v_n\}$ be a basis for V. For $i \in \{1, \dots, m\}$ we may write

$$L(u_i) = a_{1i}v_1 + \cdots + a_{ni}v_n$$

for some uniquely defined $a_{1i}, \dots, a_{ni} \in \mathbb{R}$. The nm numbers $a_{\ell i}$, $\ell = 1, \dots, n$, $i = 1, \dots, m$ are called the ***components*** of L relative to the given bases.

If $L\colon U \to V$ is a linear map, we define subspaces

$$\ker(L) = \{u \in U \mid L(u) = 0\} \subset U$$
$$\text{image}(L) = \{L(u) \mid u \in U\} \subset V$$

which we call the ***kernel*** and ***image*** of L, respectively. You may know these as "nullspace" and "range." The ***Rank-Nullity Theorem*** states that $\dim(ker(L)) + \dim(\text{image}(L)) = \dim(U)$. A linear map $L\colon U \to V$ is ***injective*** if the equality $L(u_1) = L(u_2)$ implies that $u_1 = u_2$. L is ***surjective*** if for each $v \in V$ there exists $u \in U$ so that $L(u) = v$. If L is both injective and surjective then it is ***bijective*** or ***invertible***. If L is invertible then there exists a linear map $L^{-1}\colon V \to U$ so that $L \circ L^{-1} = \text{id}_V$ and $L^{-1} \circ L = \text{id}_U$, where id denotes the identity map.

If $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ and we write $\boldsymbol{A}$ using its columns as

$$\boldsymbol{A} = \left[\ \boldsymbol{a}_1 \ \middle|\ \cdots \ \middle|\ \boldsymbol{a}_m \ \right].$$

Thus each of the vectors $\boldsymbol{a}, \dots, \boldsymbol{a}_m$ are in $\mathbb{R}^n$. We define the ***columnspace*** of $\boldsymbol{A}$ to be the subspace of $\mathbb{R}^n$ given by $\text{span}(\boldsymbol{a}, \dots, \boldsymbol{a}_m)$. We observe that the columnspace and the image of $\boldsymbol{A}$, thought of as a linear map, coincide.

If $L\colon V \to V$ is a linear transformation, a subspace $U \subset V$ is ***L-invariant*** if $L(u) \in U$ for every $u \in U$. Suppose that V is finite-dimensional and that U is an invariant subspace for L and let $\{v_1, \dots, v_n\}$ be a basis for $\mathbb{R}^n$ with the property that $\{v_1, \dots, v_k\}$ is a basis for U. Since U is L-invariant we must have

$$\begin{aligned}
L(v_1) &= a_{11}v_1 + \cdots + a_{k1}v_k \\
&\vdots \\
L(v_k) &= a_{1k}v_1 + \cdots + a_{kk}v_k \\
L(v_{k+1}) &= a_{1,k+1}v_1 + \cdots + a_{k,k+1}v_k + a_{k+1,k+1}v_{k+1} + \cdots + a_{n,k+1}v_n \\
&\vdots \\
L(v_n) &= a_{1n}v_1 + \cdots + a_{kn}v_k + a_{k+1,n}v_{k+1} + \cdots + a_{nn}v_n.
\end{aligned}$$

Thus the matrix representation for L in this basis has the form

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{n-k,k} & \boldsymbol{A}_{22} \end{bmatrix}. \tag{A.3}$$

A.4 Change of basis

One on occasion wishes to ascertain how the components of vectors and linear maps change when one changes basis. We restrict our consideration to linear transformations. So let V be a vector space with $\{v_1, \dots, v_n\}$ and $\{\tilde{v}_1, \dots, \tilde{v}_n\}$ two bases for V. Since these are both bases, there exists an invertible $n \times n$ matrix $\boldsymbol{T}$ so that

$$v_i = \sum_{j=1}^{n} t_{ji}\tilde{v}_j, \quad i = 1, \dots, n.$$

$\boldsymbol{T}$ is called the ***change of basis matrix***. For $v \in V$, let $c_1, \dots, c_n$ and $\tilde{c}_1, \dots, \tilde{c}_n$ be the components of v in the respective bases $\{v_1, \dots, v_n\}$ and $\tilde{v}_1, \dots, \tilde{v}_n\}$. One readily determines that $\tilde{\boldsymbol{c}} = \boldsymbol{T}\boldsymbol{c}$. Also, let $L\colon V \to V$ be a linear transformation and let a_{ij}, $i, j = 1, \dots, n$, and $\tilde{a}_{ij}$, $i, j = 1, \dots, n$, be the components of L in the respective bases $\{v_1, \dots, v_n\}$ and $\tilde{v}_1, \dots, \tilde{v}_n\}$. A simple computation using the definition of components shows that $\tilde{\boldsymbol{A}} = \boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}$. We shall most frequently encounter the change of basis in the scenario when $V = \mathbb{R}^n$, $\{v_1 = \boldsymbol{e}_1, \dots, v_n = \tilde{e}_n\}$, and $\{\tilde{v}_1 = \boldsymbol{x}_1, \dots, \tilde{v}_n = \boldsymbol{x}_n\}$. Thus we are changing from the standard basis for $\mathbb{R}^n$ to some nonstandard basis. In this case, the change of basis matrix is simply defined by

$$\boldsymbol{T}^{-1} = \left[\ \boldsymbol{x}_1 \ \middle|\ \cdots \ \middle|\ \boldsymbol{x}_n\ \right].$$

The transformation that sends $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ to the matrix $\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}$ is called a ***similarity transformation***. One readily verifies both trace and determinant are unchanged under similarity transformations. That is, for any invertible $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ we have

$$\operatorname{tr}(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}) = \operatorname{tr}(\boldsymbol{A}), \quad \det(\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1}) = \det(\boldsymbol{A}).$$

For this reason, the notion of trace and determinant can be applied to any linear transformation by simply defining them in an arbitrary basis.

A.5 Eigenvalues and eigenvectors

For a linear transformation $L\colon V \to V$, if we have a pair $(\lambda, v) \in \mathbb{R} \times (V \setminus \{0\})$ which satisfies $L(v) = \lambda\, v$ then we say λ_0 is an ***eigenvalue*** for L with ***eigenvector*** v. The collection of all eigenvectors for λ, along with the zero vector, forms the ***eigenspace*** for λ. The eigenspace is easily seen to be a subspace. When V is finite-dimensional, the matter of computing eigenvalues and eigenvectors is most easily done in a basis $\{v_1, \dots, v_n\}$ where L is represented by its component matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Thus the problem is reduced to finding $(\lambda, \boldsymbol{x}) \in \mathbb{R} \times (\mathbb{R}^n \setminus \{\boldsymbol{0}\})$ satisfying $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. This implies that $\boldsymbol{A} - \lambda\boldsymbol{I}_n$ must be singular. Thus eigenvalues must be roots of the ***characteristic polynomial***

$$P_{\boldsymbol{A}}(\lambda) = \det(\lambda\boldsymbol{I}_n - \boldsymbol{A}).$$

Such roots may well be complex, and recall that a complex root of a real polynomial always occurs along with its complex conjugate. We denote by

$$\operatorname{spec}(\boldsymbol{A}) = \{\lambda \in \mathbb{C} \mid \ \lambda \text{ is an eigenvalue for } A\},$$

which we call the ***spectrum*** of $\boldsymbol{A}$. If λ is an eigenvalue of $\boldsymbol{A}$, the ***algebraic multiplicity*** of λ_0, denoted $m_a(\lambda_0)$, is the largest integer k for which we can write $P_{\boldsymbol{A}}(\lambda) = (\lambda - \lambda_0)^k Q(\lambda)$ for some polynomial $Q(\lambda)$ satisfying $Q(\lambda_0) \neq 0$. The ***geometric multiplicity*** of λ_0 is the maximum number of linearly independent eigenvectors possessed by λ_0. Recall that $m_g(\lambda_0) \leq m_a(\lambda_0)$. If we have a real matrix $\boldsymbol{A}$, we will often talk about multiplicities thinking of $\boldsymbol{A}$ as a complex matrix, since $\boldsymbol{A}$ may very well have complex eigenvalues. Note that the above discussion of finding eigenvalues for L using its matrix of components in a basis is obviously a basis independent operation.

Also recall the ***Cayley-Hamilton Theorem*** which says that a matrix satisfies its own characteristic polynomial. That is to say, if

$$P_{\boldsymbol{A}}(\lambda) = \lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0$$

then

$$\boldsymbol{A}^n + p_{n-1}\boldsymbol{A}^{n-1} + \cdots + p_1\boldsymbol{A} + p_0\boldsymbol{I}_n = \boldsymbol{0}_n.$$

You'll also recall that $\operatorname{spec}(\boldsymbol{A}) = \operatorname{spec}(\boldsymbol{A}^t)$, but that the eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{A}^t$ will generally differ.

A linear transformation $L\colon V \to V$ on a finite-dimensional vector space is ***diagonalisable*** if it possesses a basis of eigenvectors. One can easily check that if L is diagonalisable then there exists a basis $\{v_1, \ldots, v_n\}$ for V in which the matrix of components for L is diagonal. Note that a matrix with complex eigenvalues can never be diagonalised if V is a $\mathbb{R}$-vector space.

A.6 Inner products

An ***inner product*** on a $\mathbb{R}$-vector space V assigns to a pair of vectors $v_1, v_2 \in V$ a number $\langle v_1, v_2\rangle \in \mathbb{R}$, and the assignment satisfies

1. $\langle v_1, v_2\rangle = \langle v_2, v_1\rangle$ (***symmetry***),
2. $\langle a_1\, v_1 + a_2\, v_2, v\rangle = a_1\langle v_1, v\rangle + a_2\langle v_2, v\rangle$ (***bilinearity***), and
3. $\langle v, v\rangle = 0$ if and only if $v = 0$ (***positive-definiteness***).

On a $\mathbb{C}$-vector space, we do things a little differently. If V is a $\mathbb{C}$-vector space, a ***Hermitian inner product*** is an assignment $(v_1, v_2) \mapsto \langle v_1, v_2\rangle \in \mathbb{C}$ with the assignment now satisfying

1. $\langle v_1, v_2\rangle = \overline{\langle v_2, v_1\rangle}$ (***symmetry***),
2. $\langle a_1\, v_1 + a_2\, v_2, v\rangle = \bar{a}_1\langle v_1, v\rangle + \bar{a}_2\langle v_2, v\rangle$ (***bilinearity***), and
3. $\langle v, v\rangle = 0$ if and only if $v = 0$ (***positive-definiteness***).

Here $\bar{z}$ denotes the complex conjugate of $z \in \mathbb{C}$. On $\mathbb{R}^n$ there is the ***standard inner product*** defined by

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^n x_i y_i,$$

i.e., the "dot product." For $\mathbb{C}^n$ the ***standard Hermitian inner product*** is defined by

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^n \bar{x}_i y_i.$$

Given $\mathbb{R}$-inner product spaces U and V and a linear map $L\colon U \to V$, we define the ***adjoint*** of L as the linear map $L^*\colon V \to U$ satisfying

$$\langle L^*(v), u\rangle_U = \langle v, L(u)\rangle_V, \quad u \in U,\ v \in V.$$

If $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ is thought of as a linear map between the $\mathbb{R}$-vector spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ with their standard inner products, then the adjoint of $\boldsymbol{A}$ is simply the standard transpose: $\boldsymbol{A}^* = \boldsymbol{A}^t$. If $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is thought of as a linear map between the $\mathbb{R}$-vector spaces $\mathbb{C}^m$ and $\mathbb{C}^n$ with their standard Hermitian inner products, then the adjoint of $\boldsymbol{A}$ is the standard transpose with each element of the matrix additional conjugated: $\boldsymbol{A}^* = \bar{\boldsymbol{A}}^t$. Given a linear map $L\colon U \to V$ between inner product spaces (or Hermitian inner product spaces), a ***singular value*** for L is an eigenvalue of $L \circ L^*\colon V \to V$.

A linear transformation $L\colon V \to V$ of a $\mathbb{R}$-inner product space (resp. a $\mathbb{C}$-Hermitian inner product space) is ***symmetric*** (resp. ***Hermitian***) if $L = L^*$. One easily deduces

that the eigenvalues of a symmetric or Hermitian linear transformation are always real. If $V = \mathbb{R}^n$ so that L is an $n \times n$ matrix, then symmetry of L is exactly the condition that L be a symmetric matrix. In this case, there exists an orthogonal matrix $\boldsymbol{R}$ so that the matrix $\boldsymbol{R}L\boldsymbol{R}^t$ is diagonal.

A linear transformation $L\colon V \to V$ of a $\mathbb{R}$-inner product space (resp. a $\mathbb{C}$-Hermitian inner product space) is ***skew-symmetric*** (resp. ***skew-Hermitian***) if $L = -L^*$. One readily checks that any transformation L on a $\mathbb{R}$-inner product space is a sum of its ***symmetric part*** $\frac{1}{2}(L + L^*)$ and its ***skew-symmetric part*** $\frac{1}{2}(L - L^*)$.

On an inner product space or a Hermitian inner product space V we also have defined a ***norm*** which assigns to a vector $v \in V$ a real number defined by $\|v\| = \sqrt{\langle v, v\rangle}$. A norm may be verified to satisfy the ***triangle inequality***:

$$\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|.$$

Exercises

EA.1 Let $V = \text{span}((1, 2, 0, 0), (0, 1, 0, 1)) \subset \mathbb{R}^4$. Construct a 4×4 matrix that leaves the subspace V invariant.

Appendix B

Ordinary differential equations

In this appendix we provide a very quick review of differential equations at the level prerequisite to use this book. The objective, as in the other appendices, is not to provide a complete overview, but to summarise the main points. The text itself covers various aspects of differential equations theory beyond what we say here; indeed, it is possible to think of control theory as a sub-discipline of differential equations, although I do not like to do so. We deal with scalar equations in Section B.1, and with systems of equations, using the matrix exponential, in Section B.2.

Contents

B.1 Scalar ordinary differential equations

Let us consider a differential equation of the form

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) = u(t). \tag{B.1}$$

Here $y^{(i)}(t) = \frac{\mathrm{d}^i y(t)}{\mathrm{d}t^i}$. and $u(t)$ is a known function of t. In the text, we will establish various methods for solving such equations for general functions u. Students are expected to be able to solve simple examples of such equations with alacrity. Let us recall how this is typically done.

First one obtains a solution, called a ***homogeneous solution*** and denoted here by y_h, to the equation

$$y_h^{(n)}(t) + p_{n-1}y_h^{(n-1)}(t) + \cdots + p_1 y_h^{(1)}(t) + p_0 y_h(t) = 0.$$

To do this, one seeks solutions of the form $e^{\lambda t}$ for $\lambda \in \mathbb{C}$. Substitution into the homogeneous equations then gives the polynomial

$$\lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0 = 0.$$

This is the ***characteristic equation*** and it will have n roots, counting multiplicities. Thus it will have roots $\lambda_1, \ldots, \lambda_\ell$, with respective multiplicities $k_1, \ldots, k_\ell$, with $k_1 + \cdots + k_\ell = n$. Corresponding to a root λ_a, one constructs k_a linearly independent solutions if the root is real, and $2k_a$ linearly independent solutions if the root is complex (a complex root and its conjugate yield the same $2k_a$ solutions). These solutions are constructed as follows.

B.1 Construction of solutions corresponding to a root Let $\lambda \in \mathbb{C}$ be a root of the characteristic equation of multiplicity k.

1. If $\lambda \in \mathbb{R}$ then k linearly independent solutions are

$$y_1(t) = e^{\lambda t},\ y_2(t) = te^{\lambda t},\ \dots,\ y_k(t) = t^{k-1}e^{\lambda t}.$$

2. If $\lambda \notin \mathbb{R}$ write $\lambda = \sigma + i\omega$. Then $2k$ linearly independent solutions are

$$y_1(t) = e^{\sigma t}\sin\omega t,\ y_2(t) = e^{\sigma t}\cos\omega t,\ y_3(t) = te^{\sigma t}\sin\omega t,\ y_4(t) = te^{\sigma t}\cos\omega t,\ \dots,$$
$$y_{2k-1}(t) = t^{k-1}e^{\sigma t}\sin\omega t,\ y_{2k}(t) = t^{k-1}e^{\sigma t}\cos\omega t.$$

These are the solutions corresponding to the root λ. •

Applying the construction to all roots of the characteristic equation yields n linearly independent solutions $y_1(t), \dots, y_n(t)$ to the homogeneous equation. Furthermore, *every* solution of the homogeneous equation is a linear combination of these n linearly independent solutions. Thus we take y_h to be a general linear combination of the n linearly independent solutions, with at the moment unspecified coefficients:

$$y_h(t) = c_1 y_1(t) + \cdots + c_n y_n(t).$$

Next one determines what is called a ***particular solution***, which we shall denote y_p. A particular solution is, by definition, *any* solution to the differential equation (B.1). Of course, one cannot expect to arrive at such a solution in a useful way for arbitrary right-hand side functions u. What we shall describe here is an essentially *ad hoc* procedure, useful for certain types of functions u. This procedure typically goes under the name of the ***method of undetermined coefficients***. The idea is that one uses the fact that certain types of functions of time—polynomial functions, trigonometric functions, and exponential functions—have their form unchanged by the act of differentiation. Thus if one takes y_p to be of this form, then

$$y_p^{(n)}(t) + p_{n-1}y_p^{(n-1)}(t) + \cdots + p_1 y_p^{(1)}(t) + p_0 y_p(t) \tag{B.2}$$

will also have the same form. Therefore, one hopes to be able to determine y_p by "comparing" u to the expression (B.2). This hope is founded, and leads to the following procedure.

B.2 Method of undetermined coefficients Suppose that $u(t) = t^\ell e^{\sigma t}(a_1\sin\omega t + a_2\cos\omega t)$. Let k be the smallest integer with the property that $t^k y_{\text{test}}(t)$ is a not solution to the homogeneous equation, where y_{test} is any function of t of the form

$$y_{\text{test}}(t) = P(t)e^{\sigma t}(\alpha_1\sin\omega t + \alpha_2\cos\omega t),$$

where α_1 and α_2 are arbitrary constants, and P is an arbitrary nonzero polynomial of degree at most ℓ. Often we will have $k = 0$. Then seek a particular solution of the form

$$y_p(t) = t^k Q(t)e^{\sigma t}(b_1\sin\omega t + b_2\cos\omega t),$$

where Q is a polynomial of degree ℓ. Substitute into (B.1) to determine b_1, b_2, and the coefficients of Q. If u is a linear combination,

$$u(t) = \mu_1 u_1(t) + \cdots + \mu_m u_m(t),$$

of terms of the form $t^\ell e^{\sigma t}(a_1 \sin \omega t + a_2 \cos \omega t)$, then the above procedure may be applied separately to each term in the linear combination separately, yielding separate particular solutions $y_{p,1}, \dots, y_{p,m}$. Then the particular solution is

$$y_p(t) = \mu_1 y_{p,1}(t) + \cdots + \mu_m y_{p,m}(t). \qquad \bullet$$

Once one has a particular solution, the ***general solution*** is then simply the sum of the homogeneous and particular solution:

$$y(t) = y_h(t) + y_p(t) = c_1 y_1(t) + \cdots + c_n y_n(t) + y_p(t).$$

The constants $c_1, \dots, c_n$ can be determined if we have an ***initial value problem*** where the initial values $y(0), y^{(1)}(0), \dots, y^{(n-1)}(0)$ are specified.

It is worth demonstrating the procedure to this point on a simple example.

B.3 Example We consider the differential equation

$$\dddot{y}(t) + 3\ddot{y}(t) + 3\dot{y} + y(t) = te^{-t}.$$

First let us find a solution to the homogeneous equation

$$\dddot{y}_h(t) + 3\ddot{y}_h(t) + 3\dot{y}_h + y_h(t) = 0.$$

The characteristic equation is $\lambda^3 + 3\lambda^2 + 3\lambda + 1 = 0$, which has the single root $\lambda = -1$ of multiplicity 3. Our recipe for the homogeneous solution then gives

$$y_h(t) = c_1 e^{-t} + c_2 t e^{-t} + c_3 t^2 e^{-t}.$$

Now let us obtain a particular solution. By the above procedure, we take

$$y_{\text{test}} = (\beta_1 t + \beta_0) e^{-t}.$$

We calculate that $y_{\text{test}}(t)$ and ty_{test} identically solve the differential equation for any β_1 and β_0. Also, we compute that $t^2 y_{\text{test}}(t)$ satisfies the homogeneous provided that $\beta_1 = 0$. In exasperation, we compute that the only way that $t^3 y_{\text{test}}(t)$ will satisfy the homogeneous equation if and only if $\beta_1 = \beta_0 = 0$. Thus we have $k = 3$, using the notation of our procedure for obtaining particular solutions. Therefore, we seek a particular solution of the form

$$y_p(t) = t^3(q_1 t + q_0) e^{-t}.$$

Substitution into the ode gives

$$(6q_0 + 24q_1 t) e^{-t} = te^{-t},$$

from which we ascertain that $q_0 = 0$ and $q_1 = \frac{1}{24}$. Thus the particular solution is

$$y_p(t) = \frac{t^4 e^{-t}}{24}.$$

The general solution is then

$$y(t) = y_h(t) + y_p(t) = c_1 e^{-t} + c_2 t e^{-t} + c_3 t^2 e^{-t} + \frac{t^4 e^{-t}}{24}.$$

To determine the unknown constants c_1, c_2, and c_3, we use the initial conditions. One may determine that

$$c1 = y(0), \quad c_2 = y(0) + \dot{y}(0), \quad c_3 = \tfrac{1}{2}(y(0) + 2\dot{y}(0) + \ddot{y}(0)). \qquad \bullet$$

B.2 Systems of ordinary differential equations

Next we look at first-order systems of ordinary differential equations:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t), \tag{B.3}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. The possibility of an additional term on the right-hand side of the form $\boldsymbol{f}(t)$ is interesting, but this is handled in the text. The solution to (B.3) is given to us by the ***matrix exponential***. For any $n \times n$ matrix $\boldsymbol{A}$ define

$$e^{\boldsymbol{A}} = \boldsymbol{I}_n + \boldsymbol{A} + \frac{\boldsymbol{A}^2}{2!} + \cdots + \frac{\boldsymbol{A}^k}{k!} + \cdots$$

One may verify that this series converges for every $\boldsymbol{A}$, and it is a straightforward calculation to verify that $\boldsymbol{x}(t) = e^{\boldsymbol{A}t}\boldsymbol{x}_0$ is the solution to the differential equation (B.3) with the initial condition $\boldsymbol{x}(0) = \boldsymbol{x}_0$. Let us present a way to compute the matrix exponential. Corresponding to the linear system (B.3) will be n linearly independent vector solutions $\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_n(t)$. If $\boldsymbol{A}$ is diagonalisable, it is relatively easy to compute these solutions, but if $\boldsymbol{A}$ is *not* diagonalisable, there is a bit more work involved. Nonetheless, there is a recipe for doing this, and we present it here. Note that this recipe will *always* work, and it can be applied, even when $\boldsymbol{A}$ is diagonalisable—in this case it just simplifies.

First one computes the eigenvalues for $\boldsymbol{A}$. There will be n of these in total, counting algebraic multiplicities and complex conjugate pairs. One treats each eigenvalue separately. For a real eigenvalue λ_0 with algebraic multiplicity $k = m_a(\lambda_0)$, one must compute k linearly independent solutions. For a complex eigenvalue λ_0 with algebraic multiplicity $\ell = m_a(\lambda_0)$, one must compute 2ℓ linearly independent solutions, since $\bar{\lambda}_0$ is also necessarily an eigenvalue with algebraic multiplicity ℓ.

We first look at how to deal with real eigenvalues. Let λ_0 be one such object with algebraic multiplicity k. It is a fact that the matrix $(\boldsymbol{A} - \lambda_0\boldsymbol{I}_n)^k$ will have rank $n - k$, and so will have a kernel of dimension k by the Rank-Nullity Theorem. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ be a basis for $\ker((\boldsymbol{A} - \lambda_0\boldsymbol{I}_n)^k)$. We call each of these vectors a ***generalised eigenvector***. If the geometric multiplicity of λ_0 is also k, then the generalised eigenvectors will simply be the usual eigenvectors. If $m_g(\lambda_0) < m_a(\lambda_0)$ then a generalised eigenvector may or may not be an eigenvector. Corresponding to each generalised eigenvector $\boldsymbol{u}_i$, $i = 1, \ldots, k$, we will define a solution to (B.3) by

$$\boldsymbol{x}_i(t) = e^{\lambda_0 t}\exp((\boldsymbol{A} - \lambda_0\boldsymbol{I}_n)t)\boldsymbol{u}_i. \tag{B.4}$$

Note that because $\boldsymbol{u}_i$ is a generalised eigenvector, the infinite series $\exp((\boldsymbol{A} - \lambda_0\boldsymbol{I}_n)t)\boldsymbol{u}_i$ will have only a finite number of terms—at most k in fact. Indeed we have

$$\exp((\boldsymbol{A} - \lambda_0\boldsymbol{I}_n)t)\boldsymbol{u}_i = \left(\boldsymbol{I}_n + t(\boldsymbol{A} - \lambda_0\boldsymbol{I}_n) + \frac{t^2}{2!}(\boldsymbol{A} - \lambda_0\boldsymbol{I}_n)^2 + \cdots + \frac{t^{k-1}}{(k-1)!}(\boldsymbol{A} - \lambda_0\boldsymbol{I}_n)^{k-1}\right)\boldsymbol{u}_i,$$

since the remaining terms in the series will be zero. In any case, it turns out that the k vector functions $\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_k(t)$ so constructed will be linearly independent solutions of (B.3). This tells us how to manage the real case.

Let's see how this goes in a few examples.

B.4 Examples

1. We take a simple case where
$$\boldsymbol{A} = \begin{bmatrix} -7 & 4 \\ -6 & 3 \end{bmatrix}.$$
The characteristic polynomial for $\boldsymbol{A}$ is $P_{\boldsymbol{A}}(\lambda) = \lambda^2 + 4\lambda + 3 = (\lambda + 1)(\lambda + 3)$. Thus the eigenvalues for $\boldsymbol{A}$ are $\lambda_1 = -1$ and $\lambda_2 = -3$. Since the eigenvalues are distinct, the algebraic and geometric multiplicities will be equal, and the generalised eigenvectors will simply be eigenvectors. An eigenvector for $\lambda_1 = -1$ is $\boldsymbol{u}_1 = (2, 3)$ and an eigenvector for $\lambda_2 = -3$ is $\boldsymbol{u}_2 = (1, 1)$. Our recipe then gives two linearly independent solutions as
$$\boldsymbol{x}_1(t) = e^{-t} \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{x}_2(t) = e^{-3t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

2. A more interesting case is the following:
$$\boldsymbol{A} = \begin{bmatrix} -2 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$
Since the matrix is upper triangular, the eigenvalues are the diagonal elements: $\lambda_1 = -2$ and $\lambda_2 = -1$. The algebraic multiplicity of λ_1 is 2. However, we readily see that $\dim(\ker(\boldsymbol{A} - \lambda_1 \boldsymbol{I}_3)) = 1$ and so the geometric multiplicity is 1. So we need to compute generalised eigenvectors in this case. We have
$$(\boldsymbol{A} - \lambda_1 \boldsymbol{I}_3)^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
and the generalised eigenvectors span the kernel of this matrix, and so we may take $\boldsymbol{u}_1 = (1, 0, 0)$ and $\boldsymbol{u}_2 = (0, 1, 0)$ as generalised eigenvectors. Applying the formula (B.4) gives
$$\begin{aligned} \boldsymbol{x}_1(t) &= e^{-2t} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + te^{-2t} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} e^{-2t} \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$
and
$$\begin{aligned} \boldsymbol{x}_2(t) &= e^{-2t} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + te^{-2t} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} te^{-2t} \\ e^{-2t} \\ 0 \end{bmatrix}. \end{aligned}$$
Finally we determine that $\boldsymbol{u}_3 = (0, 0, 1)$ is an eigenvector corresponding to $\lambda_2 = -1$, and so this gives the solution
$$\boldsymbol{x}_3(t) = \begin{bmatrix} 0 \\ 0 \\ e^{-t} \end{bmatrix}.$$
Thus we arrive at our three linearly independent solutions. •

Now let us look at the complex case. Thus let λ_0 be a complex eigenvalue (with nonzero imaginary part) of algebraic multiplicity ℓ. This means that $\bar{\lambda}_0$ will also be an eigenvalue of algebraic multiplicity ℓ since $\boldsymbol{A}$, and hence $P_{\boldsymbol{A}}(\lambda)$, is real. Thus we need to find 2ℓ linearly independent solutions. We do this by following the exact same idea as in the real case, except that we think of $\boldsymbol{A}$ as being a complex matrix for the moment. In this case it is still true that the matrix $(\boldsymbol{A}-\lambda_0\boldsymbol{I}_n)^\ell$ will have an ℓ-dimensional kernel, and we can take vectors $\boldsymbol{u}_1,\dots,\boldsymbol{u}_\ell$ as a basis for this kernel. Note, however, that since $(\boldsymbol{A}-\lambda_0\boldsymbol{I}_n)^\ell$ is complex, these vectors will also be complex. But the procedure is otherwise identical to the real case. One then constructs ℓ *complex* vector functions

$$\boldsymbol{z}_j(t) = e^{\lambda_0 t}\exp((\boldsymbol{A}-\lambda_0\boldsymbol{I}_n)t)\boldsymbol{u}_j. \tag{B.5}$$

Each such complex vector function will be a sum of its real and imaginary parts: $\boldsymbol{z}_j(t) = \boldsymbol{x}_j(t) + i\boldsymbol{y}_j(t)$. It turns out that the 2ℓ *real* vector functions $\boldsymbol{x}_1(t),\dots,\boldsymbol{x}_\ell(t),\boldsymbol{y}_1(t),\dots,\boldsymbol{y}_\ell(t)$ are linearly independent solutions to (B.3).

Let's see how this works in some examples.

B.5 Examples

1. An example with complex roots is

$$\boldsymbol{A} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix}.$$

The characteristic polynomial is $P_{\boldsymbol{A}}(\lambda) = r^3 + 4r^2 + 6r + 4$. One ascertains that the eigenvalues are then $\lambda_1 = -1+i$, $\lambda_2 = \bar{\lambda}_1 = -1-i$, $\lambda_3 = -2$. Let's deal with the complex root first. We have

$$\boldsymbol{A} - \lambda_1\boldsymbol{I}_3 = \begin{bmatrix} -i & 1 & 0 \\ -1 & -i & 0 \\ 0 & 0 & -1-i \end{bmatrix}$$

from which we glean that an eigenvector is $\boldsymbol{u}_1 = (-i,1,0)$. Using (B.5) the complex solution is then

$$\boldsymbol{z}_1(t) = e^{(-1+i)t}\begin{bmatrix} -i \\ 1 \\ 0 \end{bmatrix}.$$

Using Euler's formula, $e^{i\theta} = \cos\theta + i\sin\theta$, we have

$$\boldsymbol{z}_1(t) = e^{-t}\begin{bmatrix} -i\cos t + \sin t \\ \cos t + i\sin t \\ 0 \end{bmatrix} = e^{-t}\begin{bmatrix} \sin t \\ \cos t \\ 0 \end{bmatrix} + ie^{-t}\begin{bmatrix} -\cos t \\ \sin t \\ 0 \end{bmatrix}$$

thus giving

$$\boldsymbol{x}_1(t) = e^{-t}\begin{bmatrix} \sin t \\ \cos t \\ 0 \end{bmatrix}, \quad \boldsymbol{y}_1(t) = e^{-t}\begin{bmatrix} -\cos t \\ \sin t \\ 0 \end{bmatrix}.$$

Corresponding to the real eigenvalue λ_3 we readily determine that

$$\boldsymbol{x}_2 = e^{-2t}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

is a corresponding solution. This gives three linearly independent real solutions $\boldsymbol{x}_1(t)$, $\boldsymbol{y}_1(t)$, and $\boldsymbol{x}_2(t)$.

2. Let us look at a complex root with nontrivial multiplicity. The smallest matrix to possess such a feature will be one which is 4×4, and about the simplest example for which the geometric multiplicity is less than the algebraic multiplicity is

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}.$$

The eigenvalues are $\lambda_1 = i$ and $\lambda_2 = -i$, both with algebraic multiplicity 2. One readily determines that the kernel of $\boldsymbol{A} - i\boldsymbol{I}_4$ is one-dimensional, and so the geometric multiplicity of these eigenvalues is just 1. Thus we need to compute complex generalised eigenvectors. I have used Mathematica® for the computations below. We compute

$$(\boldsymbol{A} - i\boldsymbol{I}_4)^2 = 2\begin{bmatrix} -1 & -i & -i & 1 \\ i & -1 & -1 & -i \\ 0 & 0 & -1 & -i \\ 0 & 0 & i & -1 \end{bmatrix}$$

and one checks that $\boldsymbol{u}_1 = (0, 0, -i, 1)$ and $\boldsymbol{u}_2 = (-i, 1, 0, 0)$ are two linearly independent generalised eigenvectors. We compute

$$(\boldsymbol{A} - i\boldsymbol{I}_4)\boldsymbol{u}_1 = \begin{bmatrix} -i \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad (\boldsymbol{A} - i\boldsymbol{I}_4)\boldsymbol{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We now determine the two linearly independent real solutions corresponding to $\boldsymbol{u}_1$. We have

$$\begin{aligned} \boldsymbol{z}_1(t) &= e^{it}(\boldsymbol{u}_1 + t(\boldsymbol{A} - i\boldsymbol{I}_4)\boldsymbol{u}_1) \\ &= e^{it}\begin{bmatrix} 0 \\ 0 \\ -i \\ 1 \end{bmatrix} + te^{it}\begin{bmatrix} -i \\ 1 \\ 0 \\ 0 \end{bmatrix} \\ &= (\cos t + i \sin t)\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + i\begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} + t\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + it\begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) \\ &= \begin{bmatrix} t\sin t \\ t\cos t \\ \sin t \\ \cos t \end{bmatrix} + i\begin{bmatrix} -t\cos t \\ t\sin t \\ -\cos t \\ \sin t \end{bmatrix} \\ \implies \quad \boldsymbol{x}_1(t) &= \begin{bmatrix} t\sin t \\ t\cos t \\ \sin t \\ \cos t \end{bmatrix}, \quad \boldsymbol{y}_1(t) = \begin{bmatrix} -t\cos t \\ t\sin t \\ -\cos t \\ \sin t \end{bmatrix}. \end{aligned}$$

For $\boldsymbol{u}_2$ we have

$$\begin{aligned}
\boldsymbol{z}_2(t) &= e^{it}(\boldsymbol{u}_2 + t(\boldsymbol{A} - i\boldsymbol{I}_4)\boldsymbol{u}_2) \\
&= e^{it} \begin{bmatrix} -i \\ 1 \\ 0 \\ 0 \end{bmatrix} \\
&= (\cos t + i \sin t) \left(\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + i \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) \\
&= \begin{bmatrix} \sin t \\ \cos t \\ 0 \\ 0 \end{bmatrix} + i \begin{bmatrix} -\cos t \\ \sin t \\ 0 \\ 0 \end{bmatrix} \\
\implies \quad \boldsymbol{x}_2(t) &= \begin{bmatrix} \sin t \\ \cos t \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{y}_2 = \begin{bmatrix} -\cos t \\ \sin t \\ 0 \\ 0 \end{bmatrix}.
\end{aligned}$$

Thus we have the four real linearly independent solutions $\boldsymbol{x}_1(t)$, $\boldsymbol{x}_2(t)$, $\boldsymbol{y}_1(t)$, and $\boldsymbol{y}_2(t)$. •

We still haven't gotten to the matrix exponential yet, but all the hard work is done. Using the above methodology we may in principle compute for any $n \times n$ matrix $\boldsymbol{A}$, n linearly independent solutions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n(t)$.[1] If we assemble the resulting solutions into the columns of a matrix $\boldsymbol{X}(t)$:

$$\boldsymbol{X}(t) = \left[\begin{array}{c|c|c} \boldsymbol{x}_1(t) & \cdots & \boldsymbol{x}_n(t) \end{array} \right],$$

the resulting matrix is an example of a fundamental matrix. Generally, a ***fundamental matrix*** is any $n \times n$ matrix function of t whose columns form n linearly independent solutions to (B.3). What we have done above is give a recipe for computing *a* fundamental matrix (there are an infinite number of these). The following result connects the construction of a fundamental matrix with the matrix exponential.

B.6 Theorem *Given any fundamental matrix $\boldsymbol{X}(t)$ we have $e^{\boldsymbol{A}t} = \boldsymbol{X}(t)\boldsymbol{X}^{-1}(0)$.*

Thus, once we have a fundamental matrix, the computation of the matrix exponential is just algebra, although computing inverses of matrices of any size is a task best left to the computer.

Let's work this out for our four examples.

B.7 Examples

1. If

$$\boldsymbol{A} = \begin{bmatrix} -7 & 4 \\ -6 & 3 \end{bmatrix}.$$

[1] Note that the solutions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are those obtained from both real *and* complex eigenvalues. Therefore, the solutions denoted above as "$\boldsymbol{y}_i(t)$" for complex eigenvalues will be included in the n linearly independent solutions, except now I am calling everything $\boldsymbol{x}_j(t)$.

then we have determined a fundamental matrix to be

$$\boldsymbol{X}(t) = \begin{bmatrix} 2e^{-t} & e^{-3t} \\ 3e^{-2t} & e^{-3t} \end{bmatrix}.$$

It is then an easy calculation to arrive at

$$e^{\boldsymbol{A}t} = \boldsymbol{X}(t)\boldsymbol{X}^{-1}(0) = \begin{bmatrix} 3e^{-3t} - 2e^{-t} & -2e^{-3t} + 2e^{-t} \\ 3e^{-3t} - 3e^{-t} & -2e^{-3t} + 3e^{-t} \end{bmatrix}$$

2. When

$$\boldsymbol{A} = \begin{bmatrix} -2 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

we had determined a fundamental matrix to be

$$\boldsymbol{X}(t) = \begin{bmatrix} e^{-2t} & te^{-2t} & 0 \\ 0 & e^{-2t} & 0 \\ 0 & 0 & e^{-t} \end{bmatrix}.$$

It so happens that in this example we lucked out and $e^{\boldsymbol{A}t} = \boldsymbol{X}(t)$ since $\boldsymbol{X}(0) = \boldsymbol{I}_3$.

3. For

$$\boldsymbol{A} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix}.$$

we had determined the fundamental matrix

$$\boldsymbol{X}(t) = \begin{bmatrix} e^{-t}\sin t & -e^{-t}\cos t & 0 \\ e^{-t}\cos t & e^{-t}\sin t & 0 \\ 0 & 0 & e^{-2t} \end{bmatrix}.$$

A straightforward computation yields

$$e^{\boldsymbol{A}t} = \boldsymbol{X}(t)\boldsymbol{X}^{-1}(0) = \begin{bmatrix} e^{-t}\cos t & e^{-t}\sin t & 0 \\ -e^{-t}\sin t & e^{-t}\cos t & 0 \\ 0 & 0 & e^{-2t} \end{bmatrix}.$$

4. Finally, we look at the 4×4 example we worked out:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}.$$

The fundamental matrix we obtained was

$$\boldsymbol{X}(t) = \begin{bmatrix} t\sin t & -t\cos t & \sin t & -\cos t \\ t\cos t & t\sin t & \cos t & \sin t \\ \sin t & -\cos t & 0 & 0 \\ \cos t & \sin t & 0 & 0 \end{bmatrix}.$$

A little manipulation gives

$$e^{\boldsymbol{A}t} = \begin{bmatrix} \cos t & \sin t & t\cos t & t\sin t \\ -\sin t & \cos t & -t\sin t & t\cos t \\ 0 & 0 & \cos t & \sin t \\ 0 & 0 & -\sin t & \cos t \end{bmatrix}. \qquad \bullet$$

Now you know how to compute the matrix exponential. However, you can get Maple® to do this. I use Mathematica® for such chores. But you are expected to know how in principle to determine the matrix exponential. Most importantly, you should know precisely *what* it is.

Exercises

EB.1 Solve the initial value problem

$$\tau\dot{x}(t) + x(t) = A1(t), \quad x(0) = 0,$$

where $\tau > 0$, and $1(t)$ is the ***unit step function***:

$$1(t) = \begin{cases} 1, & t \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Draw a graph of the solution.

EB.2 A mass moving in a gravitational field is governed by the differential equation $m\ddot{y}(t) = -mg$. Solve this differential equation with initial conditions $y(0) = y_0$ and $\dot{y}(0) = v_0$.

EB.3 Obtain the general solution to the differential equation

$$\ddot{x}(t) + 2\zeta\omega_0\dot{x}(t) + \omega_0^2 x(t) = A\cos\omega t$$

for ω, ω_0, ζ and A positive constants. You will have to deal with three cases depending on the value of ζ.

EB.4 Compute by hand $e^{\boldsymbol{A}t}$ for the following matrices:

(a) $\boldsymbol{A} = \begin{bmatrix} -2 & 0 & 1 \\ 1 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix}$;

(b) $\boldsymbol{A} = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}$ for $\sigma \in \mathbb{R}$ and $\omega > 0$.

EB.5 Use a computer package to determine $e^{\boldsymbol{A}t}$ for the following matrices:

(a) $\boldsymbol{A} = \begin{bmatrix} -2 & 3 & 1 & 0 \\ -3 & -2 & 0 & 1 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & -3 & -2 \end{bmatrix}$;

(b) $\boldsymbol{A} = \begin{bmatrix} 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$.

Appendix C

Polynomials and rational functions

We shall be frequently be encountering and manipulating polynomials, so it will be helpful to have on hand some basic facts concerning such objects. That polynomials might be useful to us can be seen by noting that we have already defined the characteristic polynomial which will be very important to us in these notes.

Contents

C.1 Polynomials

We denote by $\mathbb{F}[\xi]$ the set of polynomials in indeterminant ξ and with coefficients in $\mathbb{F}$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. Thus a typical element of $\mathbb{F}[\xi]$ looks like

$$P(\xi) = a_k\xi^k + a_{k-1}\xi^{k-1} + \cdots + a_1\xi + a_0 \tag{C.1}$$

where $a_0, \ldots, a_k \in \mathbb{F}$ with $a_k \neq 0$. We call k the ***degree*** of P which we denote by $\deg(P)$. You should not think of ξ as being an element of $\mathbb{F}$, but rather as just being a placeholder. If we wish to plug in values from $\mathbb{F}$ into P we shall generally say when we do this. Of course, you can add and multiply polynomials in just the ways with which you are familiar.

A polynomial P of the form (C.1) is ***monic*** if $a_k = 1$; that is, a monic polynomial is one where the coefficient of the highest power of the indeterminant is $+1$. For example, the characteristic polynomial is always a monic polynomial.[1] Given two polynomials $P_1, P_2 \in \mathbb{F}[\xi]$ their ***least common multiple*** (***LCM***) is the unique monic polynomial $Q \in \mathbb{F}[\xi]$ of least degree with the property that $Q = P_1R_1$ and $Q = P_2R_2$ for some $R_1, R_2 \in \mathbb{F}[\xi]$.

C.1 Examples In each of these examples, the polynomials may be thought of as in either $\mathbb{R}[\xi]$ or $\mathbb{C}[\xi]$.

1. If
$$P_1(\xi) = \xi + 2, \quad P_2(\xi) = 2\xi - 3,$$
then the LCM of P_1 and P_2 is $Q(\xi) = \xi^2 - \frac{1}{2}\xi - \frac{3}{2}$.
2. If
$$P_1(\xi) = \xi^2 + 2\xi + 1, \quad P_2(\xi) = \xi + 1,$$
then the LCM of P_1 and P_2 is $Q(\xi) = \xi^2 + 2\xi + 1$. •

[1] This is why we defined $P_{\boldsymbol{A}}(\lambda)$ by $\det(\lambda \boldsymbol{I}_n - \boldsymbol{A})$ instead of by $\det(\boldsymbol{A} - \lambda \boldsymbol{I}_n)$.

If $P \in \mathbb{F}[\xi]$ is given by

$$P(\xi) = a_k\xi^k + a_{k-1}\xi^{k-1} + \cdots + a_1\xi + a_0,$$

a ***root*** for P is a number $\alpha \in \mathbb{F}$ with the property that

$$a_k\alpha^k + a_{k-1}\alpha^{k-1} + \cdots + a_1\alpha + a_0 = 0.$$

We denote the set of roots of P by $\operatorname{spec}(P)$, mirroring our notation for eigenvalues for matrices. If α is a root of P then there exists $Q \in \mathbb{F}[\xi]$ so that $P(\xi) = (\xi - \alpha)Q(\xi)$. If α is a root of P and if m is the unique integer with the property that $P(\xi) = (\xi - \alpha)^m Q(\xi)$ where α is not a root of Q, then m is the ***multiplicty*** of the root. If $\mathbb{F} = \mathbb{C}$, a root $\alpha \in \mathbb{C}$ is (1) in the ***positive half-plane*** if $\operatorname{Re}(\alpha) > 0$, (2) in the ***negative half-plane*** if $\operatorname{Re}(\alpha) < 0$, and (3) on the ***imaginary axis*** if $\operatorname{Re}(\alpha) = 0$. We denote the positive half-plane by $\mathbb{C}_+$ and the negative half-plane by $\mathbb{C}_-$. We shall often also denote the imaginary axis by $i\mathbb{R}$ (meaning $\{i\omega \mid \omega \in \mathbb{R}\}$). By $\overline{\mathbb{C}}_+$ we mean the positive half-plane along with the imaginary axis, and similarly be $\overline{\mathbb{C}}_-$ we mean the negative half-plane along with the imaginary axis.

The ***greatest common divisor*** (***GCD***) of two polynomials $P_1, P_2 \in \mathbb{F}[\xi]$ is the unique monic polynomial T of greatest degree so that $P_1 = TQ_1$ and $P_2 = TQ_2$ for some $Q_1, Q_2 \in \mathbb{F}[\xi]$.

C.2 Examples In each of these examples, the polynomials may be thought of as in either $\mathbb{R}[\xi]$ or $\mathbb{C}[\xi]$.

1. If
$$P_1(\xi) = \xi^2 + 1, \quad P_2(\xi) = 3\xi$$
then the GCD of P_1 and P_2 is $T(\xi) = 1$.
2. If
$$P_1(\xi) = 3\xi^2 + 6\xi + 3, \quad P_2(\xi) = 2\xi^2 - 6\xi - 8,$$
then the GCD of P_1 and P_2 is $T(\xi) = \xi + 1$. •

If the GCD of polynomials P_1 and P_2 is $T(\xi) = 1$, then P_1 and P_2 are said to be ***coprime***. A polynomial $P \in \mathbb{F}[\xi]$ with $\deg(P) > 0$ is ***irreducible*** if there are no polynomials $P_1, P_2 \in \mathbb{F}[\xi]$, both with degree less than P, with the property that $P = P_1P_2$. Thus an irreducible polynomial cannot be factored.

C.3 Examples

1. By the Fundamental Theorem of Algebra, the only irreducible monic polynomials in $\mathbb{C}[\xi]$ are of the form $P(\xi) = \xi + a$ for some $a \in \mathbb{C}$.
2. As we know, there are polynomials of degree two in $\mathbb{R}[\xi]$ which are irreducible. For example $P(\xi) = \xi^2 + 1$ is irreducible in $\mathbb{R}[\xi]$, but not in $\mathbb{C}[\xi]$. The irreducible monic polynomials in $\mathbb{R}[\xi]$ are of the form
   (a) $P(\xi) = \xi + a$ for some $a \in \mathbb{R}$ or
   (b) $P(\xi) = \xi^2 + b\xi + c$ where $b, c \in \mathbb{R}$ satisfy $b^2 - 4c < 0$.

   One can easily show that an irreducible monic polynomial of degree 2 in $\mathbb{R}[\xi]$ must have the form $(\xi - \sigma)^2 + \omega^2$ for $\sigma \in \mathbb{R}$ and $\omega > 0$ (see Exercise EC.1). •

The following result is related to the ***Euclidean algorithm*** for $\mathbb{F}[\xi]$ which, you will recall, states that for any polynomials $P, Q \in \mathbb{F}[\xi]$ there exists polynomials $F, R \in \mathbb{F}[\xi]$ with the properties

1. $P = FQ + R$ and
2. $\deg(D) < \deg(Q)$.

That is to say, a polynomial can be written as a product of a given polynomial and a remainder. With this in mind, we can prove the following result.

C.4 Lemma *Let $P_1, P_2, F \in \mathbb{F}[\xi]$ be polynomials with P_1 and P_2 coprime. Then there exists $Q_1, Q_2 \in \mathbb{F}[s]$ so that*

(i) $\deg(Q_1) < \deg(P_2)$ *and*

(ii) $Q_1P_1 + Q_2P_2 = F$.

Proof Since P_1 and P_2 are coprime it follows that there exists $\tilde{Q}_1, \tilde{Q}_2 \in \mathbb{R}[s]$ so that

$$\tilde{Q}_1P_1 + \tilde{Q}_2P_2 = 1.$$

Therefore we have

$$(\tilde{Q}_1F)P_1 + (\tilde{Q}_2F)P_2 = F,$$

from which it follows that for any $G \in \mathbb{R}[s]$ we have

$$(\tilde{Q}_1F - P_2G)P_1 + (\tilde{Q}_2F + P_1G)P_2 = F.$$

The Euclidean algorithm asserts that there exists a $G, R \in \mathbb{F}[\xi]$ so that

$$\tilde{Q}_1F = GP_2 + R$$

and $\deg(R) < \deg(P_2)$. That is to say, there exists $G \in \mathbb{F}[\xi]$ so that

$$\deg(\tilde{Q}_1F - P_2G) < \deg(P_2).$$

Choosing this G and then defining $Q_1 = \tilde{Q}_1F - P_2G$ and $Q_2 = \tilde{Q}_2F + P_1G$ gives the result. ■

C.2 Rational functions

We also wish to talk about objects which are quotients of polynomials. A ***rational function over $\mathbb{F}$ with indeterminate ξ*** is a quotient of two elements of $\mathbb{F}[\xi]$. Thus we write a rational function over $\mathbb{F}$ as

$$R(\xi) = \frac{N(\xi)}{D(\xi)}, \quad N, D \in \mathbb{F}[\xi].$$

Thus

$$R(\xi) = \frac{a_k\xi^k + a_{k-1}\xi^{k-1} + \cdots + a_1\xi + a_0}{b_\ell\xi^\ell + b_{\ell-1}\xi^{\ell-1} + \cdots + b_1\xi + b_0}$$

where $a_0, \ldots, a_k, b_0, \ldots, b_\ell \in \mathbb{F}$, and not all the b_i's are zero. We denote the set of rational functions over $\mathbb{F}$ with indeterminate ξ by $\mathbb{F}(\xi)$. One should take care not to unduly concern oneself about things like whether the rational function blows up for certain values of ξ where $D(\xi) = 0$. As a polynomial, the only polynomial which is zero is the zero polynomial. If $P \in \mathbb{F}[\xi]$ is a non-zero polynomial, then the two rational functions

$$R_1 = \frac{PN}{PD}, \quad R_2 = \frac{N}{D}$$

will in fact represent the same rational function. This is exactly the same thing we do when we say that $\frac{1}{3}$ and $\frac{2}{6}$ are the same rational number. Note that for $R \in \mathbb{F}(\xi)$ there are unique coprime monic polynomials $N, D \in \mathbb{F}[\xi]$ so that

$$R = a\frac{N}{D} \tag{C.2}$$

for some $a \in \mathbb{F}$.

C.5 Example Consider the rational function

$$R(\xi) = \frac{2\xi^2 - 2\xi - 12}{3\xi^2 - 15\xi + 18}.$$

We can write this as

$$R(\xi) = \tfrac{2}{3}\frac{\xi + 2}{\xi - 2}$$

which is the unique representation of the form (C.2). •

Given a rational function $R \in \mathbb{F}(\xi)$ with $a \in \mathbb{F}$ and $N, D \in \mathbb{F}[\xi]$ defined by (C.2), we call (aN, D) the ***canonical fractional representative*** of R. We will be frequently in need of this simple concept, so shall abbreviate if ***c.f.r.***. A rational function R with c.f.r. (N, D) is ***proper*** if $\deg(N) \leq \deg(D)$, and ***strictly proper*** if $\deg(N) < \deg(D)$. A rational function which is not proper is ***improper***.

Let $R \in \mathbb{F}(\xi)$ which we write in its unique representation (C.2) for some coprime monic polynomials $N, D \in \mathbb{F}[\xi]$. A ***zero*** of R is defined to be a root of N and a ***pole*** of R is defined to be a root of D. Note that in this way we get around the problem of the "function" R not being defined at poles. Two rational functions $R_1, R_2 \in \mathbb{F}(\xi)$ are ***coprime*** if they have no common zeros.

The final thing we do is provide a discussion of the so-called "partial fraction expansion." Recall that the idea here is to take a rational function and expand it as a sum of rational functions whose denominators are powers of an irreducible polynomial. Thus, for example

$$\frac{\xi^3 - 3\xi + 2}{\xi^3 - 5\xi^2 + 3\xi + 9} = \frac{1}{1} + 5\frac{1}{(\xi - 3)^2} + \frac{19}{4}\frac{1}{\xi - 3} + \frac{1}{4}\frac{1}{\xi + 1}.$$

It is hard to come across an accurate description of how this is done, so let us provide one here.

C.6 Theorem *Let $R \in \mathbb{F}(\xi)$ and suppose that*

$$R = \frac{N}{D}$$

where $N, D \in \mathbb{F}[\xi]$ are coprime, and take D to be monic.

There exists

(i) m irreducible monic polynomials $D_1, \dots, D_m \in \mathbb{F}[\xi]$,

(ii) positive integers $j_1, \dots, j_m$,

(iii) $j_1 + \cdots + j_m$ polynomials $N_{1,1}(x), \dots, N_{1,j_1}, \dots, N_{m,1}, \dots, N_{m,j_m} \in \mathbb{F}[\xi]$, and

(iv) a polynomial $Q \in \mathbb{F}[\xi]$ of degree $\deg(N) - \deg(D)$ (take $Q = 0$ if $\deg(N) - \deg(D) < 0$),

with the properties

(v) $D_1, \dots, D_m$ *are coprime (i.e., distinct),*

(vi) $\deg(N_{i,k}) < \deg(D_i)$ *for* $i = 1, \dots, m$ *and* $k = 1, \dots, j_i$,

(vii) $N_{i,k}$ *and* D_i *are coprime for* $i = 1, \dots, m$ *and* $k = 1, \dots, j_i$, *and*

(viii) $R = \displaystyle\sum_{i=1}^{m} \sum_{k=1}^{j_i} \frac{N_{i,k}}{(D_i)^k} + Q.$

Furthermore the objects described in (i)–(iv) are the unique such objects with the properties (v)–(viii).

The expression in part (viii) is called the ***partial fraction expansion*** of R. The proof of this is straightforward, but requires some buildup, and we refer to [Lang 2005, Theorems 5.2 and 5.3].

It will turn out that we are primarily interested in the case when $\deg(N) \leq \deg(D)$, and in this case Q will be a constant, possibly zero, when $\deg(N) = \deg(D)$, and zero when $\deg(N) < \deg(D)$. For a rational function $R \in \mathbb{C}(\xi)$ which satisfies our degree condition, Theorem C.6 tells us that we may write

$$R(\xi) = \sum_{i=1}^{m} \sum_{k=1}^{j_m} \frac{\beta_{i,k}}{(\xi - \alpha_i)^k} + \beta \tag{C.3}$$

for some uniquely defined complex numbers β, α_i, $i = 1, \dots, m$, and $\beta_{i,k}$, $i = 1, \dots, m$, $k = 1, \dots, j_i$. For $R \in \mathbb{R}(\xi)$, things are a bit more complicated. We may write

$$R(\xi) = \sum_{i=1}^{r} \sum_{k=1}^{j_i} \frac{\beta_{i,k}}{(\xi - \alpha_i)^k} + \sum_{i=1}^{m} \sum_{k=1}^{\ell_i} \frac{a_{i,k}\xi + b_{i,k}}{\left((\xi - \sigma_i)^2 + \omega_i^2\right)^k} + b \tag{C.4}$$

for real numbers α_i, $i = 1, \dots, r$, $\beta_{i,k}$, $i = 1, \dots, r$, $k = 1, \dots, j_i$, $a_{i,k}, b_{i,k}$, $i = 1, \dots, m$, $k = 1, \dots, \ell_i$, σ_i, ω_i, $i = 1, \dots, \ell$, and b.

This sort of leaves open how we compute the constants in the partial fraction expansion for $R = N/D$. We shall say here how to do it when $R \in \mathbb{C}(\xi)$. In this case $\alpha_1, \dots, \alpha_m$ are the poles of R, and j_i is the multiplicity of the pole α_i. That is, $D(\xi) = (\xi - \alpha_i)^{j_i} Q(\xi)$ where Q and $(\xi - \alpha_i)$ are coprime. It turns out that

$$\beta_{i,j_i} = \frac{1}{(k - j_i)!} \frac{\mathrm{d}^{k-j_i}}{\mathrm{d}\xi^{k-j_i}} \big((\xi - \alpha)^k R(\xi)\big)\Big|_{\xi=\alpha_i} \tag{C.5}$$

As usual, this is self-explanatory in examples.

C.7 Examples

1. First let us show that one cannot dispense with the constant term if the numerator and denominator polynomials have the same degree. If we take

$$R(\xi) = \frac{\xi + 1}{\xi + 2}$$

then its partial fraction expansion is

$$R(\xi) = 1 - \frac{1}{\xi + 2}.$$

2. We take
$$R(\xi) = \frac{5\xi + 4}{\xi^2 + \xi - 2}.$$
The first thing to do is factor the denominator: $\xi^2 + \xi - 2 = (\xi - 1)(\xi + 2)$. Thus in the parlance of (C.3) we have $\alpha_1 = 1$ and $\alpha_2 = -2$. These roots have multiplicity 1 and so $j_1 = j_2 = 1$. By (C.5) we then have
$$\beta_{1,1} = (\xi - 1)\frac{5\xi + 4}{(\xi - 1)(\xi + 2)}\Big|_{\xi=1} = \frac{9}{3} = 3$$
and
$$\beta_{2,1} = (\xi + 2)\frac{5\xi + 4}{(\xi - 1)(\xi + 2)}\Big|_{\xi=-2} = \frac{-6}{-3} = 2.$$
Thus the partial fraction expansion is
$$R(\xi) = \frac{3}{\xi - 1} + \frac{2}{\xi + 2}.$$

3. We take
$$R(\xi) = \frac{-3\xi^2 + 5\xi + 2}{\xi^3 - 3\xi^2 + \xi - 3}.$$
The roots of the denominator polynomial are $\alpha_1 = 3$, $\alpha_2 = i$, and $\alpha_3 = -i$. Since we have complex roots, there will be different partial fraction expansions, depending on whether we are thinking of $R \in \mathbb{C}(\xi)$ or $R \in \mathbb{R}(\xi)$. Let us take the complex case first. Using (C.5) we have
$$\begin{aligned}
\beta_{1,1} &= (\xi - 3)\frac{-3\xi^2 + 5\xi + 2}{(\xi + 3)(\xi - i)(\xi + i)}\Big|_{\xi=3} = -1 \\
\beta_{2,1} &= (\xi - i)\frac{-3\xi^2 + 5\xi + 2}{(\xi + 3)(\xi - i)(\xi + i)}\Big|_{\xi=i} = -1 + \tfrac{i}{2} \\
\beta_{3,1} &= (\xi + i)\frac{-3\xi^2 + 5\xi + 2}{(\xi + 3)(\xi - i)(\xi + i)}\Big|_{\xi=-i} = -1 - \tfrac{i}{2}.
\end{aligned}$$
Thus the partial fraction expansion over $\mathbb{C}$ is
$$R(\xi) = -\frac{1}{\xi - 3} - \frac{1 - \frac{i}{2}}{\xi - i} - \frac{1 + \frac{i}{2}}{\xi + i}.$$
The partial fraction expansion over $\mathbb{R}$ turns out to be
$$R(\xi) = -\frac{1}{\xi - 3} - \frac{2\xi + 1}{\xi^2 + 1}.$$
Thus, employing the symbols in (C.4) we have $\alpha_1 = 3$, $\sigma_1 = 0$, and $\omega_1 = 1$. The easiest way to determine this is to compute the complex partial fraction expansion, and then recombine the complex conjugate pairs over a common denominator.

4. We take
$$R(\xi) = \frac{2\xi^2 + 1}{\xi^3 + 3\xi^2 + 3\xi + 1}.$$

The root of the denominator polynomial is -1 which has multiplicity 3. We use (C.5) to get

$$\beta_{1,1} = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}\xi^2}\Big((\xi+1)^3\frac{2\xi^2+1}{\xi^3+3\xi^2+3\xi+1}\Big)\Big|_{\xi=-1} = 2$$
$$\beta_{1,2} = \frac{\mathrm{d}}{\mathrm{d}\xi}\Big((\xi+1)^3\frac{2\xi^2+1}{\xi^3+3\xi^2+3\xi+1}\Big)\Big|_{\xi=-1} = -4$$
$$\beta_{1,3} = (\xi+1)^3\frac{2\xi^2+1}{\xi^3+3\xi^2+3\xi+1}\Big|_{\xi=-1} = 3.$$

Thus the partial fraction expansion is

$$R(\xi) = \frac{2}{\xi+1} - \frac{4}{(\xi+1)^2} + \frac{3}{(\xi+1)^3}.$$

•

Heavyside coverup?

Exercises

EC.1 Let $P(\xi) \in \mathbb{R}[\xi]$ be monic, irreducible, and degree two. Show that there exists $\sigma \in \mathbb{R}$ and $\omega > 0$ so that $P(\xi) = (\xi - \sigma)^2 + \omega^2$.

EC.2 Note that $\mathbb{R}[\xi]$ is naturally a subset of $\mathbb{C}[\xi]$. If $P(\xi) \in \mathbb{R}[\xi]$, denote by $\bar{P}(\xi)$ the same polynomial, except thought of as being in $\mathbb{C}[\xi]$. Show that polynomials $P_1(\xi), P_2(\xi) \in \mathbb{R}[\xi]$ are coprime if and only if $\bar{P}_1(\xi), \bar{P}_2(\xi)$ are coprime.

EC.3 For

$$P(\xi) = \xi^n + p_{n-1}\xi^{n-1} + \cdots + p_1\xi + p_0 \in \mathbb{F}[\xi], \quad \mathbb{F} \in \{\mathbb{R}, \mathbb{C}\},$$

show that sum of the roots of P, counting multiplicities, is equal to $-p_{n-1}$.

EC.4 Determine the c.f.r. of the following rational functions:

(a) $\dfrac{\xi + 1}{3\xi^2 + 6}$;

(b) $\dfrac{-3\xi^2 + 6\xi + 9}{\xi^3 + 5\xi^2 + 7\xi + 3}$;

(c) $\dfrac{2\xi + 2}{(\xi + 1)^3}$;

(d) $\dfrac{2\xi^2 + 4}{3\xi^3 + 9\xi^2 + 3\xi + 9}$.

EC.5 Determine the partial fraction expansion of the following rational functions (for functions with complex poles, determine both the real and complex partial fraction expansions):

(a) $\dfrac{\xi^2 - 1}{\xi + 2}$;

(b) $\dfrac{3\xi + 4}{\xi^2 + 3\xi + 2}$;

(c) $\dfrac{2\xi^2 - \xi + 1}{2\xi^3 + 18\xi^2 + 48\xi + 32}$;

(d) $\dfrac{\xi^2 + 2}{(\xi^2 + 1)^2}$.

EC.6 Let $R \in \mathbb{C}(\xi)$. Show that $R \in \mathbb{R}(\xi)$ if and only if $R(s) = \overline{R(\bar{s})}$ for every $s \in \mathbb{C}$ which is not a pole for R.

Appendix D

Complex variable theory

We shall require some basic facts about functions of a complex variable. We assume the reader to have some background in such matters. Certainly we anticipate that the reader is fully functional in manipulating complex numbers. For more details on what we say here, we refer to [Lang 2003]. An excellent introduction to those topics in complex variable theory that are useful in control may be found, complete with all details, in Appendix A of [Seron, Braslavsky, and Goodwin 1997].

Contents

D.1 The complex plane and its subsets

The ***complex plane*** $\mathbb{C}$ is the set of ordered pairs (x, y) of complex numbers. We will follow the usual convention of writing $(x, y) = x + iy$ where $i = \sqrt{-1}$. Note that we do not use the symbol j for $\sqrt{-1}$. Only electrical engineers, and those under their influence, use this crazy notation. Complex numbers of the form $x + i0$ for $x \in \mathbb{R}$ are ***real*** and complex numbers of the form $z = 0 + iy$ for $y \in \mathbb{R}$ are called ***imaginary***. For $z = x + iy$ we denote $\mathrm{Re}(z) = x$ and $\mathrm{Im}(z) = y$ as the ***real part*** and ***imaginary part***, respectively, of z. We will assume the reader knows how to add and multiply complex numbers:

$$(x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2)$$
$$(x_1 + iy_1)(x_2 + iy_2) = (x_1x_2 - y_1y_2) + i(x_2y_1 + x_1y_2).$$

The ***complex conjugate*** of $z = x + iy$ is the complex number $\bar{z} = x - iy$. By $|x + iy| = \sqrt{x^2 + y^2}$ we denote the ***modulus*** of z. If z is real, then $|z|$ is the usual absolute value. By $\measuredangle x + iy = \mathrm{atan2}(x, y)$ we denote the ***argument*** of z. Here $\mathrm{atan2}\colon \mathbb{R}^2 \to (-\pi, \pi]$ is the intelligent arctangent that knows which quadrant one is in. This is illustrated in Figure D.1.

Let us make a few definitions concerning the nature of subsets of $\mathbb{C}$. First we denote by $D(z, r)$ the ***open disk*** of radius r centered at z:

$$D(z, r) = \{z' \in \mathbb{C} \mid |z' - z| < r\}.$$

The ***closed disk*** of radius r centered at z is denoted:

$$\bar{D}(z, r) = \{z' \in \mathbb{C} \mid |z' - z| \leq r\}.$$

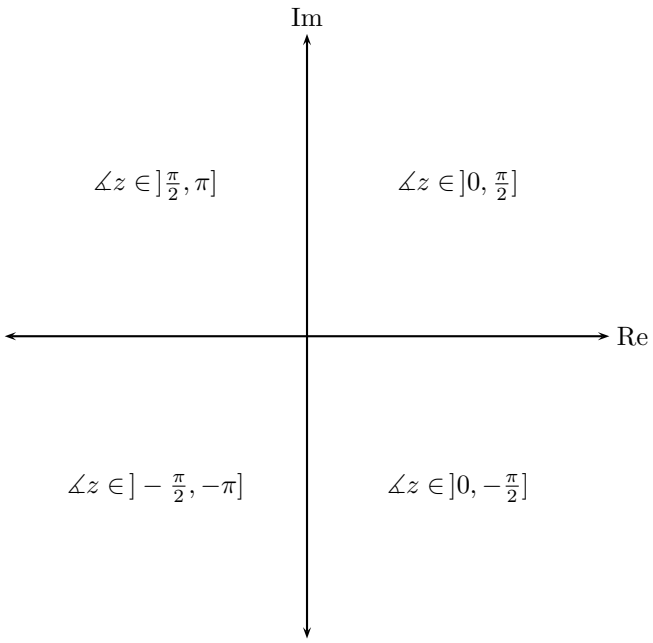


Figure D.1 The values for the argument of a complex number

Now we have the following.

D.1 Definition Let $S \subset \mathbb{C}$.

(i) S is ***open*** if for any $z \in S$ there exists an $\epsilon > 0$ so that $D(z, \epsilon) \subset S$.

(ii) S is ***closed*** if its complement, $\mathbb{C} \setminus S$, is open.

(iii) A ***boundary point*** for S is a point $s \in \mathbb{C}$ so that for every $\epsilon > 0$ there exists $s_1, s_2 \in D(s, \epsilon)$ so that $s_1 \in S$ and $s_2 \notin S$.

(iv) $S \subset \mathbb{C}$ is ***connected*** any two points in S can be connected by a polygonal path consisting of a finite number of line segments.

(v) D is ***simply connected*** if every closed curve in D can be continuously contracted to a point.[1].

(vi) S is a ***domain*** if it is open and connected.

(vii) S is a ***region*** if it is a domain together with a possibly empty subset of its boundary.

(viii) A region is ***closed*** if it contains all of its boundary points.

(ix) A region S is ***bounded*** if there exists $R > 0$ so that $S \subset D(0, R)$. •

D.2 Functions

Let $D \subset \mathbb{C}$ be a domain. A function $f\colon D \to \mathbb{C}$ is ***continuous at z_0*** if for every $\delta > 0$ there exists $\epsilon > 0$ so that $|z - z_0| < \epsilon$ implies that $|f(z) - f(z_0)| < \delta$. If f is continuous at every point in D then f is simply ***continuous***. The function f is ***differentiable*** at z_0 if the limit

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

[1]This definition refers ahead to Section D.3 for the notion of a closed curve. We do not often use the idea of simple connectivity, but it intuitively means "no holes."

exists and is independent of the manner in which the limit is taken.[2] The limit when it so exists is the ***derivative*** and denoted $f'(z_0)$. If we write $z = x + iy$ and $f(z) = u(x, y)+iv(x, y)$ for $\mathbb{R}$-valued functions u and v, then it may be shown that f is differentiable at $z_0 = x_0 + iy_0$ if and only if (1) u and v are differentiable at (x_0, y_0) and (2) the ***Cauchy-Riemann equations*** are satisfied at z_0:

$$\frac{\partial u}{\partial x}(x_0, y_0) = \frac{\partial v}{\partial y}(x_0, y_0), \quad \frac{\partial u}{\partial y}(x_0, y_0) = -\frac{\partial v}{\partial x}(x_0, y_0).$$

A function $f\colon D \to \mathbb{C}$ on a domain D is ***analytic*** at $z_0 \in D$ if there exists $\epsilon > 0$ so that f is differentiable at every point in $D(z_0, \epsilon)$. If $R \subset D$ is a region, we say f is analytic in R if it is analytic at each point in R. Note that this may necessitate differentiability of f at points outside R.

Analytic functions may fail to be defined at isolated points. Let us be systematic about characterising such points.

D.2 Definition Let $f\colon D \to \mathbb{C}$ be analytic.

(i) A point $z_0 \in D$ is an ***isolated singularity*** for f if there exists $\epsilon > 0$ so that f is defined and analytic on $D(z_0, r) \setminus \{z_0\}$ but is not defined on $D(z_0, r)$.

(ii) An isolated singularity z_0 for f is ***removable*** if there exists an $r > 0$ and an analytic function $g\colon D(z_0, r) \to \mathbb{C}$ so that $g(z) = f(z)$ for $z \neq z_0$.

(iii) An isolated singularity z_0 for f is a ***pole*** if

  (a) $\lim_{z\to z_0} |f(z)| = \infty$ and

  (b) there exists $k > 0$ so that the function g defined by $g(z) = (z-z_0)^k f(z)$ is analytic at z_0. The smallest $k \in \mathbb{Z}$ for which this is true is called the ***order*** of the pole.

(iv) An isolated singularity z_0 for f is ***essential*** if it is neither a pole nor a removable singularity.

(v) A function $f\colon D \to \mathbb{C}$ is ***meromorphic*** if it analytic except possibly at a finite set of poles. •

Another important topic in the theory of complex functions is that of series expansions. Let D be a domain. If $f\colon D \to \mathbb{C}$ is analytic at $z_0 \in D$ then one can show that all derivatives of f exist at z_0. The ***Taylor series*** for f at z_0 is then the series

$$f(z) = \sum_{j=0}^{\infty} a_j (z - z_0)^j.$$

where the coefficients are defined by

$$a_j = \frac{f^{(j)}(z_0)}{j!}.$$

Analyticity of f guarantees pointwise convergence of the Taylor series in a closed disk of positive radius. If z_0 is an isolated singularity for f then the Taylor series is not a promising approach to representing the function. However, one can instead use the ***Laurent series*** given by

$$f(z) = \sum_{j=0}^{\infty} a_j (z - z_0)^j + \sum_{j=1}^{\infty} \frac{b_j}{(z - z_0)^j}.$$

[2] Thus for any sequence $\{z_k\}$ converging to z_0, the sequence $\left\{\frac{f(z_k)-f(z_0)}{z_k-z_0}\right\}$ should converge, and should converge to the same complex number.

The matter of expressing the coefficients in terms of f obviously cannot be done by evaluations of f and its derivatives at z_0. However, there are formulas for the coefficients involving contour integrals. So...

D.3 Integration

Much of what interests in complex variable theory centres around integration. In this section we give a rapid overview of the essential facts.

A ***curve*** in $\mathbb{C}$ is a continuous map $c\colon [a,b] \to \mathbb{C}$. A ***closed curve*** in $\mathbb{C}$ is a curve $c\colon [a,b] \to \mathbb{C}$ for which $c(a) = c(b)$. Thus a closed curve forms a loop with no intersections (see Figure D.2). A curve c defined on $[a,b]$ is ***simple*** if the restriction of c to (a,b) is

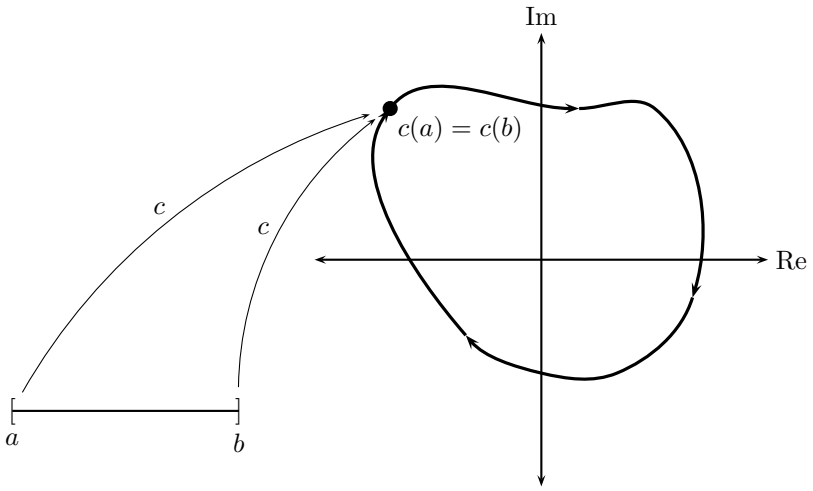


Figure D.2 A closed curve in $\mathbb{C}$

injective. Thus for each $t_1, t_2 \in (a,b)$ the points $c(t_1)$ and $c(t_2)$ are distinct. Sometimes a simple closed curve is called a ***Jordan curve***. The Jordan Curve Theorem then states that a simple closed curve separates $\mathbb{C}$ into two domains, the interior and the exterior. This also allows us to make sense of the ***orientation*** of a simple closed curve. We shall speak of simple closed curves as having "clockwise orientation" or "counterclockwise orientation." Let us agree not to give these precise notation as the meaning will be obvious in any application we encounter.

Sometimes we will wish for a curve to have more smoothness, and so speak of a ***differentiable curve*** as one where the functions $u, v\colon [a,b] \to \mathbb{R}$ defined by $c(t) = u(t) + iv(t)$ are differentiable. For short, we shall call a differentiable curve an ***arc***. In such cases we denote

$$c'(t) = \frac{\mathrm{d}u}{\mathrm{d}t} + i\frac{\mathrm{d}v}{\mathrm{d}t}.$$

The ***length*** of a differentiable curve $c\colon [a,b] \to \mathbb{C}$ is given by

$$\int_a^b |c'(t)|\,\mathrm{d}t.$$

A ***contour*** is a curve that is a concatenation of a finite collection of disjoint differentiable curves.

If $c\colon [a,b] \to \mathbb{C}$ is a curve then we define

$$\int_a^b c(t)\,\mathrm{d}t = \int_a^b u(t)\,\mathrm{d}t + i\int_a^b v(t)\,\mathrm{d}t,$$

where u and v are defined by $c(t) = u(t)+iv(t)$. Now we let D be a domain in $\mathbb{C}$, $c\colon [a,b] \to D$ be an arc, and $f\colon D \to \mathbb{C}$ be a continuous function. We define

$$\int_c f(z)\,\mathrm{d}z = \int_a^b f(c(t))c'(t)\,\mathrm{d}t. \tag{D.1}$$

One may verify that this integral does not in fact depend on the parameterisation of c, and so really only depends on the "shape" of the image of c in $U \subset \mathbb{C}$. We shall typically denote $C = \text{image}(c)$ and write $\int_C = \int_c$. If c is a contour, then one may similarly define the integral by defining it over each of the finite arcs comprising c. If $F\colon D \to \mathbb{C}$ is differentiable with continuous derivative f, then one verifies that

$$\int_c f(z)\,\mathrm{d}z = F(c(b)) - F(c(a)),$$

for a contour $c\colon [a,b] \to \mathbb{C}$.

The following theorem lies at the heart of much of complex analysis, and will be useful for us here.

D.3 Theorem **(Cauchy's Integral Theorem)** *Let $D \subset \mathbb{C}$ be a simply connected domain, suppose that $f\colon D \to \mathbb{C}$ is analytic on the closure of D, and let C be a simple closed contour contained in D. Then*

$$\int_C f(z)\,\mathrm{d}z = 0.$$

D.4 Applications of Cauchy's Integral Theorem

Cauchy's Integral Theorem forms the basis for much that is special in the theory of complex variables. We shall give a few of the applications that are of interest to us in this book.

Let us begin by providing formulas for the coefficients in the Laurent expansion in terms of contour integrals. The following result does the job.

D.4 Proposition *Let $f\colon D \to \mathbb{C}$ be analytic and let $z_0 \in D$ be an isolated singularity for f. Let C_0 and C_1 be circular contours centred at z_0 with C_1 smaller than C_0 (see Figure D.3). If*

$$f(z) = \sum_{j=-\infty}^{\infty} c_j(z-z_0)^j$$

is the Laurent series for f at z_0 then we have

$$c_j = \frac{1}{2\pi i}\int_{C_0} \frac{f(z)}{(z-z_0)^{j+1}}\,\mathrm{d}z, \quad j = 0, 1, \ldots$$

$$c_j = \frac{1}{2\pi i}\int_{C_1} \frac{f(z)}{(z-z_0)^{j+1}}\,\mathrm{d}z, \quad j = -1, -2, \ldots$$

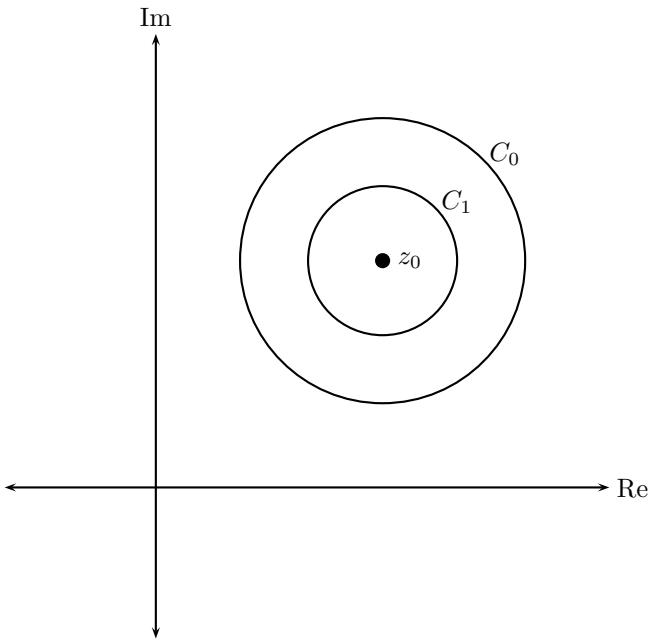


Figure D.3 Contours for definition of Laurent series coefficients

The ***residue*** of an analytic function f at an isolated singularity z_0 is the coefficient c_{-1} in the Laurent series for f at z_0. We denote the residue by

$$\operatorname*{Res}_{z=p_j} f(z) = \frac{1}{2\pi i}\int_C f(z)\,\mathrm{d}z,$$

where C is some sufficiently small circular contour centred at z_0. The Residue Theorem is also important for us.

D.5 Theorem **(Residue Theorem)** *Let $D \subset \mathbb{C}$ be a domain with C a simple, clockwise-oriented, closed contour in D. Let $f\colon D \to \mathbb{C}$ be meromorphic in the interior of C and analytic on C. Denote the poles of f in the interior of C by $p_1, \ldots, p_k$. Then*

$$\int_C f(s)\,\mathrm{d}s = 2\pi i \sum_{j=1}^{k} \operatorname*{Res}_{s=p_j} f(s).$$

D.6 Theorem

Another useful result is the ***Poisson Integral Formula***.

D.7 Theorem **(Poisson Integral Formula)** *Let $D \subset \mathbb{C}$ be a domain containing the positive complex plane $\overline{\mathbb{C}}_+$ and let $f\colon D \to \mathbb{C}$ be analytic in $\overline{\mathbb{C}}_+$. Additionally, we will suppose that if for $R > 0$ we define $m(R) > 0$ by*

$$m(R) = \sup_{\theta\in[-\frac{\pi}{2},\frac{\pi}{2}]} |f(Re^{i\theta})|, \qquad (D.2)$$

then f has the property that

$$\lim_{R\to\infty} \frac{m(R)}{R} = 0.$$

If $z_0 = x_0 + iy_0 \in \mathbb{C}_+$ *then we have*

$$f(z_0) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(iy)\frac{x_0}{x_0^2 + (x - x_0)^2}\,\mathrm{d}y.$$

The Poisson Integral Formula has the following useful corollary, stated by Freudenberg and Looze [1985].

D.8 Corollary *Suppose that* D *is a domain containing* $\overline{\mathbb{C}}$ *and that* $f\colon D \to \mathbb{C}$ *is analytic and nonzero in* $\overline{\mathbb{C}}$*, with the possible exception of zeros on the imaginary axis. Also, assume that* $\ln f$ *satisfies the equality (D.2). Then for each* $z_0 = x_0 + iy_0 \in \mathbb{C}_+$ *we have*

$$\ln|f(z_0)| = \frac{1}{\pi}\int_{-\infty}^{\infty} \ln|f(iy)|\frac{x_0}{x_0^2 + (x - x_0)^2}\,\mathrm{d}y.$$

Finally, we state a sort of stray result, but one that is standard in complex variable theory, the ***Maximum Modulus Principle***.

D.9 Theorem *If* $f\colon D \to \mathbb{C}$ *is an analytic function on a domain* D*, then* $|f|$ *has no maximum on* D *unless* f *is constant.*

From this result it follows that if f is analytic in a closed bounded region R, then the maximum value taken by $|f|$ must occur on the boundary of R.

analytic continuation

D.5 Algebraic functions and Riemann surfaces

In our discussion in Section 11.4 we shall need some not quite elementary concepts from the theory of complex variables. An ***algebraic function*** is a function f of a complex variable z satisfying an equation of the form

$$a_n(z)f(z)^n + \cdots + a_1(z)f(z) + a_0(z) = 0$$

where $a_0, a_1, \ldots, a_n \in \mathbb{C}[z]$. If a_n is not the zero polynomial then n is the ***degree*** of the algebraic function f. Upon reflection, one sees that there is a problem with making this definition precise since an algebraic function will not have as many as n possible solutions for each $z \in \mathbb{C}$. Thus an algebraic function is multi-valued. Since this is not an entirely clear notion, one should attempt to come up with a framework in which an algebraic function can be defined in a precise manner. The way in which this is done is by asking that an algebraic function take its values not in $\mathbb{C}$, but in what is called a "Riemann surface." A classical introductory reference is [Springer 1957].

Let us not formally define a what we mean by a Riemann surface, but proceed by example. We consider first the degree 1 case where f satisfies the equation

$$a_1(z)f(z) + a_0(z) = 0 \quad \Longrightarrow \quad f(z) = -\frac{a_0(z)}{a_1(z)} \in \mathbb{C}(z).$$

Thus degree 1 algebraic functions are simply rational functions. Let us examine some of the properties of such functions that may be helpful in our examination of higher-order Riemann surfaces. Suppose that f has poles at $p_1, \ldots, p_k \in \mathbb{C}$ with respective multiplicities $m_1, \ldots, m_k$. Write f using a complex partial fraction expansion:

$$f(z) = f_0(z) + f_1(z) + \cdots + f_n(z)$$

where $f_0 \in \mathbb{C}[z]$ and where

$$f_j(z) = \frac{c_{1,j}}{z - p_j} + \cdots + \frac{c_{m_j,j}}{(z - p_j)^{m_j}}, \quad j = 1, \ldots, k.$$

Now let R be a rational function in z and w; thus

$$R(z, w) = \frac{a_k(z)w^k + \cdots + a_1(z)w + a_0(z)}{b_\ell(z)w^\ell + \cdots + b_1(z)w + b_0(z)}$$

for $a_0, a_1, \ldots, a_k, b_0, b_1, \ldots, b_\ell \in \mathbb{C}[z]$. In the study of Riemann surfaces it is useful to examine integrals of the type

$$\int_{z_0}^{z} R(\zeta, f(\zeta))\, \mathrm{d}\zeta. \tag{D.3}$$

It is not clear why we should be interested in this, but let us look at this in the degree 1 case. In this case, since $f(z)$ is itself a rational function, $R(\zeta, f(\zeta))$ is a rational function in ζ, so has a partial fraction expansion. Using this, one may then explicitly evaluate the integral (D.3) as being the sum of rational functions and logarithmic functions.

Exercises

ED.1 Graphical calculation of residues from Truxal (page 27)

Appendix E

Fourier and Laplace transforms

For the most part, our use of transforms will be rather pedestrian. However, some of the technical material, especially in Chapter 15, requires that we actually know a little more than is often classically covered. Thus this appendix is a broader, although not terribly detailed, treatment of Fourier and Laplace transforms than may be required of students only engaging in the more straightforward parts of the book. Some uses of the Fourier or Laplace transforms benefit from a cursory knowledge of distributions. We begin our discussion with a presentation of distributions along these lines.

Contents

E.1 Delta-functions and distributions

When rigour is not a concern, a delta-function at 0 is a function $\delta(t)$ with the property that

$$\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1, \quad \int_{-\infty}^{\infty} f(t)\delta(t)\,\mathrm{d}t = f(0)$$

for any function f of unspecified character. It is actually not difficult to show that the existence of such a function is an impossibility. However, there is a way to rectify this in such a way that all the improper manoeuvres typically done with delta-functions are legal. The idea on making this precise is due to Schwartz [1950-1951].

E.1.1 Test functions

If $f\colon \mathbb{R} \to \mathbb{R}$ is a function, the ***support*** of f, denoted $\operatorname{supp}(f)$, is the closure of the set

$$\{x \in \mathbb{R} \mid f(x) \neq 0\}.$$

A ***test function*** on $\mathbb{R}$ is a function $\phi\colon \mathbb{R} \to \mathbb{R}$ with the properties that

1. ϕ is infinitely differentiable and

2. $\operatorname{supp}(\phi)$ is bounded.

The second condition normally is stated as ϕ having ***compact support***. We see that a test function vanishes except on some interval of finite length. Note that this precludes ϕ from being analytic. This makes it somewhat non-obvious how to define test functions. However, here is a common example of one such.

E.1 Example Let $\epsilon > 0$ and define

$$\phi_\epsilon(t) = \begin{cases} \exp(-\frac{\epsilon^2}{\epsilon^2 - t^2}), & |t| < \epsilon \\ 0, & |t| \geq \epsilon. \end{cases}$$

One may verify that this function is indeed infinitely differentiable, and it clearly has compact support. The function is plotted in Figure E.1. •

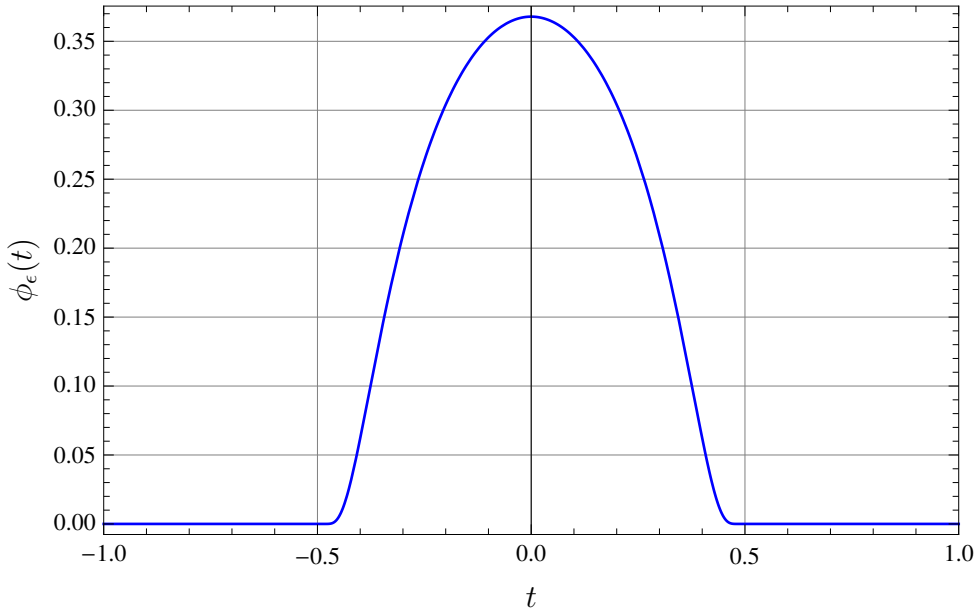


Figure E.1 The test function ϕ_ϵ when $\epsilon = 0.5$

Note that the set of test functions forms a vector space since the sum of two test functions is also a test function, and any scalar multiple of a test function is also a test function. This is then an infinite-dimensional vector space, and we denote it by $\mathscr{T}$. Let us define the notion of convergence in this vector space $\mathscr{T}$. A sequence of test functions $\{\phi_j\}_{j\in\mathbb{N}}$ ***converges to zero*** if

1. there exists an interval I of finite length so that $\operatorname{supp}(\phi_j) \subset I$ for all $j \in \mathbb{N}$ and
2. for each $k \in \mathbb{N}_0 = \{0\} \cup \mathbb{N}$, the sequence of functions $\{\phi_j^{(k)}\}_{j\in\mathbb{N}}$ converges uniformly to the zero function.

We then say that a sequence of test functions $\{\phi_j\}_{j\in\mathbb{N}}$ ***converges*** to a test function ϕ if the sequence $\{\phi_j - \phi\}_{j\in\mathbb{N}}$ converges to zero. A linear map $L\colon \mathscr{T} \to \mathbb{R}$ is ***continuous*** if the sequence $\{L(\phi_j)\}_{j\in\mathbb{N}}$ of numbers converges for every convergent sequence $\{\phi_j\}_{j\in\mathbb{N}}$ of test functions. If $I \subset \mathbb{R}$ is a closed interval of finite length, then $\mathscr{T}_I$ denotes the subspace of $\mathscr{T}$ consisting of those test functions ϕ for which $\operatorname{supp}(\phi) \subset I$.

E.1.2 Distributions

Finally, with the above terminology in place, a ***distribution*** is a linear map $L\colon \mathscr{T} \to \mathbb{R}$ having the property that the restriction of L to $\mathscr{T}_I$ is continuous for each closed interval I of finite length.

Let us consider some examples of distributions.

E.2 Example

1. Let $f\colon \mathbb{R} \to \mathbb{R}$ have the property that f is integrable over any interval of finite length. Then associated to f is the distribution L_f defined by
$$L_f(\phi) = \int_{-\infty}^{\infty} f(t)\phi(t)\,\mathrm{d}t.$$
Thus integrable functions may themselves be thought of as distributions. That is to say, distributions generalise functions.
2. Let us indicate that we may think of the delta-function as a distribution. Consider the linear map $\delta\colon \mathscr{T} \to \mathbb{R}$ defined by $L(\phi) = \phi(0)$. If $I \subset \mathbb{R}$ is a closed interval of finite length which does not contain 0 then the restriction of δ to $\mathscr{T}_I$ is obviously continuous: it is identically zero. Now let I be a closed finite-length interval containing 0. Let $\{\phi_j\}_{j\in\mathbb{N}}$ be a sequence of test functions converging to a test function ϕ and so that $\phi_j \in \mathscr{T}_I$ for $j \in \mathbb{N}$. By definition of convergence in $\mathscr{T}$, the sequence $\{\phi_j(0)\}_{j\in\mathbb{N}}$ converges to $\phi(0)$. This shows that the sequence $\{\delta(\phi_j)\}_{j\in\mathbb{N}}$ converges, so showing that δ is continuous on $\mathscr{T}_I$ as desired. Thus δ is indeed a distribution as we have defined it.
Note that the notion of a distribution suggests that we write $\delta(\phi)$ for the value of δ applied to a test function ϕ. However, custom dictates that we write
$$\delta(\phi) = \int_{-\infty}^{\infty} \delta(t)\phi(t)\,\mathrm{d}t = \phi(0).$$
3. If L is a distribution, then one may verify that the linear mapping $\dot{L}\colon \mathscr{T} \to \mathbb{R}$ defined by
$$\dot{L}(\phi) = -L(\dot{\phi})$$
is itself a distribution. This is the ***derivative*** of L, indicating that one can *always* differentiate a distribution. If f is a continuously differentiable function, then one may verify that $\dot{L}_f = L_{\dot{f}}$. Also, the derivative of the delta-function is defined by $\dot{\delta}(\phi) = -\dot{\phi}(0)$.
4. Let us combine 2 and 3 to show that $\delta(t)$ is the derivative of the unit step function $1(t)$. By definition of the derivative we have, for every test function ϕ,
$$\dot{1}(\phi) = -1(\dot{\phi}) = -\int_{-\infty}^{\infty} 1(t)\dot{\phi}(t)\,\mathrm{d}t = -\int_{0}^{\infty} \dot{\phi}(t) = -\phi(t)\big|_0^\infty = \phi(0),$$
as desired. •

It turns out that many of the manipulations valid for functions are valid for distributions. As we saw in Example 3 above, one can differentiate an arbitrary distribution in straightforward manner. Since distributions are generalisations of locally integrable functions, this by implication means that it is possible to define the derivative, *in the distributional sense*, of functions that are not differentiable! One says that a distribution has ***order*** k if k is the

smallest integer for which $L = f^{(k+1)}$ for an integrable function f. Thus the order measures how far away a distribution is from a being function. One may show that every distribution has a finite order. For example, since δ is the derivative of the locally integrable function $t \mapsto 1(t)$, δ has order 0. Likewise, $\dot{\delta}$ is order 1. If L is a distribution of order k and if f is at least k-times continuously differentiable, then it is possible to define the product of L with f to be the distribution fL given by $(fL)(\phi) = L(f\phi)$.

E.3 Example To multiply δ by a function and get a distribution, f has to be at least continuous since δ has order 0, In this case, if f is continuous, then $(f\delta)(\phi) = f(0)\phi(0)$ for $\phi \in \mathscr{T}$. •

If one can define the product fL for a function f and a distribution L, then the result can be differentiated in the distributional sense. One can easily show that the derivative of this product satisfies the usual product rule: $\frac{\mathrm{d}}{\mathrm{d}t}(fL) = \dot{f}L + f\dot{L}$, provided that the product $\dot{f}L$ makes sense as a distribution.

E.4 Example Let $f\colon \mathbb{R} \to \mathbb{R}$ be k-times continuously differentiable and define $g(t) = 1(t)f(t)$. This function may be differentiated n times in a distributional sense, and the derivatives are computed using the product rule:

$$\begin{aligned} g(t) &= 1(t)f(t) \\ g^{(1)}(t) &= f(0)\delta(t) + 1(t)f^{(1)}(t) \\ g^{(2)}(t) &= f(0)\delta^{(1)}(t) + \dot{f}(0)\delta(t) + 1(t)f^{(2)}(t) \\ &\vdots \\ g^{(n)}(t) &= \sum_{j=1}^{n} f^{(n-j)}(0)\delta^{(j-1)}(t) + 1(t)f^{(n)}(t). \end{aligned}$$

This formula is very useful in Section 3.6.2 when discussing the solution of differential equations using the left causal Laplace transform. •

As we have seen, a distribution is generally not representable as a function. However, what is true is that any distribution is a limit of a sequence of functions. This necessitates saying what we mean by convergence of distributions. A sequence $\{L_j\}_{j\in\mathbb{N}}$ of distributions converges to a distribution L if for every $\phi \in \mathscr{T}$, the sequence of numbers $\{L_j(\phi)\}_{j\in\mathbb{N}}$ converges to $L(\phi)$. What is then true is that every distribution is the limit of a sequence $\{f_j\}_{j\in\mathbb{N}}$ of infinitely differentiable functions, with these functions being regarded as distributions.

E.5 Example One can show that $\delta = \lim_{j\to\infty} f_j$ where

$$f_j(t) = \frac{j}{\sqrt{2\pi}} \exp\Big(-\frac{n^2t^2}{2}\Big).$$

In Figure E.2 we plot f_j for a couple of values of j, and we can see the anticipated behaviour of concentration of the function near 0. •

E.2 The Fourier transform

Before we begin talking about transforms, we need a few basic notions of integration. Some of this is discussed in more generality and detail in Section 5.3. A function

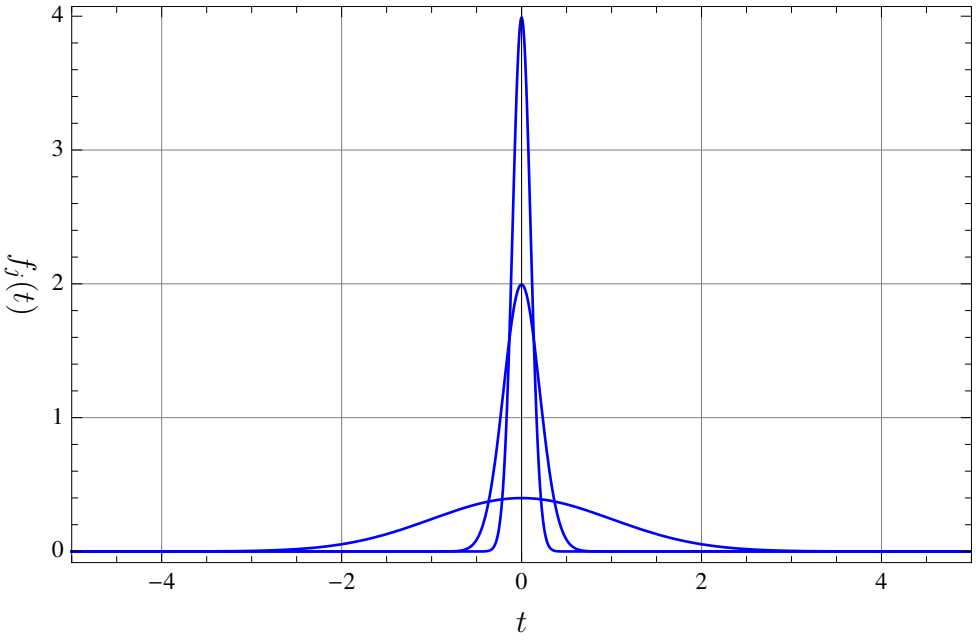


Figure E.2 A sequence of functions converging to δ

$f\colon(-\infty,\infty)\to\mathbb{C}$[1] is $\mathbf{L_2}$-***integrable*** if

$$\Big(\int_{-\infty}^{\infty}|f(t)|^2\,\mathrm{d}t\Big)^{1/2}<\infty.$$

In this case, we denote by $\|f\|_2$ the quantity on the left-hand side of the inequality. In like manner, a function $f\colon(-\infty,\infty)\to\mathbb{C}$ is $\mathbf{L_1}$-***integrable*** if

$$\int_{-\infty}^{\infty}|f(t)|\,\mathrm{d}t<\infty.$$

In this case, we denote by $\|f\|_1$ the quantity on the left-hand side of the inequality. Finally, a function $f\colon(-\infty,\infty)\to\mathbb{C}$ is $\mathbf{L_\infty}$-***integrable*** if $|f(t)|<\infty$ for almost every $t\in\mathbb{R}$. In this case, we denote by $\|f\|_\infty$ the least upper bound of $|f(t)|$, $t\in\mathbb{R}$.

With these notions of integrability at hand, we may define the ***Fourier transform*** of an L_1-integrable function $f\colon(-\infty,\infty)\to\mathbb{C}$ as the function $\check{f}\colon(-\infty,\infty)\to\mathbb{C}$ given by given by

$$\check{f}(\omega)=\int_{-\infty}^{\infty}f(t)e^{-i\omega t}\,\mathrm{d}t.$$

Note that $|\check{f}(\omega)|\le\|f\|_1$, so that $\check{f}$ is L_∞-integrable. If we wish to emphasise that we are transforming an L_1-integrable function, we shall state that we are using the $\mathbf{L_1}$-***Fourier transform***. It is also possible to take the Fourier transform of an L_2-integrable function $f\colon(-\infty,\infty)\to\mathbb{C}$, and this is done as follows. For $T>0$ define $f_T\colon(-\infty,\infty)\to\mathbb{C}$ by

$$f_T(t)=\begin{cases}f(t), & |t|\le T\\ 0, & \text{otherwise.}\end{cases}$$

[1]We deal in this appendix always with $\mathbb{C}$-valued functions of t. In the majority of instances in the text, the functions are $\mathbb{R}$-valued. However, for the general presentation here, it is convenient to consider $\mathbb{C}$-valued functions.

One may verify that f_T is L_1-integrable for any finite T, so that its L_1-Fourier transform, $\check{f}_T$, exists. What's more, it can be shown that there exists an L_2-integrable function $\check{f}\colon(-\infty,\infty)\to\mathbb{C}$ so that

$$\lim_{T\to\infty}\int_{-\infty}^{\infty}|\check{f}(\omega)-\check{f}_T(\omega)|\,\mathrm{d}t=0.$$

Thus the functions $\check{f}_T$ ***converge in mean*** to the function $\check{f}$ which we call the **L_2-*Fourier transform*** of the L_2-integrable function f. Note that the L_2-Fourier transform has the interesting property of mapping L_2-integrable functions to L_2-integrable functions. The inverse of the L_2-Fourier transform must therefore exist. Indeed, given an L_2-integrable function $\check{f}\colon\mathbb{R}\to\mathbb{C}$, its ***inverse Fourier transform*** is given by the L_2-integrable function $f\colon(-\infty,\infty)\to\mathbb{C}$ defined by

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\check{f}(\omega)e^{i\omega t}\,\mathrm{d}\omega.$$

It will on occasion be convenient to use the notation $\mathscr{F}$ for the map that sends an L_1-integrable function to its Fourier transform. Thus $\mathscr{F}(f)=\check{f}$. We likewise denote the inverse for the L_2-Fourier transform by $\mathscr{F}^{-1}$.

For L_2-integrable functions f and g, one can readily verify ***Parseval's Theorem*** which states that

$$\int_{-\infty}^{\infty}\bar{f}(t)g(t)\,\mathrm{d}t=\frac{1}{2\pi}\int_{-\infty}^{\infty}\bar{\check{f}}(\omega)\check{g}(\omega)\,\mathrm{d}\omega.$$

In particular, it follows that $\|f\|_2=\frac{1}{\sqrt{2\pi}}\|\check{f}\|_2$. Also important to us is the notion of ***convolution***. Given L_1-integrable functions f and g, we define a new L_1-integrable function, denoted $f*g$ and called the convolution of f with g, defined by

$$(f*g)(t)=\int_{-\infty}^{\infty}f(t-\tau)g(\tau)\,\mathrm{d}\tau.$$

Convolution in the time-domain is equivalent to multiplication in the frequency domain:

$$(f*g)\check{}=\check{f}\check{g}.$$

We shall encounter convolution in various contexts using both Fourier and Laplace transforms. It is important to realise just what kind of convolution with which one is dealing!

E.3 The Laplace transform

The Laplace transform is related to the Fourier transform in a sort of simple way. However, since the relationship can get a little complicated, we try to be clear about the notion of the Laplace transform. Also, we will consider Laplace transforms of various sorts of functions, so we must take care to distinguish these with proper notation. We remark that this is not normally done in control texts, with the result that there are sometimes contradictions present that apparently go unnoticed. A good introductory discussion of the Laplace transform, minus any detailed discussion of distributions, can be found in [Schiff 1999]. The distributional theory is carried out in detail by Zemanian [1965].

E.3.1 Laplace transforms of various flavours

Let us first establish some notation. Let $f\colon (-\infty,\infty)\to\mathbb{C}$ be an L_1-integrable function. If $\mathrm{Re}(s)=0$ then the integral

$$\int_{-\infty}^{\infty} f(t)e^{-st}\,\mathrm{d}t \tag{E.1}$$

exists; it is, of course, simply the Fourier transform. By continuity of the integral, if $\mathrm{Re}(s)$ is sufficiently close to zero, the integral (E.1) exists. To define the Laplace transform, we wish to ascertain a subset of $\mathbb{C}$ so that if s lies in this set, the integral (E.1) exists. In doing so, we do not necessarily require that f itself be L_1-integrable.

Let f have the property that the function $f(t)e^{-ct}$ is L_1-integrable for some $c\in\mathbb{R}$. That is, for some $c\in\mathbb{R}$ we have

$$\int_{-\infty}^{\infty} |f(t)|e^{-ct}\,\mathrm{d}t<\infty. \tag{E.2}$$

For such functions, call the real number

$$\sigma_{\min}(f)=\inf\{c\in\mathbb{R}\mid \text{ the inequality (E.2) is satisfied}\}$$

the ***minimum abscissa of absolute convergence***. Similarly we define

$$\sigma_{\max}(f)=\sup\{c\in\mathbb{R}\mid \text{ the inequality (E.2) is satisfied}\}.$$

We call $\sigma_{\max}(f)$ the ***maximum abscissa of absolute convergence***. We then define

$$\mathscr{R}_{\mathrm{c}}(f)=\{s\in\mathbb{C}\mid \sigma_{\min}(f)<\mathrm{Re}(s)<\sigma_{\max}(f)\}$$

which we call the ***region of absolute convergence***. We define the ***two-sided Laplace transform*** of f as $\mathscr{L}(f)\colon \mathscr{R}_{\mathrm{c}}(f)\to\mathbb{C}$ defined by

$$\mathscr{L}(f)(s)=\int_{-\infty}^{\infty} f(t)e^{-st}\,\mathrm{d}t. \tag{E.3}$$

For functions $f\colon(-\infty,\infty)\to\mathbb{C}$ that have the property that $f(t)e^{-ct}$ is L_2-integrable, the ***inverse Laplace transform*** exists, and is defined by

$$\mathscr{L}^{-1}(\mathscr{L}(f))=\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty}\mathscr{L}(f)(s)e^{st},$$

where σ is any number satisfying $\sigma_{\min}(f)<\sigma<\sigma_{\max}(f)$.

Most commonly we will be dealing with the Laplace transform of functions that are zero for negative times. We will also want to be able to take Laplace transforms of distributions involving delta-functions and derivatives of delta-functions at 0. In these cases, it matters when performing an integral over $[0,\infty)$ whether we take the lower limit as $0+$ or $0-$. To be precise, suppose that $f_0\colon(-\infty,\infty)\to\mathbb{C}$ possesses a two-sided Laplace transform and has the property that $f_0(t)=0$ for $t<0$. Now consider the distribution

$$f(t)=f_0(t)+\sum_{j=0}^{k}c_j\delta^{(j)}(t). \tag{E.4}$$

Thus f is a distribution that is a sum of a function which vanishes for negative times and a finite sum of delta-functions and derivatives of delta-functions. One can consider taking the

Laplace transform of more general distributions, but we shall not need this level of generality, and its use necessitates a significant diversion [see Zemanian 1965]. A distribution of the form (E.4), and for which f_0 vanishes for negative times and possesses a two-sided Laplace transform, will be called ***simple***. For simple distributions we define two kinds of Laplace transforms. The ***left causal Laplace transform*** for a simple distribution f as given in (E.4) is the map $\mathscr{L}^+_{0-}(f)\colon \mathscr{R}_{\mathrm{c}}(f_0) \to \mathbb{C}$ defined by

$$\mathscr{L}^+_{0-}(f)(s) = \lim_{\epsilon \downarrow 0} \int_{-\epsilon}^{\infty} f(t)e^{-st}\,\mathrm{d}s = \int_{0-}^{\infty} f(t)e^{-st}\,\mathrm{d}s.$$

Note that the left causal transform includes the effect of the delta-functions. Indeed, if f is k-times continuously differentiable we can simply evaluate the contributions of the delta-functions and see that

$$\mathscr{L}^+_{0-}(f)(s) = \mathscr{L}(f_0) + \sum_{j=0}^{k} c_j s^j.$$

In contrast to this, the ***right causal Laplace transform*** does not include the contributions of the delta-functions. It is defined to be the map $\mathscr{L}^+_{0+}(f)\colon \mathscr{R}_{\mathrm{c}}(f_0) \to \mathbb{C}$ defined by

$$\mathscr{L}^+_{0+}(f)(s) = \lim_{\epsilon \downarrow 0} \int_{\epsilon}^{\infty} f(t)e^{-st}\,\mathrm{d}s = \int_{0+}^{\infty} f(t)e^{-st}\,\mathrm{d}s.$$

In this case, since the delta-functions do not contribute, we simply have $\mathscr{L}^+_{0+}(f)(s) = \mathscr{L}(f_0)$. Thus for genuine functions f (i.e., that do not involve distributions), we have

$$\mathscr{L}(f) = \mathscr{L}^+_{0-}(f) = \mathscr{L}^+_{0+}(f),$$

provided that $f(t) = 0$ for $t < 0$. Note that for functions f vanishing for negative times we always have $\sigma_{\max}(f) = \infty$.

We will also have occasion to consider functions of time that are zero for *positive* times. As above, let f be a simple distribution as given by (E.4), but now assume that $f_0(t) = 0$ for $t > 0$. We shall of course assume that the two-sided Laplace transform of f_0 still exists. The ***left anticausal Laplace transform*** of f is the map $\mathscr{L}^-_{0-}(f)\colon \mathscr{R}_{\mathrm{c}}(f_0) \to \mathbb{C}$ defined by

$$\mathscr{L}^-_{0-}(f)(s) = \int_{-\infty}^{0-} f(t)e^{-st}\,\mathrm{d}s.$$

Note that the left anticausal transform does not include the effect of the delta-functions. Thus we may write $\mathscr{L}^-_{0-}(f)(s) = \mathscr{L}(f_0)$. Proceeding in the natural manner the ***right anticausal Laplace transform*** is defined to be the map $\mathscr{L}^-_{0+}(f)\colon \mathscr{R}_{\mathrm{c}}(f_0) \to \mathbb{C}$ defined by

$$\mathscr{L}^-_{0+}(f)(s) = \int_{-\infty}^{0+} f(t)e^{-st}\,\mathrm{d}s.$$

This transform does include the effects of the delta-functions. If f is k-times continuously differentiable we can simply evaluate the contributions of the delta-functions and see that

$$\mathscr{L}^-_{0+}(f)(s) = \mathscr{L}(f_0) + \sum_{j=0}^{k} c_j s^j.$$

Thus for genuine functions f (i.e., that do not involve distributions), we have

$$\mathscr{L}(f) = \mathscr{L}^-_{0-}(f) = \mathscr{L}^-_{0+}(f),$$

provided that $f(t) = 0$ for $t > 0$. For anticausal functions f we always have $\sigma_{\min}(f) = -\infty$.

E.6 Notation *When there is no possibility of confusion, we shall denote the Laplace transform of a function f by $\hat{f}$. If f is a simple distribution that is not a function, we shall always specify which transform we use. However, if f is a function with no distributional component, the various transforms are all equal, and there is no potential for confusion, unless one wants to consider the Laplace transform of the derivative, cf. Theorem E.7.* •

E.3.2 Properties of the Laplace transform

An important property of the Laplace transform is how it acts with respect to the derivative. This is a place where, as we see, the left and right transforms differ.

E.7 Theorem *Let $f : \mathbb{R} \to \mathbb{R}$ be causal and continuous on $[0, \infty)$ and suppose that $\dot{f}$ is piecewise continuous on $[0, \infty)$. If the Laplace transform of f exists then*

$$\mathscr{L}^+_{0+}(\dot{f})(s) = s\mathscr{L}^+_{0+}(f)(s) - f(0+)$$

and

$$\mathscr{L}^+_{0-}(\dot{f})(s) = s\mathscr{L}^+_{0+}(f)(s) = s\mathscr{L}^+_{0-}(f)(s)$$

for $s \in \mathscr{R}_{\mathrm{c}}(f)$.

Proof We use integration by parts to compute

$$\begin{aligned}\mathscr{L}^+_{0+}(\dot{f})(s) &= \lim_{\substack{\epsilon\to 0+\\ R\to\infty}} \int_\epsilon^R \dot{f}(t)e^{-st}\,\mathrm{d}t \\ &= \lim_{\substack{\epsilon\to 0+\\ R\to\infty}} f(t)e^{-st}\Big|_\epsilon^R + \lim_{\substack{\epsilon\to 0+\\ R\to\infty}} \int_\epsilon^R f(t)se^{-st}\,\mathrm{d}t.\end{aligned}$$

Now, if $s \in \mathscr{R}_{\mathrm{c}}(f)$ then it must be the case that $\lim_{R\to\infty} f(R)e^{-sR} = 0$. Taking also the limit as $\epsilon \to 0+$ we see that

$$\mathscr{L}^+_{0+}(\dot{f})(s) = s\mathscr{L}^+_{0+}(f)(s) - f(0+)$$

as stated.

For the second part of the theorem, define $\tilde{f} : \mathbb{R} \to \mathbb{R}$ by

$$\tilde{f}(x) = \begin{cases} f(x), & x \geq 0 \\ f(0), & x < 0. \end{cases}$$

Thus f is continuous on $\mathbb{R}$ and $\dot{f}$ is piecewise continuous on $\mathbb{R}$. We then have $f(t) = 1(t)\tilde{f}(t)$ where $1(t)$ is the unit step function, and so we obtain

$$\dot{f}(t) = \dot{1}(t)\tilde{f}(t) + 1(t)\dot{\tilde{f}}(t),$$

where $\dot{1}$ is to be understood in the distribution sense, i.e., $\dot{1}(t) = \delta(t)$. We now compute

$$\begin{aligned}
\mathscr{L}^+_{0-}(\dot{f})(s) &= \lim_{\substack{\epsilon\to 0+\\ R\to\infty}} \int_{-\epsilon}^{R} \dot{f}(t)e^{-st}\,\mathrm{d}t \\
&= \lim_{\substack{\epsilon\to 0+\\ R\to\infty}} \int_{-\epsilon}^{R} \big(\delta(t)\tilde{f}(t) + 1(t)\dot{\tilde{f}}(t)\big)e^{-st}\,\mathrm{d}t \\
&= \tilde{f}(0) + \lim_{\substack{\epsilon\to 0+\\ R\to\infty}} \int_{-\epsilon}^{R} 1(t)\dot{\tilde{f}}(t)e^{-st}\,\mathrm{d}t \\
&= f(0) + \lim_{\substack{\epsilon\to 0+\\ R\to\infty}} \int_{-\epsilon}^{R} f(t)e^{-st}\,\mathrm{d}t \\
&= f(0) + s\mathscr{L}^+_{0-}(f)(s) - f(0) = s\mathscr{L}^+_{0-}(f)(s).
\end{aligned}$$

Here we use the fact that since f does not involve a delta-function at $t = 0$, we have $\mathscr{L}^+_{0-}(f)(s) = \mathscr{L}^+_{0+}(f)(s)$. ■

We see that the right causal Laplace transform has the capacity to involve the value of f at $0+$. For this reason, it is useful to use this transform when solving initial value problems, if that is your preferred way to do such things. However, for general control theoretic discussions, the left causal Laplace transform is often the preferred tool since it provides a simple relationship between the Laplace transform of a function and its derivative. In any case, both shall appear at certain times in the text.

A repeated application of Theorem E.7 gives the following result.

E.8 Corollary *Let $f\colon \mathbb{R} \to \mathbb{R}$ be causal and suppose that $f, f^{(1)}, \dots, f^{(n-1)}$ are continuous on $[0,\infty)$ and that $f^{(n)}$ is piecewise continuous on $[0,\infty)$. Then*

$$\mathscr{L}^+_{0+}(f)(s) = s^n\mathscr{L}^+_{0+}(f)(s) - \sum_{j=0}^{n-1} s^j y^{(n-j-1)}(0+)$$

and

$$\mathscr{L}^+_{0-}(f)(s) = s^n\mathscr{L}^+_{0-}(f)(s)$$

if $s \in \mathscr{R}_{\mathrm{c}}(f)$.

The convolution also has the same useful properties for Laplace transforms as for Fourier transforms. To be clear, suppose that the Laplace transforms of $f, g\colon (-\infty,\infty) \to \mathbb{C}$ exist and that $\mathscr{R}_{\mathrm{c}}(f) \cap \mathscr{R}_{\mathrm{c}}(g) \neq \emptyset$. If the convolution $f * g$ is defined, then its Laplace transform is defined, and we further have

$$\mathscr{L}(f * g) = \mathscr{L}(f)\mathscr{L}(g);$$

thus the Laplace transform of the convolution is the product of the Laplace transforms, and it is defined on the region $\mathscr{R}_{\mathrm{c}}(f) \cap \mathscr{R}_{\mathrm{c}}(g) \subset \mathbb{C}$. In the text, we shall consider the cases when f and g are both zero for either positive or negative times. For example, if f is a causal function then

$$(f * g)(t) = \int_0^\infty f(\tau)g(t-\tau)\,\mathrm{d}\tau$$

and if g is a causal function then

$$(f * g)(t) = \int_0^\infty f(t-\tau)g(\tau)\,\mathrm{d}\tau.$$

Similar statements hold for anticausal functions.

The following result is often helpful.

E.9 Proposition *Let $f\colon \mathbb{R} \to \mathbb{R}$ be a causal function, continuous on $[0,\infty)$, and suppose that $\dot f$ is piecewise continuous on $[0,\mathbb{R})$. If the Laplace transform of f exists, then the following statements hold:*

(i) $f(0+) = \lim_{s\to\infty} s\mathscr{L}^+_{0+}(f)(s)$ *(s real);*

(ii) $\lim_{t\to\infty} f(t) = \lim_{s\to 0} s\mathscr{L}^+_{0+}(f)(s)$ *(s real), provided that the limit on the left exists.*

Proof (i) By Theorem E.7 we have

$$\int_{0+}^\infty \dot f(t)e^{-st}\,\mathrm{d}t = s\mathscr{L}^+_{0+}(f)(s) - f(0+).$$

If we take the limits as $s \to \infty$ we may switch the limit and the integral since the integral converges absolutely when $s \geq 0$. This gives

$$0 = \lim_{s\to\infty} s\mathscr{L}^+_{0+}(f)(s) - f(0+)$$

from which our first assertion follows.

(ii) By Theorem E.7 we have

$$\int_{0+}^\infty \dot f(t)e^{-st}\,\mathrm{d}t = s\mathscr{L}^+_{0+}(f)(s) - f(0+).$$

We take the limit as $s \to 0$ of both sides, and move the limit inside the integral since the integral is absolutely convergent in a neighbourhood of $s = 0$. This gives

$$\begin{aligned} &\int_{0+}^\infty \dot f(t)\,\mathrm{d}t = \lim_{s\to 0} s\mathscr{L}^+_{0+}f)(s) - f(0+) \\ \implies \quad &\lim_{t\to\infty} f(t) - f(0+) = \lim_{s\to 0} s\mathscr{L}^+_{0+}(f)(s) - f(0+) \end{aligned}$$

from which the result follows, under the proviso that $\lim_{t\to\infty} f(t)$ exists. ■

The second assertion of the proposition is often called the ***Final Value Theorem***, and it does require the hypothesis that $\lim_{t\to\infty} f(t)$ exist.

The following result is one of a similar nature, but involves the integral of the function of time in terms of its Laplace transform. These results will be interesting for us in Section 8.3 when we discuss various aspects of controller performance.

E.10 Proposition *Let $f(t)$ have the property that its Laplace transform $\mathscr{L}^+_{0+}(f)(s)$ is a strictly proper rational function with the property that there exists $\alpha > 0$ so that if s is a pole of $\mathscr{L}^+_{0+}(f)(s)$ then $\mathrm{Re}(s) \leq -\alpha$. Then, for any s_0 with $\mathrm{Re}(s_0) > -\alpha$ we have*

$$\int_0^\infty e^{-s_0 t} f(t)\,\mathrm{d}t = \lim_{s\to s_0} \mathscr{L}^+_{0+}(f)(s).$$

Proof Note that $\sigma_{\min}(f) \leq -\alpha$, and so if $\mathrm{Re}(s_0) > -\alpha$, then s_0 is in the domain of definition of the transform. Therefore the integral

$$\int_0^\infty e^{-s_0 t} f(t)\,\mathrm{d}t$$

exists and is equal to $\mathscr{L}^+_{0+}(f)(s_0)$. ■

E.3.3 Some useful Laplace transforms

You will recall some standard Laplace transforms which we collect in Table E.1. Here

$f(t)$	$\hat{f}(s)$	$\sigma_{\min}(f)$
$f(t)+g(t)$	$\hat{f}(s)+\hat{g}(s)$	unknown
$af(t)$	$a\hat{f}(s)$	$\sigma_{\min}(f)$
$\dot{f}(t)$	$-f(0+)+s\hat{f}(s)$	unknown
$1(t)$	$\dfrac{1}{s}$	0
$R_b(t)$	$\dfrac{b}{s^2}$	0

$f(t)$	$\hat{f}(s)$	$\sigma_{\min}(f)$
e^{at}	$\dfrac{1}{s-a}$	a
$\sin\omega t$	$\dfrac{\omega}{s^2+\omega^2}$	0
$\cos\omega t$	$\dfrac{s}{s^2+\omega^2}$	0

Table E.1 Some common Laplace transforms (an "unknown" means that is it not generally determinable in terms of $\sigma_{\min}(f)$)

$1(t)$ is the unit step function defined by

$$1(t) = \begin{cases} 1, & t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

and $R_b(t)$ is the ramp function defined by

$$R_b(t) = \begin{cases} bt, & t \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

We will also need some not so common Laplace transforms in order to make a conclusive stability analysis using the impulse response. You can look up that

ref

$$\begin{aligned} & f(t) = t^k e^{at} \sin\omega t \\ \Longrightarrow \quad & \hat{f}(s) = \left(\sum_{0 \leq 2m \leq k} (-1)^m \binom{k+1}{2m+1} \omega^{2m+1} (s-a)^{k-2m} \right) \frac{k!}{\left((s-a)^2+\omega^2\right)^{k+1}} \end{aligned}$$

and

$$\begin{aligned} & f(t) = t^k e^{at} \cos\omega t \\ \Longrightarrow \quad & \hat{f}(s) = \left(\sum_{0 \leq 2m \leq k+1} (-1)^m \binom{k+1}{2m} \omega^{2m} (s-a)^{k-2m+1} \right) \frac{k!}{\left((s-a)^2+\omega^2\right)^{k+1}} \end{aligned}$$

where, you recall, for integers k and ℓ with $k \geq \ell$ we have

$$\binom{k}{\ell} = \frac{k!}{\ell!(k-\ell)!}.$$

These are pretty ugly expressions. Let's observe some general features. The Laplace transforms $\hat{f}(s)$ are polynomials where the degree of the numerator is strictly less than that of

the denominator. As special cases of these formulas we have

$$\begin{aligned}
f(t) = t^k e^{at} \quad &\Longrightarrow \quad \hat{f}(s) = \frac{k!}{(s-a)^{k+1}} \\
f(t) = e^{at} \sin \omega t \quad &\Longrightarrow \quad \hat{f}(s) = \frac{\omega}{(s-a)^2 + \omega^2} \\
f(t) = e^{at} \cos \omega t \quad &\Longrightarrow \quad \hat{f}(s) = \frac{s-a}{(s-a)^2 + \omega^2} \\
f(t) = t \sin \omega t \quad &\Longrightarrow \quad \hat{f}(s) = \frac{2\omega s}{(s^2 + \omega^2)^2} \\
f(t) = t \cos \omega t \quad &\Longrightarrow \quad \hat{f}(s) = \frac{s^2 - \omega^2}{(s^2 + \omega^2)^2}.
\end{aligned}$$

We shan't fiddle much with the definition of the inverse Laplace transform, but it will be helpful to know a couple of them which cannot be obviously gleaned from Table E.1. We have

$$\begin{aligned}
&\hat{f}(s) = \frac{1}{\big((s-\sigma)^2 + \omega^2\big)^k} \\
\Longrightarrow \quad &f(t) = \frac{-e^{\sigma t}}{4^{k-1}\omega^{2k}} \sum_{i=1}^{k} \binom{2k-i-1}{k-1} (-2t)^{i-1} \frac{\mathrm{d}^i}{\mathrm{d}t^i} \cos \omega t
\end{aligned}$$

and

$$\begin{aligned}
&\hat{f}(s) = \frac{s}{\big((s-\sigma)^2 + \omega^2\big)^k} \\
\Longrightarrow \quad &f(t) = \frac{e^{\sigma t}}{4^{k-1}\omega^{2k}} \Big(\sum_{i=1}^{k} \binom{2k-i-1}{k-1} \frac{(-2t)^{i-1}}{(i-1)!} \frac{\mathrm{d}^i}{\mathrm{d}t^i} \big(\sigma \cos \omega t + \omega \sin \omega t\big) - \\
&2\omega \sum_{i=1}^{k-1} \binom{2k-i-2}{k-1} \frac{(-2t)^{i-1}}{(i-1)!} \frac{\mathrm{d}^i}{\mathrm{d}t^i} \sin \omega t \Big)
\end{aligned}$$

The import of these ridiculous expressions is contained in the following, now self-evident, statement.

E.11 Proposition *The functions*

$$\frac{1}{\big((s-\sigma)^2 + \omega^2\big)^k}, \quad \frac{s}{\big((s-\sigma)^2 + \omega^2\big)^k}, \qquad \sigma, \omega \in \mathbb{R},\ k \in \mathbb{N},$$

are in one-to-one correspondence with the Laplace transforms of functions which are linear combinations of

$$t^\ell e^{\sigma t} \sin \omega t \ \text{ and } \ t^\ell e^{\sigma t} \cos \omega t, \quad \sigma, \omega \in \mathbb{R},\ k \in \mathbb{N},\ \ell = 0, \dots, k-1. \tag{E.5}$$

Note that we have in this section established (without proof) a perfect correspondence between functions which are rational in the Laplace transform variable s, and functions of time of the form (E.5). That is to say, a strictly proper rational function in s is the Laplace transform of a linear combination of functions like (E.5) and the Laplace transform of a function like (E.5) is a rational function in s. This fact will be useful in our later investigations of the behaviour of SISO linear systems.

Exercises

EE.1 What are the abscissa of absolute convergence, $\sigma_{\min}(f)$ and $\sigma_{\max}$, for the function $f(t) = 1(t)e^{at}$? How are the values of $\sigma_{\min}(f)$ and $\sigma_{\max}(f)$ reflected in the properties of the Laplace transform for f?

EE.2 Show that $\sigma_{\min}(f) = \sigma$ and $\sigma_{\max}(f) = \infty$ for functions f of the form $f(t) = 1(t)t^\ell e^{\sigma t} \sin \omega t$ or $f(t) = 1(t)t^\ell e^{\sigma t} \sin \omega t$.

EE.3 Let $f(t) = 1(t) \sin t$. Compute $\lim_{s\to 0} s\hat{f}(s)$ and determine whether $\lim_{t\to\infty} f(t) = \lim_{s\to 0} s\hat{f}(s)$. Explain your conclusion in terms of Proposition E.9(ii).

EE.4 Show that the Laplace transform of $e^{\boldsymbol{A}t}$ is $(s\boldsymbol{I}_n - \boldsymbol{A})^{-1}$. What do you think are the abscissa of absolute convergence?

EE.5 Let f and g be functions whose causal right Laplace transforms exist and whose domains have a nonempty intersection. Prove that the inverse Laplace transform of $\hat{f}(s)\hat{g}(s)$ is either of the expressions

$$\int_0^t f(t-\tau)g(\tau)\,\mathrm{d}\tau, \quad \int_0^t g(t-\tau)f(\tau)\,\mathrm{d}\tau.$$

EE.6 Fix $T > 0$ and a function g whose causal right Laplace transform exists and is $\hat{g}$. Show that the causal right Laplace transform of the function $f(t) = g(t - T)$ is $\hat{f}(s) = e^{-Ts}\hat{g}(s)$. The function f is called the ***time delay*** of g by T.

EE.7 Using the adjugate (see Section A.3.1), determine the inverse of the matrix $s\boldsymbol{I}_3 - \boldsymbol{A}$ where $s \in \mathbb{R}$ and

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ -1 & 1 & -1 \end{bmatrix}.$$

You may suppose that s is such that $\det(s\boldsymbol{I}_3 - \boldsymbol{A}) \neq 0$.

Bibliography

Ackermann, J. E. [1972]. "Der entwurf linearer regelungsysteme im zustandsraum". *Regelungstech Prozess-Datenverarbeitung* 7, pages 297–300. ISSN: 0034-3226.

Adamjan, V. M., Arov, D. Z., and Krein, M. G. [1968a]. "Infinite Hankel matrices and generalized problems of Caratheodory–Fejer and F. Riesz". *Functional Analysis and its Applications. Translation of* Rossiĭskaya Akademiya Nauk. Funktsional′nyĭ Analiz i ego Prilozheniya 2(1), pages 1–19. ISSN: 0016-2663. DOI: 10.1007/BF01075356.

— [1968b]. "Infinite Hankel matrices and generalized problems of Caratheodory–Fejer and I. Schur". *Functional Analysis and its Applications. Translation of* Rossiĭskaya Akademiya Nauk. Funktsional′nyĭ Analiz i ego Prilozheniya 2(4), pages 1–17. ISSN: 0016-2663. DOI: 10.1007/BF01075679.

— [1971]. "Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur–Takagi problem". *Mathematics of the USSR-Sbornik* 15(1), pages 31–73. ISSN: 0025-5734. DOI: 10.1070/SM1971v015n01ABEH001531.

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.

Basile, G. and Marro, G. [1968]. "Luoghi caratteristici dello spazio degli stati relativi al controllo dei sistemi lineari". *L'Elettrotecnica* 55(12), pages 1–7. ISSN: 0013-6131.

Blondel, V. D. and Tsitsiklis, J. M. [1997]. "NP-hardness of some linear control design problems". *SIAM Journal on Control and Optimization* 35(6), pages 2118–2127. ISSN: 0363-0129. DOI: 10.1137/S0363012994272630.

Bode, H. W. [1945]. *Network Analysis and Feedback Amplifier Design*. Reprint: [Bode 1975]. Van Nostrand Reinhold Co.: London.

— [1975]. *Network Analysis and Feedback Amplifier Design*. Original: [Bode 1945]. Robert E. Krieger Publishing Company: Huntington/New York. ISBN: 978-0-88275-242-6.

Bott, R. and Duffin, R. J. [1949]. "Impedance synthesis without use of transformers". *Journal of Applied Physics* 20, page 816. ISSN: 0021-8979. DOI: 10.1063/1.1698532.

Boyce, W. E. and Diprima, R. C. [1972]. *Elementary Differential Equations with Boundary Value Problems*. New edition: [Boyce and Diprima 2012]. John Wiley and Sons: NewYork, NY.

— [2012]. *Elementary Differential Equations with Boundary Value Problems*. 10th edition. First edition: [Boyce and Diprima 1972]. John Wiley and Sons: NewYork, NY. ISBN: 978-0-470-45831-0.

Brockett, R. W. [1970]. *Finite Dimensional Linear Systems*. John Wiley and Sons: NewYork, NY. ISBN: 978-0-471-10585-5.

Bryson, A. E. and Ho, Y. C. [1975]. *Applied Optimal Control. Optimization, Estimation and Control*. John Wiley and Sons: NewYork, NY. ISBN: 978-0-470-11481-0.

Cannon, Jr., R. H. [1967]. *Dynamics of Physical Systems*. Reprint: [Cannon, Jr. 2003]. McGraw-Hill: New York, NY.

— [2003]. *Dynamics of Physical Systems*. Original: [Cannon, Jr. 1967]. Dover Publications, Inc.: New York, NY. ISBN: 978-0-486-42865-9.

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.

Chen, C.-T. [1984]. *Linear System Theory and Design*. HRW Series in Electrical and Computer Engineering. Holt, Rinehart and Winston: New York/Chicago/San Francisco/-Philadelphia. ISBN: 0-03-060289-0.

Chen, J. [1993]. "Static output feedback stabilization for SISO systems and related problems: Solutions via generalised eigenvalues". *Control Theory and Advanced Technology* 10(4), pages 2233–2244. ISSN: 2095-6983.

Chen, M. J. and Desoer, C. A. [1982]. "Necessary and sufficient condition for robust stability of linear distributed feedback systems". *International Journal of Control* 35(2), pages 255–267. ISSN: 0020-7179. DOI: 10.1080/00207178208922617.

Chen, W. K. [1976]. *Applied Graph Theory*. 2nd edition. North-Holland Series in Applied Mathematics and Mechanics 13. North-Holland: Amsterdam/New York. ISBN: 978-0-444-57003-1.

Cook, S. A. [1970]. "The complexity of theorem-proving techniques". In: *Conference Record of 3rd Annual ACM Symposium on Theory of Computing*. ACM Symposium on Theory of Computing. (Shaker Heights, OH, May 1970). Association for Computing Machinery, pages 151–158.

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.

Davis, J. H. [2002]. *Foundations of Deterministic and Stochastic Control*. Systems & Control: Foundations & Applications. Birkhäuser: Boston/Basel/Stuttgart. ISBN: 978-0-8176-4257-0.

Desoer, C. A. and Chan, W. S. [1975]. "The feedback interconnection of lumped linear time-invariant systems". *Journal of the Franklin Institute. Engineering and Applied Mathematics* 300(5/6), pages 335–351. ISSN: 0016-0032. DOI: 10.1016/0016-0032(75)90161-1.

Dorf, R. C. and Bishop, R. H. [2010]. *Modern Control Systems*. 12th edition. Prentice-Hall: Englewood Cliffs, NJ. ISBN: 978-0-13-602458-3.

Doyle, J. C. [1978]. "Guaranteed margins for LQG regulators". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 23(4), pages 756–757. ISSN: 0018-9286. DOI: 10.1109/TAC.1978.1101812.

— [1982]. "Analysis of feedback systems with structured uncertainty". *Institution of Electrical Engineers. Proceedings. D. Control Theory and Applications* 129(6), pages 242–250. ISSN: 1350-2379. DOI: 10.1049/ip-d:19820053.

Doyle, J. C., Francis, B. A., and Tannenbaum, A. R. [1990]. *Feedback Control Theory*. Reprint: [Doyle, Francis, and Tannenbaum 2009]. Macmillian: New York, NY. ISBN: 0-02-330011-6.

— [2009]. *Feedback Control Theory*. Original: [Doyle, Francis, and Tannenbaum 1990]. Dover Publications, Inc.: New York, NY. ISBN: 978-0486469331.

Doyle, J. C. and Stein, G. [1981]. "Multivariable feedback design: Concepts for a classical modern synthesis". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 26(1), pages 4–16. ISSN: 0018-9286. DOI: 10.1109/TAC.1981.1102555.

Dullerud, G. E. and Paganini, F. [1999]. *A Course in Robust Control Theory. A Convex Approach*. Texts in Applied Mathematics 36. Springer-Verlag: New York/Heidelberg/-Berlin. ISBN: 978-0-387-98945-7.

El Ghaoui, L. and Niculescu, S.-I. [1987]. *Advances in Linear Matrix Inequality Methods in Control*. Advances in Design and Control. Society for Industrial and Applied Mathematics: Philadelphia, PA. ISBN: 978-0-89871-438-8.

Evans, W. R. [1948]. "Graphical analysis of control systems". *Transactions of the American Institute of Electrical Engineers* 67, pages 547–551. ISSN: 0096-3860. DOI: 10.1109/T-AIEE.1948.5059708.

— [1950]. "Control systems synthesis by root locus method". *Transactions of the American Institute of Electrical Engineers* 69, pages 66–69. ISSN: 0096-3860. DOI: 10.1109/T-AIEE.1950.5060121.

Francis, B. A. [1987]. *A Course in H_∞ Control Theory*. Lecture Notes in Control and Information Sciences 88. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-0-387-17069-5.

Francis, B. A. and Zames, G. [1984]. "On H_∞-optimal sensitivity theory for SISO feedback systems". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 29(1), pages 9–16. ISSN: 0018-9286. DOI: 10.1109/TAC.1984.1103357.

Franklin, G. F., Powell, J. D., and Emani-Naeini, A. [2009]. *Feedback Control of Dynamic Systems*. 6th edition. Prentice-Hall: Englewood Cliffs, NJ. ISBN: 978-0-13-601969-5.

Freudenberg, J. S. and Looze, D. P. [1985]. "Right half plane poles and zeros and design tradeoffs in feedback systems". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 30(6), pages 555–565. ISSN: 0018-9286. DOI: 10.1109/TAC.1985.1104004.

Fuhrmann, P. A. [2012]. *A Polynomial Approach to Linear Algebra*. 2nd edition. Universitext. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-1-4614-0337-1.

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. [1959a]. *The Theory of Matrices*. Translated by K. A. Hirsch. Volume 1. Reprint: [Gantmacher 2000a]. Chelsea: New York, NY.

— [1959b]. *The Theory of Matrices*. Translated by K. A. Hirsch. Volume 2. Reprint: [Gantmacher 2000b]. Chelsea: New York, NY.

— [2000a]. *The Theory of Matrices*. Translated by K. A. Hirsch. Volume 1. Original: [Gantmacher 1959a]. American Mathematical Society: Providence, RI. ISBN: 978-0-8218-1393-5.

— [2000b]. *The Theory of Matrices*. Translated by K. A. Hirsch. Volume 2. Original: [Gantmacher 1959b]. American Mathematical Society: Providence, RI. ISBN: 978-0-8218-2664-5.

Geromel, J. C., Souza, C. C. de, and Skelton, R. E. [1998]. "Static output feedback controllers: Stability and convexity". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 43(1), pages 120–125. ISSN: 0018-9286. DOI: 10.1109/9.654912.

Goodwin, G. C., Graebe, S. F., and Salgado, M. E. [2001]. *Control System Design*. Prentice-Hall: Englewood Cliffs, NJ. ISBN: 978-0-13-958653-8.

Hang, C. C., Åstrom, K. J., and Ho, W. K. [1990]. *Refinements of the Ziegler–Nicols Tuning Formula*. Technical Report CI-90-1. Department of Electrical Engineering, National University of Singapore.

Hautus, M. L. J. [1969]. "Controllability and observability conditions of linear autonomous systems". *Koninklijke Nederlandse Akademie van Wetenschappen. Proceedings. Series A. Mathematical Sciences* 31, pages 443–448. ISSN: 0023-3358.

Helton, J. W. and Merrino, O. [1998]. *Classical Control Using H_∞ Methods, Theory, Optimization, and Design*. Society for Industrial and Applied Mathematics: Philadelphia, PA. ISBN: 978-0-89871-419-7.

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.

Horowitz, I. M. [1963]. *Synthesis of Feedback Systems*. Academic Press: New York, NY. ISBN: 978-0-12-355950-0.

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).

Kailath, T. [1980]. *Linear Systems*. Information and System Sciences Series. Prentice-Hall: Englewood Cliffs, NJ. ISBN: 978-0-13-536961-6.

Kalman, R. E. [1960]. "A new approach to linear filtering and prediction theory". *Transactions of the ASME. Series D. Journal of Basic Engineering* 82, pages 35–45. ISSN: 0021-9223.

Kalman, R. E. and Bucy, R. S. [1960]. "New results in linear filtering and prediction theory". *Transactions of the ASME. Series D. Journal of Basic Engineering* 83, pages 95–108. ISSN: 0021-9223.

Kalman, R. E., Ho, Y.-C., and Narendra, K. S. [1963]. "Controllability of linear dynamical systems". *Contributions to Differential Equations* 1, pages 189–213. DOI: 10.1137/0301010.

Kamen, E. W. and Su, J. K. [1999]. *Introduction to Optimal Estimation*. Advanced Textbooks in Control and Signal Processing. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-1-85233-133-7.

Khalil, H. K. [2001]. *Nonlinear Systems*. 3rd edition. Prentice-Hall: Englewood Cliffs, NJ. ISBN: 978-0-13-067389-3.

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.

Krall, A. M. [1961]. "An extension and proof of the root-locus method". *Journal of the Society for Industrial and Applied Mathematics* 9(4), pages 644–653. URL: http://www.jstor.org/stable/2098888 (visited on 07/10/2014).

Kučera, V. and Souza, C. E. de [1995]. "A necessary and sufficient condition for output feedback stabilizability". *Automatica. A Journal of IFAC, the International Federation of Automatic Control* 31(9), pages 1357–1359. ISSN: 0005-1098. DOI: 10.1016/0005-1098(95)00048-2.

Lang, S. [2003]. *Complex Analysis*. 4th edition. Graduate Texts in Mathematics 103. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-0-387-98592-3.

— [2005]. *Algebra*. 3rd edition. Graduate Texts in Mathematics 211. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-0-387-95385-4.

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.

Lynch, W. A. [1961]. "Linear control systems. A signal-flow-graph viewpoint". In: *Adaptive Control Systems*. Edited by E. Mishkin and L. Braun. McGraw-Hill: New York, NY. Chapter 2, pages 23–49.

MacFarlane, A. G. J. [1982]. "Complex variable methods in feedback systems analysis and design". In: *Design of Modern Control Systems*. Edited by D. J. Bell, P. A. Cook, and N. Munro. IEE Control Engineering Series 20. Peter Peregrinus, Ltd.: Stevanage. Chapter 2, pages 18–45. ISBN: 0-906048-74-5.

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.

Marshall, D. E. [1975]. "An elementary proof of the Pick–Nevanlinna interpolation theorem". *The Michigan Mathematical Journal* 21(3), pages 219–223. ISSN: 0026-2285. DOI: 10.1307/mmj/1029001307.

Mason, S. J. [1953a]. "Feedback theory: Further properties of signal flow agraphs". *Proceedings of the IRE* 44(7), pages 920–926. ISSN: 0096-8390. DOI: 10.1109/JRPROC.1956.275147.

— [1953b]. "Feedback theory: Some properties of signal flow graphs". *Proceedings of the IRE* 41(9), pages 1144–1156. ISSN: 0096-8390. DOI: 10.1109/JRPROC.1953.274449.

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).

Middleton, R. H. and Goodwin, G. C. [1990]. *Digital Control and Estimation. A Unified Approach*. Prentice-Hall: Englewood Cliffs, NJ. ISBN: 978-0-13-211665-7.

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.

Morris, K. A. [2000]. *Introduction to Feedback Control*. Harcourt Brace & Company: New York, NY. ISBN: 978-0-12-507660-9.

Mulligan Jr., J. R. [1949]. "The effect of pole and zero location on the transient response of linear dynamics systems". *Institution of Electrical Engineers. Proceedings. D. Control Theory and Applications* 37(5), pages 516–529. ISSN: 1350-2379. DOI: 10.1109/JRPROC.1949.232649.

Nehari, Z. [1957]. "On bounded bilinear forms". *Annals of Mathematics. Second Series* 65(1), pages 153–162. ISSN: 0003-486X. DOI: 10.2307/1969670.

Nevanlinna, R. [1919]. "Über beschränkte Funktionen, die in gegebenen Punkten vorgeschriebene Werte annehmen". *Annales Academae Scientiarium Fenniae. Mathematica* 13(1), pages 1–71. ISSN: 1239-629X.

— [1929]. "Über beschränkte analytisch Funktionen". *Annales Academae Scientiarium Fenniae. Mathematica* 32(7), pages 1–75. ISSN: 1239-629X.

Nyquist, H. [1932]. "Regeneration theory". *Bell Labs Technical Journal* 11, pages 126–147. ISSN: 1538-7305. URL: http://www3.alcatel-lucent.com/bstj/vol11-1932/articles/bstj11-1-126.pdf (visited on 07/10/2014).

Oară, C. and Varga, A. [2000]. "Computation of general inner-outer and spectral factorizations". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 45(12), pages 2307–2325. ISSN: 0018-9286. DOI: 10.1109/9.895566.

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.

Pick, G. [1916]. "Über die Beschränkunger analytischer Funktionen, welche durch vorgegebebe Funkionswerte bewirkt werden". *Mathematische Annalen* 77, pages 7–23. ISSN: 0025-5831. DOI: 10.1007/BF01456817.

Polderman, J. W. and Willems, J. C. [1998]. *Introduction to Mathematical Systems Theory. A Behavioral Approach*. Texts in Applied Mathematics 26. Springer-Verlag: New York/-Heidelberg/Berlin. ISBN: 978-0-387-98266-3.

Reinschke, K. J. [1988]. *Multivariable Control. A Graph-Theoretic Approach*. Lecture Notes in Control and Information Sciences 108. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-0-387-18899-7.

Routh, E. J. [1877]. *A Treatise on the Stability of a Given State of Motion*. Adam's Prize Essay. Cambrige University.

Schiff, J. L. [1999]. *The Laplace Transform. Theory and Applications*. Undergraduate Texts in Mathematics. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-0-387-98698-2.

Schwartz, L. [1950-1951]. *Théorie des distributions*. 2 volumes. Publications de l'Institut de Mathématique de l'Université de Strasbourg. Reprint: [Schwartz 1997]. Hermann: Paris.

— [1997]. *Théorie des distributions*. 2nd edition. Original: [Schwartz 1950-1951]. Hermann: Paris. ISBN: 978-2-7056-5551-8.

Seron, M. M., Braslavsky, J. H., and Goodwin, G. C. [1997]. *Fundamental Limitations in Filtering and Control*. Communications and Control Engineering Series. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-3-540-76126-6.

Smale, S. [1998]. "Mathematical problems for the next century". *The Mathematical Intelligencer* 20(2), pages 7–15. ISSN: 0343-6993. DOI: 10.1007/BF03025291.

Sontag, E. D. [1998]. *Mathematical Control Theory. Deterministic Finite Dimensional Systems*. 2nd edition. Texts in Applied Mathematics 6. Springer-Verlag: New York/Heidelberg/Berlin. ISBN: 978-0-387-98489-6.

Springer, G. [1957]. *Introduction to Riemann Surfaces*. Reprint: [Springer 2000]. Addison Wesley: Reading, MA.

— [2000]. *Introduction to Riemann Surfaces*. 2nd edition. Original: [Springer 1957]. Chelsea: New York, NY. ISBN: 978-0-8218-3156-4.

Syrmos, V. L., Abdallah, C. T., Dorato, P., and Grigoriadis, K. M. [1987]. "Static output feedback—A survey". *Automatica. A Journal of IFAC, the International Federation of Automatic Control* 33(2), pages 125–137. ISSN: 0005-1098. DOI: 10.1016/S0005-1098(96)00141-0.

Truxal, J. G. [1955]. *Automatic Feedback Control System Synthesis*. McGraw-Hill Electrical and Electronic Engineering Series. McGraw-Hill: New York, NY.

van der Woude, J. [1988]. "A note on pole placement by static output feedback for single-input systems". *Systems & Control Letters* 11(4), pages 285–287. ISSN: 0167-6911. DOI: 10.1016/0167-6911(88)90072-2.

Vidyasagar, M. [1981]. *Input-Output Analysis of Large-Scale Interconnected Systems*. Lecture Notes in Control and Information Sciences 29. Springer-Verlag: New York/Heidelberg/-Berlin. ISBN: 978-3-540-10501-5.

— [1986]. "On undershoot and nonminimum phase zeros". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 31(5), page 440. ISSN: 0018-9286. DOI: 10.1109/TAC.1986.1104289.

— [1987]. *Control System Synthesis. A Factorization Approach*. MIT Press Series in Signal Processing, Optimization, and Control 7. MIT Press: Cambridge, MA. ISBN: 978-0-262-72012-0.

Wang, Q.-G., Lee, T.-H., and He, J.-B. [1999]. "Internal stability of interconnected systems". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 44(3), pages 593–596. ISSN: 0018-9286. DOI: 10.1109/9.751358.

Willems, J. C. [1971]. *Analysis of Feedback Systems*. MIT Press: Cambridge, MA. ISBN: 978-0-26-273160-7.

Youla, D. C. [1961]. "On the factorization of rational matrices". *IRE Transactions on Information Theory* 7, pages 172–189. ISSN: 0096-1000.

Youla, D. C., Jabr, H. A., and Bongiorno, J. J. [1976]. "Modern Wiener–Hopf design of optimal controllers. I. The single-input-output case". *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 21(1), pages 3–13. ISSN: 0018-9286. DOI: 10.1109/TAC.1976.1101139.

Zadeh, L. A. and Desoer, C. A. [1963]. *Linear System Theory. The State Space Approach*. McGraw-Hill Series in System Science. McGraw-Hill: New York, NY.

— [1979]. *Linear System Theory. The State Space Approach*. Original: [Zadeh and Desoer 1963]. Robert E. Krieger Publishing Company: Huntington/New York. ISBN: 978-0-88275-809-1.

Zemanian, A. H. [1965]. *Distribution Theory and Transform Analysis. An Introduction to Generalized Functions, with Applications*. Reprint: [Zemanian 1987]. McGraw-Hill: New York, NY.

— [1987]. *Distribution Theory and Transform Analysis. An Introduction to Generalized Functions, with Applications*. 2nd edition. Original: [Zemanian 1965]. Dover Publications, Inc.: New York, NY. ISBN: 978-0-486-65479-9.

Zhou, K., Doyle, J. C., and Glover, K. [1996]. *Robust and Optimal Control*. Prentice-Hall: Englewood Cliffs, NJ. ISBN: 978-0-13-456567-5.

Ziegler, J. G. and Nicols, N. B. [1942]. "Optimum settings for automatic controllers". *Transactions of the American Society for Mechanical Engineers* 64(8), page 759. DOI: 10.1115/1.2899060.

Symbol Index