Optimization and Control of Stochastic Systems

Serdar Yüksel
Queen's University, Mathematics and Statistics

Lecture Notes

April 2, 2026

These notes have been been prepared for *MTHE/MATH 472 / MATH 872: Optimization and Control of Stochastic Systems* at Queen's University and also used for *EEE 446/546: Control and Optimization of Stochastic Systems* at Bilkent University.

Contents

Introduction

In a typical *differential equations* or a *systems* course, one learns about the behaviour of a system described by differential or difference equations. For such systems, under mild regularity conditions, a given initial condition (in the absence of disturbances) leads to a unique solution/output. Even when one cannot obtain an explicit analytical solution, it is often possible to establish stability properties of solutions.

In many engineering or applied mathematics areas, one also has the liberty to affect the flow of the system through a control term. Control theory is concerned with shaping the input-output behaviour of a system by possibly utilizing feedback from system outputs under various design criteria and constraints. The way control actions or variables are generated based on the information available at the controller is called the *control policy* or *control law*. In deterministic control theory, under mild conditions, a given initial state and a given control policy uniquely specifies the realized path. Such a policy may be designed to stabilize a system, under stabilizability conditions; or control a system, under controllability conditions. The deterministic theory has had tremendous impact and success in many applications with commonly considered criteria being system stability (e.g. convergence to a point or a set with respect to initial state conditions, or boundedness of the output corresponding to any bounded input), reference tracking, robustness to incorrect models and disturbances (which may appear in the system itself or in measurements available at the controller), and optimal control.

However, in many applications, the deterministic theory is not directly applicable, as there may be disturbances in a given system. In such systems, disturbances may appear in the dynamics of a system or in the information available to the controller. Some general differences between the deterministic and stochastic setups are the following:

- The criteria on stability for stochastic systems require a different approach since stabilization to a point, or to a compact set or formation, is often too much to ask in a stochastic system.
- The solution concepts for stochastic systems can be significantly different from those in a deterministic setup.
- For optimization problems, the optimality criteria need to be probabilistic in general.
- Informational aspects of control and decision making in the presence of uncertainty lead to significant mathematical complexity but also versatility in applications (including those in decentralized setups where multiple decision makers are present either in a cooperative or in an adversarial context, as in game theory). Such informational dependence of stochastic control is perhaps what particularly distinguishes the stochastic theory from its deterministic counterpart.

Nonetheless, we will see that many concepts and principles from deterministic control theory carry over to the stochastic setup. For a stochastic system, we will see that even though a control policy and an initial condition does not uniquely determine the path that a controlled process may take, the probability measure on the future paths is uniquely specified given a policy. Likewise, the concepts of stability, optimality and observability will all find corresponding interpretations (though with significant generalizations, refinements, but also limitations). Results from geometric control theory and robust control theory will lead to remarkable insights.

However, these connections require a strong foundation on probability (and several other areas of mathematics and engineering): before we proceed with the technical study of the subject, which will also touch on the aforementioned application areas, in the first chapter of these notes a concise but sufficiently detailed review of probability theory will be presented.

Some application areas include: optimal regulation and tracking; optimal filtering of noisy measurements with respect to a hidden dynamical system and control of such systems; operations research; mathematical finance and optimal investment; stochastic and data-driven learning methods for optimization (including reinforcement and stochastic learning theoretic problems and applications); stability and optimization of communication networks (e.g. in optimal routing and scheduling); information theory (in particular for setups involving causality and feedback); robust design of control systems under approximation errors, incorrect models and priors; stability analysis and stabilization of stochastic dynamical systems; decentralized stochastic control of systems; stochastic control in the presence of adverse decision makers (as in stochastic game theory); and stochastic networked control (control under information constraints between various components of a control system).

In the lecture notes, following a review chapter on probability, we will first proceed with stochastic stability, optimization under various criteria, the problems with partial information, and stochastic learning theory. A basic course in stochastic

control could cover the topics mentioned so far. If further time is available, the additional material presented on decentralized stochastic control, stochastic control in continuous-time, and robustness to incorrect models and learning, can be covered.

1

Review of Probability

1.1 Introduction

Before discussing controlled Markov chains, we first discuss some preliminaries about probability theory.

Many events in the physical world are uncertain; that is, with a given prior knowledge (such as an initial condition) regarding a process, the future values of the process are not exactly predictable. Probability theory attempts to develop an understanding for such uncertainty in a consistent way given a number of properties to be satisfied.

Examples of stochastic processes include: a) The temperature in a city at noon throughout some October: This process takes values in $\mathbb{R}^{31}$, b) The sequence of outputs of a communication channel modeled by an additive scalar Gaussian noise when the input sequence is given by $x = \{x_1, \ldots, x_n\} \in \mathbb{R}^n$ (the output process lives in $\mathbb{R}^n$), c) Infinite copies of a discrete-time coin flip process (living in $\{H, T\}^{\mathbb{Z}_+}$, where H denotes the *head* and T denotes the *tail* outcome), d) The trajectory of a plane flying from point A to point B (taking values in $C([t_0, \infty); \mathbb{R}^3)$, the space of all continuous paths in $\mathbb{R}^3$ with $x_{t_0} = A, x_{t_f} = B$ for some $t_0 < t_f \in \mathbb{R}$), e) The exchange rate between the Canadian dollar and the American dollar on a given time index T.

Some of these processes take values in countable spaces, some do not. If the state space $\mathbb{X}$ in which a random variable takes values is finite or countably infinite, it suffices to associate with each point $x \in \mathbb{X}$ a number which determines the likelihood of the *event* that x is the realized value of the variable. However, when $\mathbb{X}$ is uncountable, only focusing on such realizations is not sufficiently descriptive and further technical intricacies arise. Accordingly, the notion of an *event* needs to be carefully defined. First, if some event A takes place, it must be that the complement of A (that is, this event not happening) must also be defined. Furthermore, if A and B are two events, then the intersection must also be an event. This line of thought will motivate us for a more formal analysis below. In particular, one needs to construct probability values by first defining values for certain events and extending such probabilities to a larger class of events in a consistent fashion (in particular, one does not first associate probability values to single points as we do in countable state spaces). These issues are best addressed with a precise characterization of probability and random variables.

Probability theory is a versatile mathematical construction to model and study uncertainty in the real world. In the following, we will present a rigorous, though concise, review of probability. For a more complete exposition the reader could consult with several comprehensive texts on probability theory, such as [43, 57, 111, 130, 316] and texts on stochastic processes, such as [151, 158, 332].

1.2 Measures and Integration

Let $\mathbb{X}$ be a collection of points. Let $\mathcal{F}$ be a collection of subsets of $\mathbb{X}$ with the following properties such that $\mathcal{F}$ is a σ-field (also called a σ-algebra), that is:

- $\mathbb{X} \in \mathcal{F}$

- If $A \in \mathcal{F}$, then $\mathbb{X} \setminus A \in \mathcal{F}$
- If $A_k \in \mathcal{F}, k = 1, 2, 3, \ldots$, then $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$ (that is, the collection is closed under countably many unions).

By De Morgan's laws, and set properties, it can be shown that the collection has to be closed under countable intersections as well.

For example, the full power-set of any set is a σ-field.

If the third item above holds for only finitely many unions or intersections, then the collection of subsets is said to be a *field* or $algebra$ over $\mathbb{X}$.

With the above, $(\mathbb{X}, \mathcal{F})$ is termed a measurable space (that is we can associate a measure to this space; which we will discuss shortly).

Remark 1.1. Subsets in σ-fields can be interpreted to represent *information* that a controller has with regard to an underlying process. We will discuss this interpretation further and this will be a recurring theme in our discussions in the context of stochastic control.

A σ-field $\mathcal{J}$ is generated by a collection of sets $\mathcal{A}$, if $\mathcal{J}$ is the smallest σ-field containing the sets in $\mathcal{A}$, and in this case, we write $\mathcal{J} = \sigma(\mathcal{A})$.

Exercise 1.2.1 *Let $\mathbb{X} = \{a, b, c\}$. (i) Find $\sigma(\{a\})$. (ii) Find $\sigma(\{a\}, \{b\}, \{c\})$.*

We consider an important special case in the following.

1.2.1 Borel σ-field

An important class of σ-fields is the Borel σ-field on a metric (or more generally, topological) space. Such a σ-field is the one which is generated by open sets. The term *open* naturally depends on the space being considered. For this course, we will mainly consider spaces which are complete, separable and metric spaces (such as the space of real numbers $\mathbb{R}$, or countable sets) see Appendix A). Recall that in a metric space with metric d, a set U is open if for every $x \in U$, there exists some $\epsilon > 0$ such that $\{y : d(x, y) < \epsilon\} \subset U$. We note also that the empty set is a special open set.

The Borel σ-field on $\mathbb{R}$ is then the one generated by sets of the form $(a, b) \subset \mathbb{R}$, that is, open intervals (it is important to note here that every open set in $\mathbb{R}$ can be expressed a union of countably many open intervals). It is also important to note that not all subsets of $\mathbb{R}$ are Borel sets, that is, elements of the Borel σ-field; see e.g. Exercise 1.6.7.

We will denote the Borel σ-field on a space $\mathbb{X}$ as $\mathcal{B}(\mathbb{X})$.

Exercise 1.2.2 *Show that for every $a \in \mathbb{R}$, $\{a\} \in \mathcal{B}(\mathbb{R})$, by writing $a = \cap_{n \in \mathbb{N}}(a - \frac{1}{n}, a + \frac{1}{n})$.*

We can also define a Borel σ-field on a product space. Let $\mathbb{X}$ be a complete, separable, metric space (with metric d). Let $\mathbb{X}^{\mathbb{Z}_+}$ denote the infinite product of $\mathbb{X}$ so that $x = (x_0, x_1, x_2, \cdots) \in \mathbb{X}^{\mathbb{Z}_+}$, where $x_k \in \mathbb{X}$ for $k \in \mathbb{Z}_+$. If this space is endowed with the product metric (such a metric is defined as: $\rho(x, y) = \sum_{i=0}^{\infty} 2^{-i} \frac{d(x_i, y_i)}{1 + d(x_i, y_i)}, \quad x, y \in \mathbb{X}^{\mathbb{Z}_+}$), sets of the form $\prod_{i \in \mathbb{Z}_+} A_i$, where only finitely many of these sets are not equal to $\mathbb{X}$ and these sets are open; and unions of such sets form open sets. We define cylinder sets in this product space as:

$$B_{[A_m, m \in I]} = \{x \in \mathbb{X}^{\mathbb{Z}_+}, x_m \in A_m, m \in I\},$$

with $A_m \in \mathcal{B}(\mathbb{X})$ and where $I \subset \mathbb{Z}$ with $|I| < \infty$, that is, the set I has finitely many elements. Thus, in the above, if $x \in B_{[A_m, m \in I]}$, then, $x_m \in A_m$ for $m \in I$ and the remaining terms (that is, the x_m values for $m \notin I$) can be taken arbitrarily from $\mathbb{X}$. We can thus view a cylinder set as a pre-image of the projection operation onto finitely many coordinates. The σ-field generated by such open cylinder sets is the Borel σ-field on the product space. Such a construction

is important for stochastic processes (and is the reason why while studying certain properties of stochastic processes one often only considers finite dimensional distributions).

Remark 1.2. A space which admits a metric under which it is complete and separable is called a Polish space; while the metric is often not apriori specified for such a space, we will often assume that a metric is given and define a Polish metric space to be a complete separable and metric space. A Borel subset of a Polish space is called a *standard Borel* space [293]. A very important fact is that any Polish space is related to either a finite set, or a countably infinite set, or $\mathbb{R}$, through a bijection (that is, via a measurable function -to be defined shortly- with a measurable inverse).

1.2.2 Measurable Function

If $(\mathbb{X}, \mathcal{F})$ and $(\mathbb{Y}, \mathcal{G})$ are measurable spaces; we say a mapping from $h : (\mathbb{X}, \mathcal{F}) \to (\mathbb{Y}, \mathcal{G})$ is a measurable function if

$$h^{-1}(B) := \{x : h(x) \in B\} \in \mathcal{F} \qquad \forall B \in \mathcal{G}$$

In the particular case involving Borel σ-fields, if $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ are measurable spaces; we say a mapping from $h : \mathbb{X} \to \mathbb{Y}$ is (Borel) measurable if

$$h^{-1}(B) = \{x : h(x) \in B\} \in \mathcal{B}(\mathbb{X}), \quad \forall B \in \mathcal{B}(\mathbb{Y})$$

Theorem 1.2.1 *To show that a function is measurable, it is sufficient to check the measurability of the inverses of sets that generate the σ-algebra on the image space.*

See Section 1.5.1 for a proof. Therefore, for Borel measurability, it suffices to check the measurability of the inverse images of open sets. Furthermore, for real valued functions, to check the measurability of the inverse images of open sets, it suffices to check the measurability of the inverse images sets of the form $\{(-\infty, a], a \in \mathbb{R}\}$, $\{(-\infty, a), a \in \mathbb{R}\}$, $\{(a, \infty), a \in \mathbb{R}\}$ or $\{[a, \infty), a \in \mathbb{R}\}$, since each of these generate the Borel σ-field on $\mathbb{R}$. In fact, here we can restrict a to be $\mathbb{Q}$-valued, where $\mathbb{Q}$ is the set of rational numbers (since $\{x : x < r\} = \cup_{q \in \mathbb{Q}, q<r}\{x : x < q\}$; often this reasoning is why we call such sigma-fields *countably generated*).

It is instructive to view measurability in terms of informativeness of a σ-field. Let $\mathbb{X} = \{a, b, c\}$ and let $\mathcal{F}_1$ be as in Exercise 1.2.1(i) and $\mathcal{F}_2$ be as in Exercise 1.2.1(ii). Let $\mathbb{Y} = \{0, 1\}$ and $\mathcal{G} = \sigma(\{0\}, \{1\})$. Now, let $F : \mathbb{X} \to \mathbb{Y}$ be a map so that $F^{-1}(0) = \{a, b\}$ and $F^{-1}(1) = \{c\}$. Then, we can conclude that this map defines a measurable function from $(\mathbb{X}, \mathcal{F}_2) \to (\mathbb{Y}, \mathcal{G})$ but it is not a measurable function from $(\mathbb{X}, \mathcal{F}_1) \to (\mathbb{Y}, \mathcal{G})$: the reason is that the information on whether $F(x) = 1$ (that is $x = c$) is not an element of $\mathcal{F}_1$ (and thus, this information that $x = c$ or not, is not available as information under $\mathcal{F}_1$).

1.2.3 Measure

Let $(\mathbb{X}, \mathcal{F})$ be a measurable space. A positive measure μ on $(\mathbb{X}, \mathcal{F})$ is a map from $\mathcal{F}$ to $[0, \infty]$ which is *countably additive* such that for $A_k \in \mathcal{F}$ and $A_k \cap A_j = \emptyset$:

$$\mu\left(\cup_{k=1}^{\infty} A_k \right) = \sum_{k=1}^{\infty} \mu(A_k).$$

Definition 1.2.1 *μ is a probability measure if it is positive and $\mu(\mathbb{X}) = 1$.*

Definition 1.2.2 *A measure μ is finite if $\mu(\mathbb{X}) < \infty$, and σ-finite if there exist a collection of subsets $A_k \in \mathcal{F}$ such that $X = \cup_{k=1}^{\infty} A_k$ with $\mu(A_k) < \infty$ for all k.*

On the real line $\mathbb{R}$, the Lebesgue measure λ is defined on the Borel σ-field (in fact on a somewhat larger field obtained through adding all subsets of Borel sets of measure zero: this is known as *completion* of a σ-field) such that for $A = (a, b)$, $\lambda(A) = b-a$. Borel σ-field of subsets is a strict subset of *Lebesgue measurable* sets, that is there exist Lebesgue measurable sets which are not Borel sets. For a definition and the construction of Lebesgue measurable sets, see [43]. Countable subsets of $\mathbb{R}$ all have zero Lebesgue measure but there also exist Lebesgue measurable sets of measure zero which contain uncountably many elements, for a well-studied example see the Cantor set [111]. Not all subsets of $\mathbb{R}$ are Lebesgue measurable (and thus, not Borel either), see e.g. Exercise 1.6.7.

1.2.4 The Extension Theorem (Optional)

Theorem 1.2.2 *[The Extension Theorem (Carathéodory)] Let $\mathcal{M}$ be an algebra over $\mathbb{X}$, and suppose that there exists a map (called a pre-measure) $P : \mathcal{M} \to \mathbb{R}_+$ so that for any (possibly countably infinitely many) pairwise disjoint sets $A_n \in \mathcal{M}$, if the countable union $\cup_n A_n \in \mathcal{M}$, then $P(\cup_n A_n) = \sum_n P(A_n)$. Suppose also that there exists a countable collection of sets B_n with $\mathbb{X} = \cup_n B_n$, each with $P(B_n) < \infty$ (that is P is σ-finite). Then, there exists a unique measure P' on the σ-field generated by $\mathcal{M}$, $\sigma(\mathcal{M})$, which is consistent with P on $\mathcal{M}$.*

The above is useful since, when one states that two measures are equal it suffices to check whether they are equal on the algebra of sets which generate the σ-field, and not necessarily on the entire σ-field. More importantly, a refinement of the above can be used to define or construct a measure on a σ-field, such as the Lebesgue measure on $\mathcal{B}(\mathbb{R})$.

The following is a refinement useful for stochastic processes. It, in particular, does not require a pre-measure defined apriori before an extension [4]:

Theorem 1.2.3 *[Kolmogorov's Extension Theorem] Let $\mathbb{X}$ be a complete and separable metric space, and for all $n \in \mathbb{N}$ let μ_n be a sequence of probability measures on $\mathbb{X}^n$, the n product of $\mathbb{X}$, such that*

$$\mu_n(A_1 \times A_2 \times \cdots \times A_n) = \mu_{n+1}(A_1 \times A_2 \times \cdots \times A_n \times \mathbb{X}),$$

every sequence of Borel sets $A_k \subset \mathbb{X}$. Then, there exists a unique probability measure μ on $(\mathbb{X}^{\mathbb{N}}, \mathcal{B}(\mathbb{X}^{\mathbb{N}}))$ which is consistent with each of the μ_n's.

A further related result, which often in stochastic control is cited in the context of extensions, is the Ionescu-Tulcea Extension Theorem [165, Appendix C]; where conditional probability measures (stochastic kernels) are defined (instead of probability measures on finite dimensional product spaces) as a starting assumption, before an extension to the infinite product space is established.

Thus, if the σ-field on a product space is generated by the collection of finite dimensional cylinder sets, one can define a measure in the product space which is consistent with the finite dimensional distributions.

Likewise, we can construct the Lebesgue measure on $\mathcal{B}(\mathbb{R})$ by defining it on finitely many unions and intersections of intervals of the form (a, b), $[a, b)$, $(a, b]$ and $[a, b]$, and the empty set, thus forming an algebra (or a field), and extending this to the Borel σ-field. Thus, the relation $\mu(a, b) = b - a$ for $b > a$ is sufficient to define the Lebesgue measure.

Remark 1.3. A related general result is as follows: Let $\mathcal{S}$ be a σ-field. A class of subsets $\mathcal{A} \subset \mathcal{S}$ is called a separating class if two probability measures that agree on $\mathcal{A}$ agree on the entire $\mathcal{S}$. A class of subsets is a π-system if it is closed under finite intersections. The class $\mathcal{A}$ is a separating class if it is both a π-system and it generates the σ-field $\mathcal{S}$; see [42] or [43].

1.2.5 Integration

Let h be a non-negative measurable function from $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The Lebesgue integral of h with respect to a measure μ can be defined in three steps:

First, for $A \in \mathcal{B}(\mathbb{X})$, define $1_{\{x\in A\}}$ (or $1_{(x\in A)}$, or $1_A(x)$) as an indicator function for event $x \in A$, that is the value that the function takes is 1 if $x \in A$, and 0 otherwise. In this case, define

$$\int_{\mathbb{X}} 1_{\{x\in A\}}\mu(dx) := \mu(A).$$

Now, let us define simple functions h such that, there exist $A_1, A_2, \ldots, A_n$ all in $\mathcal{B}(\mathbb{X})$ and positive numbers $b_1, b_2, \ldots, b_n$ such that $h_n(x) = \sum_{k=1}^n b_k 1_{\{x\in A_k\}}$. For such functions, define

$$\int_{\mathbb{X}} h_n(x)\mu(dx) := \sum_{k=1}^n b_k\mu(A_k).$$

Now, for any given measurable h, there exists a sequence of simple functions h_n such that $h_n(x) \uparrow h(x)$ monotonically, that is $h_{n+1}(x) \geq h_n(x)$ (for a construction, if h only takes non-negative values, consider partitioning the positive real line to two intervals $[0, n)$ and $[n, \infty)$, and partition $[0, n)$ to $n2^n$ uniform intervals, define $h_n(x)$ to be the lower floor of the interval that contains $h(x)$: thus

$$h_n(x) = k2^{-n}, \quad \text{if} \quad k2^{-n} \leq h(x) < (k+1)2^{-n}, \quad k = 0, 1, \cdots, n2^n - 1,$$

and $h_n(x) = n$ for $h(x) \geq n$. By definition, and since $h^{-1}([k2^{-n}, (k+1)2^{-n}))$ is Borel, h_n is a simple function. If the function takes also negative values, write $h(x) = h_+(x) - h_-(x)$, where h_+ is the non-negative part and $-h_-$ is the negative part, and construct the same for $h_-(x)$. We *define* the limit (which exists as a real valued monotonically increasing sequence) as the Lebesgue integral:

$$\lim_{n\to\infty} \int h_n(x)\mu(dx) =: \int h(x)\mu(dx)$$

We note that the notation $\int hd\mu$ or $\int h(x)d\mu(x)$ can also be used in place of $\int h(x)\mu(dx)$.

1.2.6 Fatou's Lemma, the Monotone Convergence Theorem and the Dominated Convergence Theorem

Theorem 1.2.4 (Monotone Convergence Theorem) *If μ is a σ-finite positive measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $\{f_n, n \in \mathbb{Z}_+\}$ is a sequence of measurable functions from $\mathbb{X}$ to $\mathbb{R}$ which pointwise, monotonically, converges to f so that $0 \leq f_n(x) \leq f_{n+1}(x)$ for all n, and*

$$\lim_{n\to\infty} f_n(x) = f(x),$$

for $\mu-$almost every x, then

$$\int_{\mathbb{X}} f(x)\mu(dx) = \lim_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(dx)$$

The following is a consequence of the monotone convergence theorem, but is a critical result which will be utilized later in the upcoming chapters.

Theorem 1.2.5 (Fatou's Lemma) *If μ is a σ-finite positive measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $\{f_n, n \in \mathbb{Z}_+\}$ is a sequence of measurable functions, bounded from below, from $\mathbb{X}$ to $\mathbb{R}$, then*

$$\int_{\mathbb{X}} \liminf_{n\to\infty} f_n(x)\mu(dx) \leq \liminf_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(dx)$$

Theorem 1.2.6 (Dominated Convergence Theorem) *If (i) μ is a σ-finite positive measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, (ii) g is a Borel measurable function with*

$$\int_{\mathbb{X}} g(x)\mu(dx) < \infty,$$

and (iii) $\{f_n, n \in \mathbb{Z}_+\}$ *is a sequence of measurable functions from* $\mathbb{X}$ *to* $\mathbb{R}$ *which satisfy* $|f_n(x)| \leq g(x)$ *for* $\mu-$*almost every* x*, and* $\lim_{n\to\infty} f_n(x) = f(x)$*, then:*

$$\int_{\mathbb{X}} f(x)\mu(dx) = \lim_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(dx)$$

Note that for the monotone convergence theorem, there is no restriction on boundedness; whereas for the dominated convergence theorem, there is a boundedness condition. On the other hand, for the dominated convergence theorem, the pointwise convergence does not have to be monotone.

There also exist generalized versions of these theorems, where the measures themselves are time-varying, but converge to a limit measure is in some appropriate sense; see in particular Theorem D.3.1 (building on [212, 287]). These will be discussed later in further detail (and will be seen to be particularly important for stochastic control applications, and in particular on robust stochastic control and learning theory).

1.3 Probability Space and Random Variables

Let $(\Omega, \mathcal{F})$ be a measurable space. If P is a probability measure, then the triple $(\Omega, \mathcal{F}, P)$ is called *a probability space*. Here Ω is a set called the sample space. $\mathcal{F}$ is called the event space, and this is a σ-field of subsets of Ω.

Let $(E, \mathcal{E})$ be another measurable space and $X : (\Omega, \mathcal{F}, P) \to (E, \mathcal{E})$ be a measurable map. We call X an $E-$valued *random variable*. The image under X defines a probability measure on $(E, \mathcal{E})$, called the *law* of X.

The σ-field generated by the events $\{\{w : X(w) \in A\}, A \in \mathcal{E}\}$, that is $\{X^{-1}(A), A \in \mathcal{E}\}$, is called the σ-field generated by X and is denoted by $\sigma(X)$.

Consider a coin flip process, with possible outcomes $\{H, T\}$, heads or tails. We have a good intuitive understanding on the environment when someone tells us that *a coin flip leads to the value* H *with probability* $\frac{1}{2}$. Based on the definition of a random variable, we view then a coin flip outcome as a *deterministic* function from some space $(\Omega, \mathcal{F}, P)$ to the binary output space consisting of a head and a tail event. Here, P denotes the uncertainty measure (you may think of the initial condition of the coin when it is being flipped, the flow dynamics in the air, the conditions on the surface where the coin touches etc.; we encode all these aspects and all the uncertainty in the universe with the abstract space $(\Omega, \mathcal{F}, P)$). You can view then the σ-field generated by such a coin flip as a partition of Ω: if certain things take place the outcome is a H and otherwise it is a T and the outcomes give us information on (the state of) the universe.

A useful fact about measurable functions (and thus random variables) is the following result.

Theorem 1.3.1 *Let* f_n *be a sequence of measurable functions from* $(\Omega, \mathcal{F})$ *to a complete separable metric space* $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$*. Then,* $\limsup_{n\to\infty} f_n(x)$*,* $\liminf_{n\to\infty} f_n(x)$ *are measurable. In particular, if* $f(x) = \lim_{n\to\infty} f_n(x)$ *exists, then* f *is measurable.*

Similar to Theorem 1.2.1, this theorem implies that to verify whether a real valued mapping f is a Borel measurable function, it suffices to check if $f^{-1}(a, b) \in \mathcal{B}(\mathbb{R})$ for $a < b$ since one can construct a sequence of simple functions which will converge to any measurable f, as discussed earlier. It suffices then to check if $f^{-1}(-\infty, a) \in \mathcal{B}(\mathbb{R})$ for $a \in \mathbb{R}$.

1.3.1 More on Random Variables and Probability Density Functions

Consider a probability space $(\mathbb{X}, \mathcal{B}(\mathbb{X}), P)$ and consider an $\mathbb{R}$-valued random variable U measurable with respect to $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$.

This random variable induces a probability measure μ on $\mathcal{B}(\mathbb{R})$ such that for some $(a, b] \in \mathcal{B}(\mathbb{R})$:

$$\mu((a, b]) = P(U \in (a, b]) = P\Big(\{x : U(x) \in (a, b]\}\Big) = P(U^{-1}((a, b])$$

When U is $\mathbb{R}-$valued, the *expectation* of U is given with

$$E[U] = \int_{\mathbb{R}} \mu(dx)x,$$

whenever this is defined (i.e., $E[|U|] < \infty$, in which case we say that U is integrable). We define $F(x) = \mu(-\infty, x]$ as the *cumulative distribution function* of U. If $F(x) = \int_{-\infty}^{x} p(s)\lambda(ds)$ for some p, p is called the *probability density function* (with respect to the Lebesgue measure) of μ. If such a density function exists, we can then write

$$E[U] = \int_{\mathbb{R}} p(x)xdx$$

If a probability density function p exists, the measure μ is said to be *absolutely continuous with respect to the Lebesgue measure*. In particular, the density function p is the *Radon-Nikodym derivative* of μ with respect to the Lebesgue measure λ in the sense that for all Borel A: $\int_A p(x)\lambda(dx) = \mu(A)$. A probability density function does not always exist. In particular, whenever there is a probability *mass* on a given point, then a probability density function does not exist; hence in $\mathbb{R}$, if for some x, $\mu(\{x\}) > 0$, then we say there is a probability mass at x, and a density function does not exist.

However, one can also consider density functions with respect to more general positive measures (that is, different from the Lebesgue measure), we will consider such conditions later in the notes. If $\mathbb{X}$ is countable, we can write $P(\{x = m\}) = p(m)$, where p is called the *probability mass function*; this can be viewed as a density with respect to the (discrete) counting measure.

Some examples of commonly encountered random variables, with their probability density or mass functions, are as follows:

- Gaussian (with mean μ and variance σ^2: $\mathcal{N}(\mu, \sigma^2)$):

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad x \in \mathbb{R}$$

- Exponential (with parameter λ):

$$F(x) = 1 - e^{-\lambda x}, \quad p(x) = \lambda e^{-\lambda x} \qquad x \in \mathbb{R}_+$$

- Uniform on $[a, b]$ ($U([a, b])$):

$$F(x) = \frac{x-a}{b-a}, \quad p(x) = \frac{1}{b-a} \qquad x \in [a, b]$$

- Poisson with rate $\lambda > 0$ on $\mathbb{Z}_+$

$$p(m) = \frac{\lambda^m e^{-\lambda}}{m!}, \qquad m \in \mathbb{Z}_+$$

- Binomial ($B(n, p)$):

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k \in \{0, 1, \cdots, n\}$$

  If $n = 1$, we also call a binomial variable a *Bernoulli* variable.

1.3.2 Independence and Conditional Probability

Consider $A, B \in \mathcal{B}(X)$ such that $P(B) > 0$. The quantity

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

is called the conditional probability of event A given B. The measure $P(\cdot|B)$ defined on $\mathcal{B}(X)$ is itself a probability measure. If

$$P(A|B) = P(A),$$

A and B are said to be independent events. A countable collection of events $\{A_n\}$ is independent if for any finitely many sub-collections $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$, we have that

$$P(A_{i_1}, A_{i_2}, \ldots, A_{i_m}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_m}).$$

Here, we use the notation $P(A, B) = P(A \cap B)$. A sequence of events is said to be pairwise independent if for any two pairs (A_m, A_n): $P(A_m, A_n) = P(A_m)P(A_n)$. Pairwise independence is a weaker concept than independence, that is there exist examples where a collection of random variables is pairwise independent but not independent.

Conditional probability and expectation will be discussed in more detail later in *Chapter* 4.

1.4 Stochastic Processes and Markov Chains

One can define a sequence of random variables as a single random variable living in a product space; that is, we can consider $\{x_1, x_2, \cdots, x_N, \cdots\}$ as an individual random variable X which is an $\mathbb{X}^{\mathbb{Z}_+}$-valued random variable, where now the events are to be defined on the product space.

Let $\mathbb{X}$ be a complete, separable, metric space and let $T = \mathbb{Z}$ or $T = \mathbb{Z}_+$. Let $\mathcal{B}(\mathbb{X})$ denote the Borel sigma-field over $\mathbb{X}$. Let $\Sigma = \mathbb{X}^T$ denote the sequence space of all one-sided (with $T = \mathbb{Z}_+$) or two-sided (with $T = \mathbb{Z}$) infinitely many random variables drawn from $\mathbb{X}$. Thus, if $T = \mathbb{Z}$, $x \in \Sigma$ then $x = \{\ldots, x_{-1}, x_0, x_1, \ldots\}$ with $x_i \in \mathbb{X}, \quad i \in T$. Let $X_n : \Sigma \to \mathbb{X}$ denote the coordinate function such that $X_n(x) = x_n$. Let $\mathcal{B}(\Sigma)$ denote the smallest sigma-field containing all cylinder sets of the form $\{x : X_i(x) = x_i \in B_i, m \leq i \leq n\}$ where $B_i \in \mathcal{B}(\mathbb{X})$, for all integers m, n. We can define a probability measure by a characterization on these finite dimensional cylinder sets, by (the extension) Theorem 1.2.3.

A similar characterization also applies for continuous-time stochastic processes, where T is uncountable. The extension requires more delicate arguments, since finite-dimensional characterizations are too weak to uniquely define a sigma-field on a space of continuous-time paths which is consistent with such distributions. Such technicalities arise in the discussion for continuous-time Markov chains and controlled processes, typically requiring a construction where realizations take values from a separable product space (such as the space of continuous sample paths; in this case, the sample path values on a countably dense subset uniquely determine the entire sample path and hence the discrete-time theory, essentially, is applicable); see Section 11.1.1 for further discussion.

In much of these notes, our focus will primarily be on discrete-time processes; however, we will note later that the analysis for continuous-time processes essentially follows from similar constructions with further structures that one needs to impose on continuous-time processes (such as some continuity properties of the sample paths). Further discussion on this is presented in *Chapter* 11.

1.4.1 Markov Chains

If the probability measure on an $\mathbb{X}^{\mathbb{Z}_+}$-valued sequence is such that for every $k \in \mathbb{N}$, for every Borel A_{k+1} and (P-almost surely) all realizations $x_{[0,k]}$,

$$P(x_{k+1} \in A_{k+1}|x_k, x_{k-1}, \ldots, x_0) = P_k(x_{k+1} \in A_{k+1}|x_k),$$

for some conditional probability measure P_k, then $\{x_k\}$ is said to be a *Markov chain*. If P_k is a constant and does not depend on k, the chain is said to be a time-homogeneous chain, otherwise it is time-inhomogeneous. Thus, for a Markov chain, the immediate state is sufficient to predict the future (and past variables are not needed).

One way to construct a Markov chain is via the following: Let $\{x_t, t \geq 0\}$ be a random sequence with state space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, and defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, where $\mathcal{B}(\mathbb{X})$ denotes the Borel σ-field on $\mathbb{X}$, Ω is the sample space, $\mathcal{F}$ a sigma field of subsets of Ω, and $\mathbf{P}$ a probability measure. For $x \in \mathbb{X}$ and $D \in \mathcal{B}(\mathbb{X})$, we let $P(x, D) := \mathbf{P}(x_{t+1} \in D | x_t = x)$ denote the transition probability from x to D, that is the probability of the event $\{x_{t+1} \in D\}$ given that $x_t = x$. Thus, the Markov chain is completely determined by the transition probability and the probability of the initial state, $P(x_0 \in \cdot)$. The probability of the event $\{x_{t+1} \in D\}$ for any t can be computed recursively by starting at $t = 0$, with $\mathbf{P}(x_1 \in D) = \int P(x_1 \in D | x_0 = x) \mathbf{P}(x_0 \in dx)$, and iterating with a similar formula for $t = 1, 2, \ldots$ (building on the Ionescu-Tulcea Extension Theorem [165, Appendix C], which was discussed earlier).

Hence, if the probability of the same event given some history of the past and the present does not depend on the past, and hence is given by the same quantity regardless of the past realizations as long as the present realization is fixed (almost surely), the chain is a Markov chain. As an example, consider the following linear system:

$$x_{t+1} = a x_t + w_t,$$

where $\{w_t\}$ is an independent sequence of random variables for some $a \in \mathbb{R}$. The process $\{x_t\}$ is Markov. We also note that every time-homogeneous Markov chain admits a stochastic, functional and sample-path, realization of the form $x_{k+1} = f(x_k, w_k)$ where f is measurable and w_k is an i.i.d. $[0, 1]$-valued process (see [144, Lemma 1.2], [56, Lemma 3.1], or [21, Lemma F]). This realization result will be useful later on.

We will continue our discussion on Markov chains after discussing controlled Markov chains in the following chapter.

1.5 Appendix

1.5.1 Proof of Theorem 1.2.1

Observe that set operations satisfy that for any $B \in \mathcal{B}(\mathbb{Y})$: $h^{-1}(\mathbb{Y} \setminus B) = \mathbb{X} \setminus h^{-1}(B)$ and

$$h^{-1}(\cup_{i=1}^{\infty} B_i) = \cup_{i=1}^{\infty} h^{-1}(B_i), \quad h^{-1}(\cap_{i=1}^{\infty} B_i) = \cap_{i=1}^{\infty} h^{-1}(B_i).$$

Define the set of all subsets of $\mathbb{Y}$ whose inverses are Borel

$$\mathcal{M} := \{B \subset \mathbb{Y} : h^{-1}(B) \in \mathcal{B}(\mathbb{X})\}.$$

Note that $\mathbb{Y} \subset \mathcal{M}$ and by the discussion above, this set is closed under countably many unions. Thus, this $\mathcal{M}$ is a σ-algebra over $\mathbb{Y}$. Note also that this set contains open sets in $\mathbb{Y}$, by the fact that h is measurable, and since this set contains open sets (and that $\mathcal{B}(\mathbb{Y})$ is the smallest σ-algebra containing open sets), it must be that $\mathcal{B}(\mathbb{Y}) \subset \mathcal{M}$. $\diamond$

1.6 Exercises

Exercise 1.6.1 *a) Let H be some set and for all $\beta \in H$, $\mathcal{F}_\beta$ be a σ-field of subsets over some set $\mathbb{X}$. Let*

$$\mathcal{F} = \bigcap_{\beta \in H} \mathcal{F}_\beta$$

Show that $\mathcal{F}$ is also a σ-field on $\mathbb{X}$.

For a space $\mathbb{X}$, on which a metric is defined, the Borel σ-field is generated by the collection of open sets. This means that, the Borel σ-field is the smallest σ-field containing open sets, and as such it is the intersection of all σ-fields containing open sets.

b) Show that any open set in $\mathbb{R}$ under the usual distance $d(x, y) = |x - y|$, can be written as a countable union of intervals. A consequence of this result is that, on $\mathbb{R}$, the Borel σ-field is the smallest σ-field containing open intervals.

c) Is the set of rational numbers an element of the Borel σ-field on $\mathbb{R}$? Is the set of irrational numbers an element?

d) Let $\mathbb{X}$ be a countable set. On this set, let us define a metric as follows:

$$d(x,y) = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{if } x \neq y \end{cases}$$

Show that, the Borel σ-field on $\mathbb{X}$ is generated by the collection of singletons $\{\{x\}, x \in \mathbb{X}\}$; this is the power set, that is, the set of all subsets of $\mathbb{X}$.

e) Let $\mathbb{X} = \mathbb{R}$ and consider the metric d defined as above in part d). Is the σ-field generated by open sets according to this metric the same as the Borel σ-field on $\mathbb{R}$ (under the usual distance metric on $\mathbb{R}$)? Finally, consider the σ-algebra generated by individual points (singletons, that is $\{\{x\}, x \in \mathbb{X}\}$); is this the same as the Borel σ-field or is this the same as the power set on $\mathbb{R}$?

Exercise 1.6.2 *A Borel subset of a complete, separable and metric (i.e., a Polish) space is called a standard Borel space. If $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ are standard Borel spaces; we say a mapping from $h : \mathbb{X} \to \mathbb{Y}$ is (Borel) measurable if*

$$h^{-1}(B) = \{x : h(x) \in B\} \in \mathcal{B}(\mathbb{X}), \quad \forall B \in \mathcal{B}(\mathbb{Y})$$

Prove the following statement: To show that a function h is Borel measurable, it is sufficient to check the measurability of the inverses (under h) of open sets in $\mathbb{Y}$.

Exercise 1.6.3 *Investigate the following limits in view of the convergence theorems.*

a) Check if $\lim_{n\to\infty} \int_0^1 x^n dx = \int_0^1 \lim_{n\to\infty} x^n dx$.

b) Check if $\lim_{n\to\infty} \int_0^1 nx^n dx = \int_0^1 \lim_{n\to\infty} nx^n dx$.

c) Define $f_n(x) = n1_{\{0 \leq x \leq \frac{1}{n}\}}$. Find $\lim_{n\to\infty} \int f_n(x)dx$ and $\int \lim_{n\to\infty} f_n(x)dx$. Are these equal?

Exercise 1.6.4 *a) Let X and Y be real-valued random variables defined on a given probability space. Show that X^2 and $X + Y$ are also random variables.*

b) Let $\mathcal{F}$ be a σ-field of subsets over a set $\mathbb{X}$ and let $A \in \mathcal{F}$. Prove that $\{A \cap B, B \in \mathcal{F}\}$ is a σ-field over A (that is a σ-field of subsets of A).

Hint for part a: The following equivalence holds: $\{X + Y < x\} \equiv \cup_{r\in\mathbb{Q}}\{X < r, Y < x - r\}$. To check if $X + Y$ is a random variable, it suffices to check if the event $\{X + Y < x\} = \{\omega : X(\omega) + Y(\omega) < x\}$ is an element of $\mathcal{F}$ for every $x \in \mathbb{R}$.

Exercise 1.6.5 *Let f_n be a sequence of measurable functions from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Show $f(\omega) = \limsup_{n\to\infty} f_n(\omega)$ and $g(\omega) = \liminf_{n\to\infty} f_n(\omega)$ define measurable functions.*

Exercise 1.6.6 *Let X and Y be real-valued random variables defined on a given probability space $(\Omega, \mathcal{F}, P)$. Suppose that X is measurable on $\sigma(Y)$. Show that there exists a function f such that $X = f(Y)$.*

This result also holds if X and Y are standard Borel valued random variables.

Exercise 1.6.7 *Consider the interval $[0, 1]$. We have seen that the Lebesgue measure λ satisfies $\lambda([a, b)) = U([a, b]) = b - a$ for $0 \leq a \leq b \leq 1$. Consider now the following question: does every subset $S \subset [0, 1]$ admit a Lebesgue measure? In the following we will provide a counterexample, known as the Vitali set.*

Let us define an equivalence class among points in $[0, 1]$ such that $x \sim y$ if $x - y \in \mathbb{Q}$. This equivalence definition partitions $[0, 1]$ into disjoint sets. Note that there are countably many points in each equivalent class.

Let A be a subset which picks exactly one element from each equivalent class (here, we adopt what is known as the Axiom of Choice [43]). Since A contains an element from each equivalence class, each point of $[0,1]$ is contained in the union $\cup_{q\in\mathbb{Q}}(A+q)$. Furthermore, since A contains only one point from each equivalence class, the sets $A+q$, for different q, are disjoint, for otherwise there would be two sets which could include a common point: $A+q$ and $A+q'$ would include a common point, leading to the result that the difference $x-q=z$ and $x-q'=z$ are both in A, a contradiction, since there should be at most one point which is in the same equivalence class as $x-q=z$. The Lebesgue measure is shift-invariant, therefore $\lambda(A)=\lambda(A+q)$. Observe that $[0,1]\subset\cup_{q\in\mathbb{Q}\cup[-1,1]}\{A+q\}\subset[-1,2]$. Since a countable sum of identical non-negative elements can either become ∞ or 0, the contradiction follows: We can't associate a number to this set and as a result, this set is not a Lebesgue measurable set (and also not a Borel set).

2

Controlled Markov Chains

In the following, we discuss controlled Markov models under a variety of informational and dynamical setups.

2.1 Controlled Markov Models

Consider the following model.

$$x_{t+1} = f(x_t, u_t, w_t), \tag{2.1}$$

where x_t is an $\mathbb{X}$-valued state variable, u_t a $\mathbb{U}$-valued control action variable, w_t a $\mathbb{W}$-valued an i.i.d noise process, and f a measurable function. We assume that $\mathbb{X}, \mathbb{U}, \mathbb{W}$ are Borel subsets of complete, separable, metric spaces (such complete, separable and metric spaces are called *Polish metric* spaces); such subsets of these spaces are also called *standard Borel*. We assume that all random variables live in some probability space $(\Omega, \mathcal{F}, P)$.

Using stochastic realization results (see [21, Lemma F], [144, Lemma 1.2], or [56, Lemma 3.1]), it follows that the model above in (2.1) contains the class of all $(\mathbb{X} \times \mathbb{U})^{\mathbb{Z}_+}$-valued stochastic processes which satisfy the following probabilistic characterization: for all Borel sets $B \in \mathcal{B}(\mathbb{X})$, $t \geq 0$, and P-almost all realizations $x_{[0,t]}, u_{[0,t]}$:

$$P(x_{t+1} \in B | x_{[0,t]} = a_{[0,t]}, u_{[0,t]} = b_{[0,t]}) = P(x_{t+1} \in B | x_t = a_t, u_t = b_t) =: \mathcal{T}(B|a_t, b_t) \tag{2.2}$$

where $\mathcal{T}(\cdot|x, u)$ is a *stochastic kernel* from $\mathbb{X} \times \mathbb{U}$ to $\mathbb{X}$ (so that for every B, $\mathcal{T}(B|\cdot, \cdot)$ is a measurable function on $\mathbb{X} \times \mathbb{U}$, and for every fixed $(a, b) \in \mathbb{X} \times \mathbb{U}$, $\mathcal{T}(\cdot|a, b)$ is a probability measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. That is, all stochastic processes that satisfy (2.2) admit a realization in the form (2.1), almost surely. Furthermore, we may take $\mathbb{W} = [0, 1]$ without any loss. Since a system of the form (2.1) satisfies (2.2), it follows that the representations in these equations are equivalent.

A stochastic process which satisfies (2.2) is called a *controlled Markov chain*.

For the process $\{x_t, u_t\}$ to define a stochastic process, in addition to a transition kernel and an initial measure on x_0, we need to specify the dependence of u_t on the history of the process. Once this is established, through the extension theorems discussed earlier (and in particular the Ionescu-Tulcea Extension Theorem), one can construct a stochastic process $\{x_t, u_t, t \geq 0\}$. This dependence is defined by a *control policy*.

We start with Fully Observed Controlled Markov Models, otherwise known as Markov Decision Processes (or MDPs).

2.2 Fully Observed Markov Control Problem Model (MDP Models)

A Fully Observed Markov Control Problem, otherwise known as a Markov Decision Process (or MDP), is a five tuple

$$(\mathbb{X}, \mathbb{U}, \{\mathbb{U}(x), x \in \mathbb{X}\}, \mathcal{T}, c),$$

where

- $\mathbb{X}$ is the state space, assumed a standard Borel space.
- $\mathbb{U}$ is the action space, assumed a standard Borel space.
- $\mathbb{K} = \{(x,u) : u \in \mathbb{U}(x) \in \mathcal{B}(\mathbb{U}), x \in \mathbb{X}\}$ is the set of state, control pairs that are feasible. There might be different states where different control actions are possible/feasible. We will assume that $\mathbb{K}$ is standard Borel.
- $\mathcal{T}$ is a state transition kernel, that is $\mathcal{T}(A|x,u) = P(x_{t+1} \in A | x_t = x, u_t = u)$, as defined above.
- $c : \mathbb{K} \to \mathbb{R}_+$ is a cost function.

2.2.1 Classes of Control Policies

Admissible Control Policies Γ_A

Let $\mathbb{H}_0 := \mathbb{X}, \quad \mathbb{H}_t = \mathbb{H}_{t-1} \times \mathbb{K}$ for $t = 1, 2, \ldots$. We let I_t denote an element of $\mathbb{H}_t$, where $I_t = \{x_{[0,t]}, u_{[0,t-1]}\}$. A deterministic admissible control policy γ is a sequence of functions $\{\gamma_t, t \in \mathbb{Z}_+\}$ such that $\gamma : \mathbb{H}_t \to \mathbb{U}$ with $u_t = \gamma_t(I_t)$.

We can also state this as follows: Let us write U_t to emphasize that u_t is a realization of the action random variable U_t under an admissible policy, and likewise let us emphasize that H_t is a random variable with realization I_t (In the notes, we will follow this approach of using capital letters when the distinction of whether we are discussing a random variable or its realization needs to be particularly emphasized explicitly). We say that γ_t is a function measurable on $\sigma(H_t)$ in the sense that for every Borel $B \subset \mathbb{U}$, we have that

$$\{\omega : U_t(\omega) \in B\} = U_t^{-1}(B) \subset \sigma(H_t).$$

A randomized admissible control policy is a sequence $\gamma = \{\gamma_t, t \geq 0\}$ such that $\gamma : \mathbb{H}_t \to \mathcal{P}(\mathbb{U})$, with $\mathcal{P}(\mathbb{U})$ being the set of probability measures on $\mathbb{U}$, so that for every realization I_t, $\gamma_t(I_t)$ is a probability measure on $\mathbb{U}$. Once again, by stochastic realization arguments, this is equivalent to writing $u_t = \gamma_t(I_t, r_t)$ for some $[0,1]$-valued i.i.d. random variable r_t.

Markov Control Policies Γ_M

A deterministic Markov control policy γ is a sequence of functions $\{\gamma_t, t \in \mathbb{Z}_+\}$ with $\gamma_t : \mathbb{X} \times \mathbb{Z}_+ \to \mathbb{U}$ such that

$$u_t = \gamma_t(x_t),$$

for each $t \in \mathbb{Z}_+$. Hence, the control action only depends on the state and the time, and not the past history. A policy is randomized Markov if the induced strategic measure satisfies

$$P^\gamma(u_t \in C | I_t) = \gamma_t(u_t \in C | x_t), \quad C \in \mathcal{B}(\mathbb{U}),$$

for all t and P^γ-almost all x_t. Alternatively, we can write $u_t = \gamma_t(x_t, r_t)$ for some $[0,1]$-valued i.i.d. random variable r_t and measurable function γ_t for all $t \in \mathbb{Z}_+$.

Stationary Control Policies Γ_S

A deterministic stationary control policy γ is a sequence of identical functions $\{\gamma_t, t \in \mathbb{Z}_+\}$ where for some $f : \mathbb{X} \to \mathbb{U}$, $\gamma_t = f$ so that

$$u_t = f(x_t).$$

for each $t \in \mathbb{Z}_+$.

A policy is randomized stationary if

$$P^\gamma(u_t \in C|I_t) = f(u_t \in C|x_t), \quad C \in \mathcal{B}(\mathbb{U}),$$

for some stochastic kernel f. As earlier, alternatively, we can write $u_t = f(x_t, r_t)$ for some $[0,1]$-valued i.i.d. random variable r_t and measurable function f for all $t \in \mathbb{Z}_+$. Hence, the control selection is independent of the past history or time, given the current state x_t.

Often, we will simply identify the stage-wise constant map with the stationary policy γ by an abuse of notation, that is $\gamma := \{\gamma, \gamma, \cdots\}$.

As reviewed above and in *Chapter* 1, according to the Ionescu-Tulcea theorem [165] (or Kolmogorov's extension theorem), an initial probability measure μ on $\mathbb{X}$, a transition kernel $\mathcal{T}$, and a control policy γ define a unique probability measure P_μ^γ on $(\mathbb{X} \times \mathbb{U})^{\mathbb{Z}_+}$, which is called a *strategic measure* [283]. If the initial measure μ is known, sometimes we omit this subscript while discussing the strategic measure.

Consider for now that the objective to be minimized is given by: $J_\mathcal{N}(\nu_0, \gamma) := E_{\nu_0}^\gamma[(\sum_{t=0}^{T-1} c(x_t, u_t)) + c_N(x_N)]$, where ν_0 is the initial probability measure, that is $x_0 \sim \nu_0$. The goal is to find a policy γ^* so that

$$J_\mathcal{N}(\nu_0, \gamma^*) \leq J_\mathcal{N}(\nu_0, \gamma) \qquad \forall \gamma \in \mathit{\Gamma}_A.$$

Such a γ^* is called an **optimal policy**. Here γ can also be called a **strategy**, or a **law**.

2.3 Performance Criteria: Optimality and Stability

2.3.1 Several Optimality Criteria and Performance of Policy Classes

Consider a Markov control problem with an objective given as the minimization of

$$J_\mathcal{N}(\nu_0, \gamma) = E_{\nu_0}^\gamma\left[\left(\sum_{t=0}^{N-1} c(x_t, u_t)\right) + c_N(x_N)\right]$$

where ν_0 denotes the distribution on x_0 and c_N is a terminal state cost function. For the case with $x_0 = x$, so that $\nu_0 = \delta_x$, we often simply write

$$J_\mathcal{N}(\delta_x, \gamma) =: J_\mathcal{N}(x, \gamma) = E_{\delta_x}^\gamma[\sum_{t=0}^{N-1} c(x_t, u_t) + c_N(x_N)] = E^\gamma[\sum_{t=0}^{N-1} c(x_t, u_t) + c_N(x_N)|x_0 = x]$$

Such a cost problem is known as an expected *finite horizon cost criterion*.

We will also consider costs of the following form:

$$J_\beta(\nu_0, \gamma) = E_{\nu_0}^\gamma[\sum_{t=0}^{\infty} \beta^t c(x_t, u_t)],$$

for some $\beta \in (0,1)$. This is called an expected *discounted infinite horizon cost criterion*.

Finally, we will study costs of the following form:

$$J_\infty(\nu_0, \gamma) = \limsup_{N\to\infty} \frac{1}{N} E_{\nu_0}^\gamma[\sum_{t=0}^{N-1} c(x_t, u_t)]$$

Such a problem is known as an *infinite horizon average cost criterion*.

As before, let Γ_A denote the class of admissible policies, Γ_M denote the class of Markov policies, Γ_S denote the class of Stationary policies. These policies can be both randomized or deterministic. We may also denote the randomized policies with Γ_{RA}, Γ_{RM} and Γ_{RS} if randomization needs to be made explicit.

For each of the criteria above, in these notes, we will investigate existence, structural and approximation results and also computational and numerical as well as simulation based solution methods.

In a general setting, we note the following relation

$$\inf_{\gamma\in\Gamma_A} J_{\mathcal{N}}(\nu_0,\gamma) \le \inf_{\gamma\in\Gamma_M} J_{\mathcal{N}}(\nu_0,\gamma) \le \inf_{\gamma\in\Gamma_S} J_{\mathcal{N}}(\nu_0,\gamma),$$

since the sets of policies are progressively shrinking

$$\Gamma_S \subset \Gamma_M \subset \Gamma_A.$$

We will show, however, that for the optimal control of a Markov chain, under mild conditions, Markov policies are always optimal (that is there is no loss in optimality in restricting the policies to be Markov); that is, it is sufficient to consider only Markov policies. That is,

$$\inf_{\gamma\in\Gamma_A} J_{\mathcal{N}}(\nu_0,\gamma) = \inf_{\gamma\in\Gamma_M} J_{\mathcal{N}}(\nu_0,\gamma)$$

We will also show that, under somewhat more restrictive conditions, stationary policies are optimal (that is, there is no loss in optimality in restricting the policies to be stationary). This will typically exclude finite horizon problems and under mild conditions we will have that

$$\inf_{\gamma\in\Gamma_A} J_{\beta}(\nu_0,\gamma) = \inf_{\gamma\in\Gamma_S} J_{\beta}(\nu_0,\gamma), \quad \text{and} \quad \inf_{\gamma\in\Gamma_A} J_{\infty}(\nu_0,\gamma) = \inf_{\gamma\in\Gamma_{RS}} J_{\infty}(\nu_0,\gamma),$$

where we will also see that the infimum on the right hand side can be taken among those stationary policies which are deterministic under further conditions. Furthermore, we will show that, under some stronger conditions, $\inf_{\gamma\in\Gamma_S} J_{\infty}(\nu_0,\gamma)$ is independent of the initial probability measure ν_0 (or the initial condition) on x_0.

For further relations between such policies, see *Chapter* 5 and *Chapter* 7.

The last two results are computationally very important, as there are powerful computational algorithms that allow one to find such stationary policies. We will be discussing these later in the notes.

In the rest of the notes, we will first consider further properties of Markov chains, since under a Markov control policy, the controlled state becomes a Markov chain by Theorem 2.3.1 below. We will then get back to controlled Markov chains and the development of optimal control policies in *Chapter*s 5 and 7.

Further optimality criteria include *sample path optimality*, *risk-sensitive optimality* and *control up to a stopping time*. We will obtain structural results for optimal policies under these criteria as well, together with analytical results.

2.3.2 Stability as a Performance Criterion

In addition to, or instead of (depending on applications), optimality, one would like to achieve stability in a stochastic sense. Such stochastic stability may be in a variety of senses, and these will be discussed in detail in the upcoming chapters.

2.3.3 Markov Chain Induced by a Markov Policy

Theorem 2.3.1 *Let the control policy be randomized Markov. Then, the controlled Markov chain induces an $\mathbb{X}$-valued Markov chain, that is, the state process itself becomes a Markov chain:*

$$P^{\gamma}_{x_0}(x_{t+1}\in B|x_t=b_t,x_{t-1}=b_{t-1},\ldots,x_0=b_0) = Q^{\gamma}_t(x_{t+1}\in B|x_t=b_t), \quad B\in\mathcal{B}(\mathbb{X}), t\ge 1,$$

for P almost every realization of the past variables $b_t, \cdots, b_0$, where Q_t^γ is a possibly time-dependent stochastic kernel defining a Markov chain. If the control policy is a stationary policy, then the induced Markov chain $\{x_t\}$ is time-homogenous; that is, the transition kernel Q_t^γ for the induced Markov chain does not depend on time.

Proof. We will consider the case where $\mathbb{U}$ is countable, the uncountable case follows similarly. Let $B \in \mathcal{B}(\mathbb{X})$. It follows that,

$$\begin{aligned}
&P_{x_0}^\gamma(x_{t+1} \in B | x_t = b_t, x_{t-1} = b_{t-1}, \ldots, x_0 = b_0) \\
&= P_{x_0}^\gamma(x_{t+1} \in B, u_t \in \mathbb{U} | x_t = b_t, x_{t-1} = b_{t-1}, \ldots, x_0 = b_0) \\
&= P_{x_0}^\gamma(\cup_{u \in \mathbb{U}} \{x_{t+1} \in B, u_t = u\} | x_t = b_t, x_{t-1} = b_{t-1}, \ldots, x_0 = b_0) \\
&= \sum_{u \in \mathbb{U}} P_{x_0}^\gamma(x_{t+1} \in B, u_t = u | x_t = b_t, x_{t-1} = b_{t-1}, \ldots, x_0 = b_0) \\
&= \sum_{u \in \mathbb{U}} P_{x_0}^\gamma(x_{t+1} \in B | u_t = u, x_t = b_t, x_{t-1} = b_{t-1}, \ldots, x_0 = b_0) P_{x_0}^\gamma(u_t = u | x_t = b_t, x_{t-1} = b_{t-1}, \ldots, x_0 = b_0) \\
&= \sum_{u \in \mathbb{U}} \mathcal{T}(x_{t+1} \in B | u_t = u, x_t = b_t) \gamma_t(u_t = u | x_t = b_t) \\
&= \sum_{u \in \mathbb{U}} Q_t^\gamma(x_{t+1} \in B, u_t = u | x_t = b_t) \\
&= Q_t^\gamma(x_{t+1} \in B | x_t = b_t)
\end{aligned} \tag{2.3}$$

Here, Q_t^γ is a conditional probability measure defined with $Q_t^\gamma(x_{t+1} \in B, u_t = u | x_t = b_t) := \mathcal{T}(x_{t+1} \in B | u_t = u, x_t = b_t) \gamma_t(u_t = u | x_t = b_t)$. The essential issue here is that the control only depends on x_t, and since x_{t+1} depends stochastically only on x_t and u_t (being a controlled Markov chain), the desired result follows. If $\gamma_t(u_t | x_t = b_t) = \gamma(u_t | x_t = b_t)$, that is, $\gamma_t = \gamma$ for all t values so that the policy is stationary, the resulting chain satisfies

$$P_{x_0}^\gamma(x_{t+1} \in B | x_t, x_{t-1}, \ldots, x_0) = Q^\gamma(x_{t+1} \in B | x_t),$$

for some Q^γ. Thus, the transition kernel does not depend on time and the chain is time-homogenous. ◇

2.4 Partially Observed Models and Reduction to a Fully Observed Model

Consider a partially observable stochastic control problem with the following dynamics.

$$x_{t+1} = f(x_t, u_t, w_t), \qquad y_t = g(x_t, v_t).$$

Here, x_t is the $\mathbb{X}$-valued state, u_t is the $\mathbb{U}$-valued the control, y_t is the $\mathbb{Y}$-valued observation (measurement) process. Furthermore, (w_t, v_t) are i.i.d noise processes and $\{w_t\}$ is independent of $\{v_t\}$. The controller only has causal access to $\{y_t\}$.

As noted, y_t denotes an observation variable taking values in $\mathbb{Y}$, a subset of $\mathbb{R}^n$ in the context of this review. The controller only has causal access to the second component $\{y_t\}$ of the process: A **deterministic admissible control policy** γ is a sequence of functions $\{\gamma_t\}$ so that $u_t = \gamma_t(y_{[0,t]}; u_{[0,t-1]})$.

We will see in *Chapter* 6 that one could transform a partially observable Markov Decision Problem to a Fully Observed Markov Decision Problem via an enlargement of the state space.

Thus, the fully observed Markov Decision Model we will consider is sufficiently rich to be applicable to a large class of controlled stochastic systems. Partially Observable Markov Decision Problems, also known as POMDPs, will be studied in detail in *Chapter* 6.

2.5 Decentralized Stochastic Control

We will consider situations in which there are multiple decision makers acting on a system under a variety of information structures. These will be studied extensively in *Chapter* 10.

2.6 Controlled Continuous-Time Stochastic Systems

We will also study setups where the time index is a continuum. We will cover this material in *Chapter* 11.

2.7 Numerical Methods, Reinforcement Learning, and Robustness to Incorrect Models

While we will extensively study analytical methods to arrive at solutions, for many problems it is more convenient to consider numerical methods or stochastic learning methods. For some applications, this may be the only option, e.g. when a model is not known apriori. These will be studied in detail in *Chapter*s 8 and 9.

A good control design must be robust to perturbations in the model. This brings the question of continuity and robustness of optimal costs and optimal controls to perturbations in a model, where topological questions on model regularity are to be studied in detail. These also, as a special case, cover finite model approximations of systems with uncountable state and action spaces. These are studied in *Chapter* 12.

2.8 Bibliographic Notes

We are thankful to Prof. Eugene Feinberg on some historical remarks regarding stochastic realization and pointing out to Aumann's lemma: [21, Lemma F], and that this result may have also been due to Girsanov.

2.9 Exercises

Exercise 2.9.1 *a) Let f be an arbitrary measurable function from $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Show that a controlled stochastic process defined with*

$$x_{t+1} = f(x_t, u_t, w_t),$$

with $\{w_t\}$ an independent and identically distributed noise sequence is a controlled Markov chain.

b) Study Lemma 3.1 and Corollary 3.1 of [56].

Exercise 2.9.2 *A common example in mathematical finance applications is the portfolio selection problem where a controller (investor) would like to optimally allocate his wealth between a stochastic stock market and a market with a guaranteed income : Consider a stock with an i.i.d. random return σ_t and a bank account with fixed interest rate $r > 0$. These are modeled by:*

$$X_{t+1} = X_t u_t(1 + \sigma_t) + X_t(1 - u_t)(1 + r), \quad X_0 = 1$$

and

$$X_{t+1} = X_t(1 + r + u_t(\sigma_t - r))$$

Here, $u_t \in [0, 1]$ denotes the proportion of the money that the investor invests in the stock market. Suppose that the goal is to maximize $E[\log(X_T)]$. Then, we can write:

$$\log(X_T) = \log(\prod_{k=0}^{T-1} \frac{X_{k+1}}{X_k}) = \sum_{k=0}^{T-1} \log((1 + r + u_t(\sigma_t - r))) \tag{2.4}$$

Formulate the problem as an optimal stochastic control problem by clearly identifying the state and the control action spaces, the information available at the controller, the transition kernel, and a cost functional mapping the actions and states to $\mathbb{R}$.

Exercise 2.9.3 *Consider an inventory-production system given by*

$$x_{t+1} = x_t + u_t - w_t,$$

where x_t *is* $\mathbb{R}$*-valued, with the one-stage cost*

$$c(x_t, u_t, w_t) = bu_t + h \max(0, x_t + u_t - w_t) + p \max(0, w_t - x_t - u_t)$$

Here, b *is the unit production cost,* h *is the unit holding (storage) cost and* p *is the unit shortage cost; here we take* $p > b$*. At any given time, the decision maker can take* $u_t \in \mathbb{R}_+$*. The demand variable* $w_t \sim \mu$ *is a* $\mathbb{R}_+$*-valued i.i.d. process, independent of* x_0*, with a finite mean where* μ *is assumed to admit a probability density function. The goal is to minimize*

$$J(x, \gamma) = E_x^{\gamma}[\sum_{t=0}^{T-1} c(x_t, u_t, w_t)]$$

The controller at time t *has access to* $I_t = \{x_s, u_s, s \leq t-1\} \cup \{x_t\}$.

Formulate the problem as an optimal stochastic control problem by clearly identifying the state, the control action spaces, the information available at the controller, the transition kernel and a cost functional mapping the actions and states to $\mathbb{R}$.

Exercise 2.9.4 *A fishery manager annually has* x_t *units of fish and sells* $u_t x_t$ *of these where* $u_t \in [0, 1]$*. With the remaining ones, the next year's production is given by the following model*

$$x_{t+1} = w_t x_t (1 - u_t) + w_t,$$

with x_0 *is given and* w_t *is an independent, identically distributed sequence of random variables and* $w_t \geq 0$ *for all* t *and therefore* $E[w_t] = \tilde{w} \geq 0$*.*

The goal is to maximize the profit over the time horizon $0 \leq t \leq T - 1$*. At time* T*, he sells all of the fish.*

Formulate the problem as an optimal stochastic control problem by clearly identifying the state, the control actions, the information available at the controller, the transition kernel and a cost functional mapping the actions and states to $\mathbb{R}$.

Exercise 2.9.5 *An investor's wealth dynamics is given by the following:*

$$x_{t+1} = u_t w_t,$$

where $\{w_t\}$ *is an i.i.d.* $\mathbb{R}_+$*-valued stochastic process with* $E[w_t] = 1$*. The investor has access to the past and current wealth information and his previous actions. The goal is to maximize:*

$$J(x_0, \gamma) = E_{x_0}^{\gamma}[\sum_{t=0}^{T-1} \sqrt{x_t - u_t}].$$

The investor's action set for any given x *is:* $\mathbb{U}(x) = [0, x]$.

Formulate the problem as an optimal stochastic control problem by clearly identifying the state, the control action spaces, the information available at the controller, the transition kernel and a cost functional mapping the actions and states to $\mathbb{R}$.

Exercise 2.9.6 *Consider an unemployed person who will have to work for years $t = 1, 2, ..., 10$ if she takes a job at any given t.*

Suppose that each year in which she remains unemployed; she may be offered a good job *that pays 10 dollars per year (with probability $1/4$); she may be offered a* bad job *that pays 4 dollars per year (with probability $1/4$); or she may not be offered a job (with probability $1/2$). These events of job offers are independent from year to year (that is the job market is represented by an independent sequence of random variables for every year).*

Once she accepts a job, she will remain in that job for the rest of the ten years. That is, for example, she cannot switch from the bad job to the good job.

Suppose the goal is maximize the expected total earnings in ten years, starting from year 1 up to year 10 (including year 10).

State the problem as a Markov Decision Problem, identify the state space, the action space and the transition kernel.

Exercise 2.9.7 (Zero-Delay Source Coding) *Let $\{x_t\}_{t\geq 0}$ be an $\mathbb{X}$-valued discrete-time Markov process where $\mathbb{X}$ can be a finite set or $\mathbb{R}^n$. Let there be an encoder which encodes (quantizes) the source samples and transmits the encoded versions to a receiver over a discrete noiseless channel with input and output alphabet $\mathcal{M} := \{1, 2, \ldots, M\}$, where M is a positive integer.*

The encoder policy η is a sequence of functions $\{\eta_t\}_{t\geq 0}$ with

$$\eta_t : \mathcal{M}^t \times (\mathbb{X})^{t+1} \ni (q_{[0,t-1]}, x_{[0,t]}) \mapsto q_t \in \mathcal{M}.$$

A zero-delay receiver policy is a sequence of functions $\gamma = \{\gamma_t\}_{t\geq 0}$ of type

$$\gamma_t : \mathcal{M}^{t+1} \ni q_{[0,t]} \mapsto u_t \in \mathbb{U}.$$

For the finite horizon setting the goal is to minimize the average cumulative cost (distortion)

$$J^T(\pi_0, \eta, \gamma) := E^{\eta,\gamma}_{\pi_0}\left[\frac{1}{T}\sum_{t=0}^{T-1} c_0(x_t, u_t)\right], \tag{2.5}$$

for some $T \geq 1$, where $c_0 : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is a nonnegative cost (distortion) function, and $E^{\eta,\gamma}_{\pi_0}$ denotes expectation with initial distribution π_0 for x_0 and under the quantization policy η and receiver policy γ.

Express this problem as a controlled Markov chain problem. Later on, we will provide further refinements. There is a rich history behind this problem, see e.g., [330], [321], [305] and [335, 343].

Exercise 2.9.8 *Suppose that there are two decision makers DM^1 and DM^2. Suppose that the information available to to DM^1 is a random variable Y^1 and the information available to DM^2 is Y^2, where these random variables are defined on a probability space $(\Omega, \mathcal{F}, P)$. Suppose that for $i = 1, 2$, Y^i is $\mathbb{Y}^i$-valued and these are standard Borel spaces. Let X be a $\mathbb{X}$-valued random variable defined on the same probability space where $\mathbb{X}$ is also a standard Borel space.*

Suppose that the sigma-field generated by Y^1 is a subset of the sigma-field generated by Y^2, that is $\sigma(Y^1) \subset \sigma(Y^2)$. That is, the information contained in Y^1 is a subset of the information contained in Y^2 (Recall here that the σ-field generated by a random variable Y is the smallest σ-field over Ω on which Y is measurable).

Further, suppose that the decision makers wish to minimize the following cost function:

$$E[c(X, U)],$$

where $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ is a measurable cost function. Let, for $i = 1, 2$, $U^i = \gamma^i(Y^i)$ be generated by a measurable function γ^i on the sigma-field generated by the random variable Y^i. Let Γ^i denote the space of all such functions (which we will refer to as policies).

Prove that

$$\inf_{\gamma^1 \in \Gamma^1} E[c(X, U^1)] \geq \inf_{\gamma^2 \in \Gamma^2} E[c(X, U^2)].$$

Hint: Make the argument that every policy $u^1 = \gamma^1(Y^1)$ can be expressed as $u^2 = \gamma^2(Y^2)$ for some $\gamma^2 \in \Gamma^2$; see Exercise 1.6.6.

3

Classification of Markov Chains

3.1 Countable State Space Markov Chains

In this section, we study Markov chains where the Markovian state takes values in a finite or a countably infinite set $\mathbb{X}$. In the following, we will consider $(\Omega, \mathcal{F}, \mathbf{P})$ to be the probability space on which all of the random variables are defined (later on, when a particular notational distinction is not needed, we will replace the notation $\mathbf{P}$ with P as the probability measure on the events to be considered).

We assume that ν_0 is an initial distribution for the Markov chain, so that $\mathbf{P}(x_0 \in \cdot) = \nu_0(\cdot)$ (also denoted with $x_0 \sim \nu_0$). The process $\Phi = \{x_0, x_1, \dots, x_n, \dots\}$ is a (time-homogeneous) Markov chain with its probability law (or the probability measure induced on the sequence space) satisfying, for $n \in \mathbb{Z}_+$:

$$\begin{aligned} &P_{\nu_0}(x_0 = a_0, x_1 = a_1, x_2 = a_2, \dots, x_n = a_n) \\ &:= \nu_0(x_0 = a_0)\mathbf{P}(x_1 = a_1 | x_0 = a_0)\mathbf{P}(x_2 = a_2 | x_1 = a_1) \dots \mathbf{P}(x_n = a_n | x_{n-1} = a_{n-1}) \end{aligned} \tag{3.1}$$

If the initial condition is known to be a fixed state $a_0 \in \mathbb{X}$, we use $P_{a_0}(\cdots)$ in place of $P_{\delta_{a_0}}(\cdots)$. We could represent the probabilistic evolution in terms of a matrix:

$$P(i,j) := \mathbf{P}(x_{t+1} = j | x_t = i) \geq 0, \qquad i, j \in \mathbb{X}.$$

Here $P(\cdot, \cdot)$ is a *probability transition kernel*, that is for every $i \in \mathbb{X}$, $P(i, .)$ is a probability measure on $\mathbb{X}$, in particular with $\sum_j P(i,j) = 1$ for every i. Let P be the $|\mathbb{X}| \times |\mathbb{X}|$ matrix with entries given with $P(i,j) \geq 0$. Such a matrix P is called a *stochastic matrix*.

The initial condition probability and the transition kernel uniquely identify the probability measure on the product space $\mathbb{X}^{\mathbb{N}}$, by the extension theorems presented in *Chapter 1*.

Let $\pi_k(i) = \mathbf{P}(x_k = i)$ for $i \in \mathbb{X}$, and for all $k \in \mathbb{Z}_+$. Let $\pi_k = [\pi_k(i), i \in \mathbb{X}]$. It follows that

$$\pi_1(j) = \mathbf{P}(x_1 = j) = \sum_{i \in \mathbb{X}} \mathbf{P}(x_1 = j, x_0 = i) = \sum_{i \in \mathbb{X}} \mathbf{P}(x_1 = j | x_0 = i)\mathbf{P}(x_0 = i) = \sum_{i \in \mathbb{X}} \pi_0(i) P(i,j)$$

and with P denoting the transition matrix given with $P(i,j)$ as defined above, by a similar reasoning,

$$\pi_{k+1} = \pi_k P, \qquad k \in \mathbb{Z}_+ \tag{3.2}$$

Note here that we represent π_k as a row vector. By induction, we could verify that for $k \in \mathbb{N}$:

$$P^k(i,j) := P(x_{t+k} = j | x_t = i) = \sum_{m \in \mathbb{X}} P(i,m) P^{k-1}(m,j)$$

We will see that whether the sequence $\{\pi_k, k \in \mathbb{Z}_+\}$ admits a limit and the dependence properties of this limit on π_0 have significant implications on the characterization of Markov chains, and later, in stabilization and optimization of controlled Markov chains.

In the following, we characterize Markov Chains based on transience, recurrence and communication properties. We then consider the problem of the existence of an invariant probability measure. Later, we will extend the analysis to uncountable space Markov chains.

Communication

If there exists an integer $k \in \mathbb{N}$ such that $\mathbf{P}(x_{t+k} = j | x_t = i) = P^k(i,j) > 0$, and an integer $l \in \mathbb{N}$ such that $P(x_{t+l} = i | x_t = j) = P^l(j,i) > 0$ then state i **communicate**s with state j.

A set $C \subset \mathbb{X}$ is said to be communicating if every two elements (states) of C communicate with each other.

If every member of the set communicates with every other member, such a chain is said to be **irreducible**.

The period of a state $i \in \mathbb{X}$ is defined to be the greatest common divisor of $\{k > 0 : P^k(i,i) > 0\}$.

A Markov chain is called aperiodic if the period of all states is 1.

Absorbing Set

A set C is called **absorbing** if $P(i,C) = 1$ for all $i \in C$. That is, if the state is in C, then the state cannot get out of C.

The Markov chain is **irreducible** if the smallest absorbing set is the entire $\mathbb{X}$ itself.

The Markov chain is **indecomposable** if $\mathbb{X}$ does not contain two disjoint absorbing sets.

Occupation, Hitting and Stopping Times

For any set $A \subset \mathbb{X}$, the *occupation time* η_A is the number of visits of $\{x_t\}$ to set A:

$$\eta_A = \sum_{t=0}^{\infty} 1_{\{x_t \in A\}},$$

where 1_E denotes the indicator function for an event E, that is, it takes the value 1 when E takes place, and is otherwise 0.

Remark 3.1. Another common notation for the indicator function is the following: Let A be an event (a subset of some σ-field). Then $1_A(x) = 1$ if $x \in A$ and 0 otherwise.

Let $A \subset \mathbb{X}$. Define

$$\tau_A := \min\{k > 0 : x_k \in A\},$$

to be the first time that the state visits A; we call this the *return time* to set A. We also define a very similar notion, called a *hitting time*:

$$\sigma_A = \min\{k \geq 0 : x_k \in A\}.$$

The variable τ_A defined above is an example for **stopping time**s:

Definition 3.1.1 *A $\mathbb{Z}_+ \cup \{\infty\}$-valued random variable τ is a stopping time (with respect to the σ-field generated by the process $\{x_0, x_1, \cdots\}$), if for all $n \in \mathbb{Z}_+$, the event $\{\tau = n\} \in \sigma(x_0, x_1, x_2, \ldots, x_n)$, that is the event is in the sigma-field generated by the random variables up to time n.*

Any realistic decision takes place at a time which is a stopping time. Consider an optimal investment problem: if an investor claims to stop investing (e.g., purchasing houses) when the investment (value of the housing market) is at its local peak, the decision instant could not be a stopping-time in general: this peak-time is not a stopping time because to find out whether the investment value is at its peak, the next realization should be known, and this information is not available up to any given time in a causal fashion for a non-trivial (i.e., non-deterministic) stochastic process.

One important property of Markov chains is the *strong Markov property*. This says the following: If we sample a Markov chain according to a stopping time rule, the sampled Markov chain starts from the sampled instant as a Markov chain with the same transition probabilities as if the sampling instant is time 0:

Proposition 3.1.1 *For a (time-homogenous) Markov chain with a countable state space $\mathbb{X}$, the strong Markov property holds: that is, if τ is a stopping time with $P(\tau < \infty) = 1$, then almost surely for any $m \in \mathbb{N}$:*

$$P(x_{\tau+m} = a | x_\tau = b_0, x_{\tau-1} = b_1, \dots) = P(x_{\tau+m} = a | x_\tau = b_0) = P^m(b_0, a).$$

Proof. We consider $m = 1$; for larger m, the result follows from identical steps. For an event with $\{x_\tau = b_0, x_{\tau-1} = b_1, \dots\}$ with $P(x_\tau = b_0, x_{\tau-1} = b_1, \dots) > 0$, we have that

$$\begin{aligned}
&P(x_{\tau+1} = a | x_\tau = b_0, x_{\tau-1} = b_1, \dots) \\
&= \frac{P(x_{\tau+1} = a, x_\tau = b_0, x_{\tau-1} = b_1, \dots)}{P(x_\tau = b_0, x_{\tau-1} = b_1, \dots)} \\
&= \frac{\sum_{k=0}^{\infty} P(\tau = k, x_{\tau+1} = a, x_\tau = b_0, x_{\tau-1} = b_1, \dots)}{P(x_\tau = b_0, x_{\tau-1} = b_1, \dots)} \qquad (3.3) \\
&= \frac{\sum_{k=0}^{\infty} P(x_{k+1} = a | \tau = k, x_k = b_0, x_{k-1} = b_1, \dots) P(\tau = k, x_k = b_0, x_{k-1} = b_1, \dots)}{P(x_\tau = b_0, x_{\tau-1} = b_1, \dots)} \\
&= \frac{\sum_{k=0}^{\infty} P(x_{k+1} = a | x_k = b_0, x_{k-1} = b_1, \dots) P(\tau = k, x_k = b_0, x_{k-1} = b_1, \dots)}{P(x_\tau = b_0, x_{\tau-1} = b_1, \dots)} \qquad (3.4) \\
&= \frac{\sum_{k=0}^{\infty} P(x_{k+1} = a | x_k = b_0) P(\tau = k, x_k = b_0, x_{k-1} = b_1, \dots)}{P(x_\tau = b_0, x_{\tau-1} = b_1, \dots)} \\
&= P(b_0, a) \frac{\sum_{k=0}^{\infty} P(\tau = k, x_\tau = b_0, x_{\tau-1} = b_1, \dots)}{P(x_\tau = b_0, x_{\tau-1} = b_1, \dots)} \\
&= P(b_0, a) \frac{P(x_\tau = b_0, x_{\tau-1} = b_1, \dots)}{P(x_\tau = b_0, x_{\tau-1} = b_1, \dots)} \\
&= P(b_0, a) \qquad (3.5)
\end{aligned}$$

Note that the assumption $P(\tau < \infty) = 1$ is critically used in the proof in (3.3). In (3.4), we use the fact that τ is a stopping time. ◇

3.1.1 Recurrence and transience

Let us define

$$U(x, A) := E[\sum_{t=1}^{\infty} 1_{(x_t \in A)} | x_o = x] = \sum_{t=1}^{\infty} P^t(x, A) =: E_x[\sum_{t=1}^{\infty} 1_{(x_t \in A)}$$

and define

$$L(x, A) := P(\tau_A < \infty | x_0 = x) =: P_x(\tau_A < \infty),$$

which is the probability of the chain visiting set A, once the process starts at state x.

Definition 3.1.2 *(i) A set $A \subset \mathbb{X}$ is* **recurrent** *if the Markov chain visits A infinitely often in expectation, when the process starts in A:*

$$E_x[\eta_A] = \infty, \qquad \forall x \in A \tag{3.6}$$

(ii) A state $\alpha \in \mathbb{X}$ is **transient** *if*

$$U(\alpha, \alpha) = E_\alpha[\eta_\alpha] < \infty. \tag{3.7}$$

(iii)A set $A \subset \mathbb{X}$ is **positive recurrent** *if*

$$E_x[\tau_A] < \infty, \qquad \forall x \in A.$$

In particular, if a state $\alpha \in \mathbb{X}$ is not recurrent, it is transient.

Equation (3.7) can also be written as $\sum_{i=1}^{\infty} P^i(\alpha, \alpha) < \infty$, which in turn is implied by

$$P_i(\tau_i < \infty) < 1,$$

as we will show further below. The reader should connect the above with the strong Markov property: once the process hits a state, it starts from the state as if it is time 0 (regardless of the the past); the process recurs itself.

There is another important notion of recurrence, called *Harris recurrence*:

Definition 3.1.3 *(i) A set A is* **Harris recurrent** *if $P_x(\eta_A = \infty) = 1$ for all $x \in A$.*

(ii) An irreducible Markov chain is Harris recurrent if

$$P_x(\eta_A = \infty) = 1, \quad \forall x \in \mathbb{X}, A \subset \mathbb{X}.$$

Let $\tau_i(1) := \tau_i$ and for $i \geq 1$,

$$\tau_i(k+1) = \min\{n > \tau_i(k) : x_n = i\}$$

We have the following result whose proof, which builds on *continuity of probability* (Theorem B.1.2), is presented later in the chapter in a more general context in Theorem 3.2.1.

Theorem 3.1.1 *The condition $P_i(\tau_i < \infty) = 1$ is equivalent to the condition $P_i(\eta_i = \infty) = 1$.*

One can verify that (3.7) is equivalent to $L(i, i) < 1$.

Theorem 3.1.2 *If $P_i(\tau_i < \infty) < 1$, then $E_i[\eta_i] < \infty$ and thus the state $i \in \mathbb{X}$ is transient.*

To show this, it suffices to first verify the relation

$$P_i(\tau_i(k) < \infty) = P_i(\tau_i(k-1) < \infty) P_i(\tau_i(1) < \infty),$$

and then use the equality $E[\eta] = \sum_{k=1}^{\infty} P(\eta \geq k)$.

We will investigate the Harris recurrence property further while studying uncountable state space Markov chains, however one needs to note that even for countable state space chains Harris recurrence is stronger than recurrence as we make explicit next.

Remark 3.2. Harris recurrence is stronger than recurrence. In one, an expectation is considered; in the other, a probability is considered. Consider the following example: Let $\mathbb{X} = \mathbb{N}$, $P(1, 1) = 1$ and for $x > 1$: $P(x, x+1) = 1 - 1/x^2$ and $P(x, 1) = 1/x^2$. Then, for $x \geq 2$ (see Exercise 3.5.7):

$$P_x(\tau_1 = \infty) = \prod_{t \geq x, t \in \mathbb{N}} (1 - 1/t^2) > 0.$$

Thus, the set $\{1, 2\}$ is not Harris recurrent, but it is recurrent.

3.1.2 Stability and invariant measures

Stability is an important concept, but it has different meanings in different contexts. This notion will be made more precise in the following chapter. Nonetheless, perhaps the weakest form of stochastic stability in the context of these notes is the existence of an invariant probability measure.

Recall from (3.2) that the occupation probabilities satisfy the recursions:

$$\pi_1 = \pi_0 P$$

And for $t > 1$:

$$\pi_{t+1} = \pi_t P = \pi_0 P^{t+1}$$

One important property of Markov chains is whether the above iteration leads to a fixed point. Such a fixed point π is called an **invariant probability measure**. Thus, a probability measure in a countable state Markov chain is invariant if

$$\pi = \pi P$$

This is equivalent to

$$\pi(j) = \sum_{i \in \mathbb{X}} \pi(i) P(i, j), \quad \forall j \in \mathbb{X}$$

We note that, if such a π exists, it must be written in terms of $\pi = \pi_0 \lim_{t\to\infty} P^t$, for some π_0. Clearly, π_0 can be π itself, but often π_0 can be any initial probability measure under irreducibility/aperiodicity conditions (where aperiodicity can be relaxed if convergence of the averages $\lim_{t\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \pi_0 P^t$ is considered) which will be discussed further. Invariant probability measures are especially important in stochastic control, due to ergodicity theorems (which show that temporal averages converge to statistical averages with probability 1), as we will discuss later in the chapter. Finally, how fast $\frac{1}{T} \sum_{t=0}^{T-1} \pi_0 P^t$ converges to invariance is another very important question to be studied.

3.1.3 Invariant measures via an occupational characterization

The following is one of the most consequential results in this chapter.

Theorem 3.1.3 *For a Markov chain, if there exists an element i such that $E_i[\tau_i] < \infty$; the following is an invariant probability measure:*

$$\mu(j) = E\left[\frac{\sum_{k=0}^{\tau_i - 1} 1_{\{x_k = j\}}}{E_i[\tau_i]} \middle| x_0 = i \right], \qquad j \in \mathbb{X}$$

Proof. We will show for every $j \in \mathbb{X}$ that

$$E\left[\frac{\sum_{k=0}^{\tau_i - 1} 1_{\{x_k = j\}}}{E[\tau_i]} \middle| x_0 = i \right] = \sum_{s \in \mathbb{X}} P(s, j) E\left[\frac{\sum_{k=0}^{\tau_i - 1} 1_{\{x_k = s\}}}{E[\tau_i]} \middle| x_0 = i \right],$$

which establishes the desired result[1]. Note that $E[1_{\{X_{t+1}=j\}}] = P(X_{t+1} = j)$ and $P(s, j) = E[1_{\{X_{k+1}=j\}} | X_k = s] = E[1_{\{X_{k+1}=j\}} | X_k](\omega)$ with $X_k(\omega) = s$. Hence,

$$\sum_{s \in \mathbb{X}} P(s, j) E\left[\frac{\sum_{k=0}^{\tau_i - 1} 1_{\{X_k = s\}}}{E_i[\tau_i]} \middle| X_0 = i \right]$$

[1] In the following, to make the random nature of x_k terms explicit, we will use capital letters X_k to emphasize randomness. In the notes, we will occasionally follow this route, since for conditional expectations, often it is very crucial to distinguish between random variables and their realizations.

$$
\begin{aligned}
&= E\left[\frac{\sum_{k=0}^{\tau_i-1}\sum_{s\in\mathbb{X}} P(s,j)1_{\{X_k=s\}}}{E_i[\tau_i]}\middle| X_0=i\right]\\
&= \frac{E\left[\sum_{k=0}^{\tau_i-1}\sum_{s\in\mathbb{X}}\left(1_{\{X_k=s\}}E[1_{\{X_{k+1}=j\}}|X_k=s]\right)\middle| x_0=i\right]}{E_i[\tau_i]}\\
&= \frac{E\left[\sum_{k=0}^{\tau_i-1}\sum_{s\in\mathbb{X}}\left(1_{\{X_k=s\}}E[1_{\{X_{k+1}=j\}}|X_k]\right)\middle| x_0=i\right]}{E_i[\tau_i]} \qquad (3.8)\\
&= \frac{E\left[\sum_{k=0}^{\tau_i-1}\sum_{s\in\mathbb{X}}1_{\{X_k=s\}}E[1_{\{x_{k+1}=j\}}|X_k,X_{k-1},X_{k-2},\cdots,X_0=i]\middle| X_0=i\right]}{E_i[\tau_i]} \qquad (3.9)\\
&= \frac{E\left[\sum_{k=0}^{\tau_i-1}\sum_{s\in\mathbb{X}}E[1_{\{X_k=s\}}1_{\{X_{k+1}=j\}}|X_k,X_{k-1},X_{k-2},\cdots,X_0=i]\middle| X_0=i\right]}{E_i[\tau_i]} \qquad (3.10)\\
&= \frac{E\left[\sum_{k=0}^{\infty}\sum_{s\in\mathbb{X}}1_{\{k<\tau_i\}}E[1_{\{X_k=s\}}1_{\{X_{k+1}=j\}}|X_k,X_{k-1},X_{k-2},\cdots,X_0=i]\middle| X_0=i\right]}{E_i[\tau_i]}\\
&= \frac{\sum_{k=0}^{\infty}\sum_{s\in\mathbb{X}}E\left[E[1_{\{k<\tau_i\}}1_{\{X_k=s\}}1_{\{X_{k+1}=j\}}|X_k,X_{k-1},X_{k-2},\cdots,X_0=i]\middle| X_0=i\right]}{E_i[\tau_i]} \qquad (3.11)\\
&= \frac{\sum_{k=0}^{\infty}E\left[1_{\{k<\tau_i\}}\sum_{s\in\mathbb{X}}1_{\{X_k=s\}}1_{\{X_{k+1}=j\}}\middle| X_0=i\right]}{E_i[\tau_i]} \qquad (3.12)\\
&= \frac{E\left[\sum_{k=0}^{\infty}1_{\{k<\tau_i\}}1_{\{X_{k+1}=j\}}|X_0=i\right]}{E_i[\tau_i]}\\
&= \frac{E\left[\sum_{k=0}^{\tau_i-1}1_{\{X_{k+1}=j\}}|X_0=i\right]}{E_i[\tau_i]} = E_i\left[\frac{\sum_{k=0}^{\tau_i-1}1_{\{X_{k+1}=j\}}}{E_i[\tau_i]}\right]\\
&= E_i\left[\frac{\sum_{k=1}^{\tau_i}1_{\{X_k=j\}}}{E_i[\tau_i]}\right] = E_i\left[\frac{\sum_{k=0}^{\tau_i-1}1_{\{X_k=j\}}}{E_i[\tau_i]}\right]\\
&= \mu(j),
\end{aligned}
$$

where we use the fact that the total number of visits to a given set does not change whether we include either $t=0$ or τ_i, since $X_0 = X_{\tau_i} = i$. Here, (3.8) follows from the fact that $X_k = s$ is specified so that $1_{\{X_k=s\}}E[1_{\{X_{k+1}=j\}}|X_k] = 1_{\{X_k=s\}}E[1_{\{X_{k+1}=j\}}|X_k=s]$, (3.9) follows from the fact that the process is a Markov chain, (3.10) and (3.11) follow from the properties of conditional expectation and that τ_i is a stopping time (we will discuss such properties in *Chapter 4*), and (3.12) follows from the law of the iterated expectations, see Theorem 4.1.3. In the above (3.8) follows from the fact that $1_{\{X_k=s\}}E[1_{\{X_{k+1}=j\}}|X_k] = E[1_{\{X_k=s\}}1_{\{X_{k+1}=j\}}|X_k]$.

Finally, observe that if $E_i[\tau_i] < \infty$, then the above measure indeed is a probability measure, as it follows that

$$\sum \mu(j) = \sum_j E\left[\frac{\sum_{k=0}^{\tau_i-1}1_{\{X_k=j\}}}{E[\tau_i]}\middle| X_0=i\right] = 1.$$

This concludes the proof. ◇

Theorem 3.1.4 *Every finite state space Markov chain admits an invariant probability measure.*

A common proof technique on the existence of invariant probability measures for finite state Markov chains builds on an important result called the *Perron-Frobenius Theorem*. However, we will present a more comprehensive result in the context of general space Markov chains later in Theorem 3.3.1.

Theorem 3.1.5 *For an irreducible Markov chain with countable $\mathbb{X}$, there can be at most one invariant probability measure.*

Proof. Let $\pi(i)$ and $\pi'(i)$ be two different invariant probability measures. Define $D := \{i : \pi(i) > \pi'(i)\}$. Then,

$$\pi(D) = \sum_{i \in D} \pi(i)P(i, D) + \sum_{i \notin D} \pi(i)P(i, D)$$

$$\pi'(D) = \sum_{i \in D} \pi'(i)P(i, D) + \sum_{i \notin D} \pi'(i)P(i, D)$$

implies that

$$\pi(D) - \pi'(D) = \sum_{i \in D} (\pi(i) - \pi'(i))P(i, D) + \sum_{i \notin D} (\pi(i) - \pi'(i))P(i, D)$$

and thus

$$\sum_{i \in D} (\pi(i) - \pi'(i))(1 - P(i, D)) = \sum_{i \notin D} (\pi(i) - \pi'(i))P(i, D)$$

The first term is strictly positive (since $P(i, D) = 1$ cannot hold for all $i \in D$ due to irreducibility, for otherwise D would be absorbing). The second term is not positive, hence a contradiction. ◇

Remark 3.3. One can see that for any $a \in \mathbb{X}$ with $\pi(a) > 0$, it must be that $E_a[\tau_a] < \infty$. The reason will be evident once we study Theorem 3.2.7 and consider the relation:

$$1 = \pi(\mathbb{X}) = \pi(a)E_a[\sum_{k=0}^{\tau_a - 1} 1_{\{x_k \in \mathbb{X}\}}] = \pi(a)E_a[\tau_a]$$

An implication of the above is the following very important result, which is a special case of Kac's lemma (see also Theorem

Theorem 3.1.6 (Kac's Lemma) *Let $\{x_t\}$ be irreducible and π be its (unique) invariant probability measure. Then, for all $i \in \mathbb{X}$ with $\pi(i) > 0$,*

$$\pi(i) = \frac{1}{E_i[\tau_i]}, \qquad i \in \mathbb{X}.$$

Remark 3.4. Consider the random walk on $\mathbb{Z}$ given with the transition kernel $P(x, x + 1) = P(x, x - 1) = \frac{1}{2}$ for $z \in \mathbb{Z}$. In this case, we have that for every $i \in \mathbb{Z}$, $E_i[\tau_i] = \infty$, and hence there does not exist an invariant **probability measure**. But, it has an invariant *measure* defined with: $\mu(\{i\}) = K, \quad i \in \mathbb{Z}$, for an arbitrary (fixed) $K \in \mathbb{R}$. That $E_i[\tau_i] = \infty$ can be established through the following reasoning: if there were an invariant probability measure, this would be unique by irreducibility and also for every state i, the measure $\frac{1}{E_i[\tau_i]}$ would take the same value. But the sum of these (countably infinitely many) identical values would need to be 1, leading to a contradiction. Then, $E_i[\tau_i]$ cannot be finite for any i.

3.1.4 Rates of convergence to invariant measures and Dobrushin's ergodic coefficient

Consider the iteration $\pi_{t+1} = \pi_t P$, with a given π_0. We would like to know when this iteration converges to a limit and how fast this convergence is. Here, the reader is referred to Appendix A for a review of vector and function spaces.

A map T from one complete normed linear (that is, a Banach) space $\mathbb{X}$ to itself is called a contraction if for some $0 \leq \rho < 1$

$$\|T(x) - T(y)\| \leq \rho \|x - y\|, \qquad \forall x, y \in \mathbb{X}.$$

Theorem 3.1.7 *A contraction map T in a Banach space has a unique fixed point x^* with $x^* = T(x^*)$. Furthermore, the iterates $x_{n+1} = T(x_n)$, for any given x_0, converge to x^* geometrically fast in the sense that $\|x_n - x^*\| \leq L_{x_0}\rho^n$ for some $L_{x_0} < \infty$.*

Proof. $\{T^n(x)\}$ forms a Cauchy sequence: First note that, $\|T^k(x) - T^{k-1}(x)\| \leq \|T(T^{k-1}(x)) - T(T^{k-2}(x))\| \leq \rho\|T^k(x) - T^{k-1}(x)\| \leq \cdots \leq \rho^{k-1}\|T(x) - x)\|$. Then,

$$\|T^n(x) - x\| \leq \sum_{k=1}^{n} \|T^k(x) - T^{k-1}(x)\| \leq \sum_{k=1}^{n} \rho^{k-1}\|T(x) - x)\| \leq \|T(x) - x)\|\frac{1}{1-\rho}$$

implying that

$$\|T^n(x)\| \leq \|x\| + \|T^n(x) - x\| \leq \|x\| + \|T(x) - x)\|\frac{1}{1-\rho} =: M(x)$$

uniformly over all n. Now, for every $n, m \geq N$, we have that $\|T^n(x) - T^m(x)\| \leq \rho^N\|T^{n-N}(x) - T^{m-N}(x)\| \leq 2M(x)\rho^N$. This implies that the sequence is Cauchy. By completeness, the Cauchy sequence has a limit, x^*. For uniqueness, suppose that there are two (different) fixed points with $u = T(u)$ and $v = T(v)$. Then $\|u - v\| = \|T(u) - T(v)\| \leq \rho\|u - v\| < \|u - v\|$, a contradiction. Thus, $u = v$. The rate of convergence follows by writing $\|x_n - x^*\| = \|T^n(x_0) - T^n(x^*)\| \leq \rho^n\|x_0 - x^*\|$. ◇

Contraction Mapping via Dobrushin's Ergodic Coefficient Consider a countable state Markov Chain with one-step transition kernel P. Define the Dobrushin coefficient as

$$\delta(P) = \min_{i,k}\left(\sum_{j\in\mathbb{X}} \min(P(i,j), P(k,j))\right) \tag{3.13}$$

Observe that for two scalars a, b

$$|a - b| = a + b - 2\min(a, b).$$

Let us define for a vector v the l_1 norm:

$$||v||_1 = \sum_{i\in\mathbb{X}} |v_i|.$$

The set of all countable index real-valued vectors (that is functions which map $\mathbb{Z} \to \mathbb{R}$) with a finite l_1 norm

$$\{v : ||v||_1 < \infty\}$$

is a complete normed linear space, and as such, is a Banach space. With these observations, we state the following:

Theorem 3.1.8 *[Dobrushin] [106] For any two probability measures π, π', it follows that*

$$||\pi P - \pi' P||_1 \leq (1 - \delta(P))||\pi - \pi'||_1.$$

Accordingly, the sequence of iterates $\pi_{n+1} = T(\pi_n) := \pi_n P$, for any given π_0, converges to invariance geometrically fast.

Proof. Let $\psi(i) = \pi(i) - \min(\pi(i), \pi'(i))$ for all $i \in \mathbb{X}$. Further, let $\psi'(i) = \pi'(i) - \min(\pi(i), \pi'(i))$. Since

$$0 = \sum_i \pi(i) - \pi'(i) = \sum_{i:\pi(i)>\pi'(i)} \pi(i) - \pi'(i) + \sum_{i:\pi'(i)>\pi(i)} \pi(i) - \pi'(i)$$

we have that $||\psi||_1 = ||\psi'||_1$, and since

$$\sum_i |\pi(i) - \pi'(i)| = \sum_{i:\pi(i)>\pi'(i)} \psi(i) + \sum_{i:\pi'(i)>\pi(i)} \psi'(i)$$

we have that

$$\sum_i |\pi(i) - \pi'(i)| = ||\psi||_1 + ||\psi'||_1$$

and thus

$$||\pi - \pi'||_1 = ||\psi - \psi'||_1 = 2||\psi||_1 = 2||\psi'||_1$$

Now,

$$\begin{aligned}
||\pi P - \pi' P||_1 &= ||\psi P - \psi' P||_1 \\
&= \sum_j |\sum_i \psi(i)P(i,j) - \sum_k \psi'(k)P(k,j)| \\
&= \frac{1}{||\psi'||_1} \sum_j |\sum_k \sum_i \psi(i)\psi'(k)P(i,j) - \psi(i)\psi'(k)P(k,j)| && (3.14) \\
&\leq \frac{1}{||\psi'||_1} \sum_j \sum_k \sum_i \psi(i)\psi'(k)|P(i,j) - P(k,j)| && (3.15) \\
&= \frac{1}{||\psi'||_1} \sum_k \sum_i \psi(i)\psi'(k) \sum_j |P(i,j) - P(k,j)| \\
&= \frac{1}{||\psi'||_1} \sum_k \sum_i |\psi(i)||\psi'(k)| \Big\{ \sum_j P(i,j) + P(k,j) - 2\min(P(i,j), P(k,j)) \Big\} && (3.16) \\
&\leq \frac{1}{||\psi'||_1} \sum_k \sum_i |\psi(i)||\psi'(k)|(2 - 2\delta(P)) && (3.17) \\
&= ||\psi'||_1(2 - 2\delta(P)) && (3.18) \\
&= ||\pi - \pi'||_1(1 - \delta(P))
\end{aligned}$$

In the above, (3.14) follows from adding terms in the summation, (3.15) from taking the norm inside, (3.16) follows from the relation $||a - b|| = a + b - 2\min(a, b)$, (3.17) from the definition of $\delta(P)$ and finally (3.18) follows from the l_1 norms of ψ, ψ'.

Thus, the map $\pi P : \pi \in \mathcal{P}(\mathbb{X}) \mapsto \pi P \in \mathcal{P}(\mathbb{X})$, where $\mathcal{P}(\mathbb{X})$ is the set of probability measures on $\mathbb{X}$ viewed as a subset of $l_1(\mathbb{X}; \mathbb{R})$, is a contraction mapping if $\delta(P) > 0$. As a result, the sequence $\{\pi_0 P^n, n \in \mathbb{Z}_+\}$ is Cauchy by Theorem 3.1.7, and as every Cauchy sequence in a Banach space has a limit, so does this process. We emphasize that the set of probability measures is not a linear space, but viewed as a closed subset of $l_1(\mathbb{X}; \mathbb{R})$, the sequence will have a limit. Since πP is also a probability measure for every $\pi \in \mathcal{P}(\mathbb{X})$, the limit must also be a probability measure. The limit is the invariant probability measure. ◇

It should be emphasized that Dobrushin's theorem tells us how fast the sequence of probability measures $\{\pi_0 P^n\}$ converges to the invariant probability measure π for an arbitrary π_0: since $\pi P^n = \pi$, we have that

$$\|\pi_0 P^n - \pi\|_1 = \|\pi_0 P^n - \pi P^n\|_1 \leq (1 - \delta(P))^n \|\pi_0 - \pi\|_1 \leq (1 - \delta(P))^n, \qquad n \in \mathbb{Z}_+$$

3.1.5 Ergodic theorem for countable state space chains

In Exercise 4.5.11, we will prove the ergodic theorem: let $\{x_t\}$ be a Harris recurrent Markov chain with an invariant probability measure μ (such a process is called *positive Harris recurrent*, as we will define in the next section). We then have that for every fixed initial state, almost surely

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f(x_t) = \sum_i f(i)\mu(i)$$

for bounded f (if f is not bounded, we require that $\sum_i |f(i)|\mu(i) < \infty$). This is a very important theorem, as this property is what establishes an important connection with average cost stochastic control. Under a stationary control policy leading

to a unique invariant probability measure μ on the state and control process (which is a Markov chain), with a bounded function c it follows that almost surely,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}c(x_t,u_t)=\sum_{x,u}c(x,u)\mu(x,u).$$

3.2 Uncountable Standard Borel State Spaces

We now extend the discussion above to the uncountable state space setting. We will consider state spaces that are *standard Borel*; as noted earlier, these are Borel subsets of complete, separable and metric spaces. We note again that the spaces that are complete, separable and metric are also called *Polish metric* spaces.

Let $\{x_t, t\in\mathbb{Z}_+\}$ be a Markov chain with a Polish state space $\mathbb{X}$, and defined on a probability space $(\Omega,\mathcal{F},\mathbf{P})$, where Ω is the sample space, $\mathcal{F}$ a sigma field of subsets of Ω, and $\mathbf{P}$ a probability measure. Let $P(x,D):=\mathbf{P}(x_{t+1}\in D|x_t=x)$ denote the transition probability from x to D, that is the probability of the event $\{x_{t+1}\in D\}$ given that $x_t=x$.

We could compute $P(x_{t+k}\in D|x_t=x)$ inductively as follows:

$$\mathbf{P}(x_{t+k}\in D|x_t=x)=\int\cdots\int\int P(x_t,dx_{t+1})\dots P(x_{t+k-2},dx_{t+k-1})P(x_{t+k-1},D)$$

As such, we have for all $n\geq 1$, states x and Borel sets A, $P^n(x,A):=\mathbf{P}(x_{t+n}\in A|x_t=x)=\int_X P^{n-1}(x,dy)P(y,A)$, with $P^1(\cdot,\cdot):=P(\cdot,\cdot)$.

Definition 3.2.1 *A Markov chain is μ-irreducible, if for any set $B\in\mathcal{B}(\mathbb{X})$ such that $\mu(B)>0$, and any $x\in\mathbb{X}$, there exists some integer $n>0$ (possibly depending on B and x), such that $P^n(x,B)>0$, where $P^n(x,B)$ is the transition probability in n stages, that is, $P(x_{t+n}\in B|x_t=x)$.*

A maximal irreducibility measure ψ is an irreducibility measure such that for all other irreducibility measures ϕ, we have $\psi(B)=0\Rightarrow\phi(B)=0$ for any $B\in\mathcal{B}(\mathcal{X})$ (that is, all other irreducibility measures are absolutely continuous with respect to ψ). In the text, whenever a chain is said to be irreducible, irreducibility with respect to a maximal irreducibility measure is implied. We also define $\mathcal{B}^+(\mathcal{X})=\{A\in\mathcal{B}(\mathcal{X}):\psi(A)>0\}$ where ψ is a maximal irreducibility measure. A maximal irreducibility measure ψ exists for a μ-irreducible Markov chain, for example $\psi(B)=\sum_{n\in\mathbb{Z}_+}2^{-n}P^n(x,B)\mu(dx)$ (see [233, Propostion 4.2.2]).

As an example, consider the following linear system:

$$x_{t+1}=ax_t+w_t,$$

This chain is Lebesgue irreducible if w_t is a Gaussian variable. The definitions for recurrence and transience follow those in the countable state space setting:

Definition 3.2.2 *A set $A\in\mathcal{B}(\mathbb{X})$ is called recurrent if*

$$E_x[\sum_{t=1}^{\infty}1_{x_t\in A}]=\sum_{t=1}^{\infty}P^t(x,A)=\infty,\qquad\forall x\in\mathbb{A}.$$

A $\psi-$irreducible Markov chain is called recurrent if, for A with $\psi(A)>0$,

$$E_x[\sum_{t=1}^{\infty}1_{x_t\in A}]=\sum_{t=1}^{\infty}P^t(x,A)=\infty,\qquad\forall x\in\mathbb{X}.$$

Definition 3.2.3 *A set $A\in\mathcal{B}(\mathbb{X})$ is Harris recurrent if*

$$P_x(\eta_A = \infty) = 1, \qquad \forall x \in A. \tag{3.19}$$

A ψ-irreducible Markov chain is Harris recurrent if

$$P_x(\eta_A = \infty) = 1, \qquad A \in \mathcal{B}(\mathbb{X}), \psi(A) > 0, \qquad \forall x \in \mathbb{X}.$$

Theorem 3.2.1 *Harris recurrence of a set A is equivalent to*

$$P_x(\tau_A < \infty) = 1, \qquad \forall x \in A.$$

Proof. Let $\tau_A(1)$ be the first time the state hits A. Now, with $x \in A$, the probability of $\tau_A(2) < \infty$ can be computed recursively as

$$\begin{aligned}
&P_x(\tau_A(2) < \infty) = P_x(\tau_A(2) < \infty, \tau_A(1) < \infty) \\
&= \sum_{m \in \mathbb{Z}_+} P_x(\tau_A(2) < \infty, \tau_A(1) = m) \\
&= \sum_{m \in \mathbb{Z}_+} P_x(\tau_A(2) < \infty, X_{\tau_A(1)} \in A, \tau_A(1) = m) \\
&= \sum_{m \in \mathbb{Z}_+} \int_A P_x(\tau_A(2) < \infty, X_{\tau_A(1)} \in dy, \tau_A(1) = m) \\
&= \sum_{m \in \mathbb{Z}_+} P_x(\tau_A(2) < \infty | X_{\tau_A(1)} = y, \tau_A(1) = m) P_x(X_{\tau_A(1)} \in dy, \tau_A(1) = m) \\
&= \sum_{m \in \mathbb{Z}_+} \int_A P(\tau_A(1) < \infty | X_0 = y) P_x(X_{\tau_A(1)} \in dy, \tau_A(1) = m) \qquad (3.20) \\
&= \sum_{m \in \mathbb{Z}_+} \int_A P_x(X_{\tau_A(1)} \in dy, \tau_A(1) = m) \\
&= \sum_{m \in \mathbb{Z}_+} P_x(X_{\tau_A(1)} \in A, \tau_A(1) = m) \\
&= \sum_{m \in \mathbb{Z}_+} P_x(\tau_A(1) = m) = 1,
\end{aligned}$$

where (3.20) uses the strong Markov property and the next equation follows from $P(\tau_A(1) < \infty | X_0 = y) = 1$ for $y \in A$. By induction, for every $n \in \mathbb{N}$

$$P_x(\tau_A(n+1) < \infty) = 1 \tag{3.21}$$

Now,

$$P_x(\eta_A \geq k) = P_x(\tau_A(k-1) < \infty),$$

since k times visiting a set requires k times returning to a set, when the initial state x is in the set. As such,

$$P_x(\eta_A \geq k) = 1, \forall k \in \mathbb{Z}_+$$

is identically equal to 1. Define $B_k = \{\omega \in \Omega : \eta(\omega) \geq k\}$, and it follows that $B_{k+1} \subset B_k$ for all $k \in \mathbb{N}$. By the continuity of probability (see Theorem B.1.2), $P(\bigcap_{k=1}^{\infty} B_k) = \lim_{k \to \infty} P(B_k)$, it follows that $P_x(\eta_A = \infty) = 1$. The other direction for equivalence follows from the definitions of occupation time η_A and return time τ_A. ◇

Definition 3.2.4 *For a Markov chain with transition probability P, a probability measure π is invariant if*

$$\pi(D) = \int_{\mathbb{X}} P(x, D) \pi(dx), \qquad D \in \mathcal{B}(\mathbb{X}).$$

Definition 3.2.5 *If a Harris recurrent Markov chain admits an invariant probability measure, then the chain is called positive Harris recurrent.*

We will discuss the ergodic theorem for such chains further, but it may be useful to state the following:

Lemma 3.2.1 (Ergodic Theorem for Positive Harris Recurrent Markov Chains, *MeynBook] For a Markov chain* $\{X_n\}_{n\in\mathbb{N}}$ *which admits at least one invariant probability measure, the following statements are equivalent:*

(i) The chain is positive Harris recurrent.

(ii) There exists an invariant probability measure π *of such that for all* $f \in L_1(\pi)$ *and every initial distribution* μ*,*

$$\mathbb{P}\left(\lim_{n\to\infty}\frac{1}{n+1}\sum_{i=0}^{n} f(x_i) = \int_{\mathbb{X}} f(x)\pi(dx)\right) = 1,$$

where $\{x_i, i \in \mathbb{Z}_+\}$ *is Markov with* $x_0 \sim \mu$*.*

3.2.1 Invariant probability measures and split chains

Uncountable chains act like countable ones when there is a single *atom* $\alpha \subset \mathbb{X}$ which satisfies a finite mean return property to be discussed below.

Definition 3.2.6 *A set* α *is called an atom if there exists a probability measure* ν *such that*

$$P(x, A) = \nu(A), \quad \forall x \in \alpha, \forall A \in \mathcal{B}(\mathbb{X}).$$

If the chain is ψ*-irreducible and* $\psi(\alpha) > 0$*, then* α *is called an accessible atom.*

In case there is an accessible atom α, we have the following result the proof of which follows the same steps of those of Theorem 3.1.3 and 3.1.5.

Theorem 3.2.2 *For a* ψ*-irreducible Markov chain for which* $E_\alpha[\tau_\alpha] < \infty$*, the following is the invariant probability measure:*

$$\pi(A) = \frac{E_\alpha\left[\sum_{k=0}^{\tau_\alpha - 1} 1_{\{x_k \in A\}}\right]}{E_\alpha[\tau_\alpha]}, \qquad A \in \mathcal{B}(X)$$

Small Sets and Nummelin and Athreya-Ney's Splitting Technique

In case an atom is not present, we may be able to construct an *artificial atom*:

Definition 3.2.7 *A set* $A \in \mathcal{B}(\mathbb{X})$ *is (*n*-*μ*)-small on* $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ *if for some non-trivial positive (i.e., not all sets have zero measure) measure* μ *and* $n \in \mathbb{N}$

$$P^n(x, B) \geq \mu(B), \qquad \forall x \in A, \text{ and } B \in \mathcal{B}(\mathbb{X})$$

Often, we simply say that a set is small without specifying the smallness measure μ or time index n.

The results on recurrence apply to uncountable chains with no atom provided there is a small set or a petite set (to be discussed further below). In the following, we construct an artificial atom through what is commonly known as the splitting technique, see [246] [247] (see also [20]).

Suppose a set A is 1-small. Define a process $z_t = (x_t, a_t)$, where z_t is $\in$X $\times\{0,1\}$-valued and $\{a_t\}$ is i.i.d. Bernoulli with $P(a_t = 1) = \delta$. That is, we enlarge the state space and observe that z_t is also Markov. When $x_t \notin A$, $\{a_t\}, \{x_t\}$ evolve independently from each other. When $x_t \in A$, depending on the realization of a_t, with probability δ the state is mapped to $A \times \{1\}$ and with probability $1-\delta$ the state is mapped to $A \times \{0\}$. From $A \times \{1\}$, the transition for the next time stage is given by $\frac{\nu(dx_t)}{\delta}$ (for all $(x,1) \in A \times \{1\}$), and from $A \times \{0\}$, it visits the future time stage with transition probability

$$\frac{P(dx_{t+1}|x_t) - \nu(dx_{t+1})}{1-\delta},$$

where $\delta = \nu(\mathbb{X})$. That is,

$$\begin{aligned} P(x_{t+1\in B}|(x_t,a_t) = (x,1)) &= \frac{\nu(B)}{\delta}, \quad (x,1) \in A \times \{1\} \\ P(x_{t+1\in B}|(x_t,a_t) = (x,0)) &= \frac{P(x,B) - \nu(B)}{1-\delta}, \quad (x,0) \in A \times \{0\} \end{aligned} \tag{3.22}$$

In this case, $A \times \{1\}$ is an accessible atom for the extended (split) Markov chain (x_t, a_t), and the marginal distribution of the original Markov process $\{x_t\}$ has not been altered by this construction.

The following can be established using the construction above.

Proposition 3.2.1 *If*

$$\sup_{x\in A} E[\min(t > 0 : x_t \in A)|x_0 = x] < \infty$$

then,

$$\sup_{z\in(A\times\{1\})} E[\min(t > 0 : z_t \in (A \times \{1\}))|z_0 = z] < \infty.$$

Now suppose that a set A is m-small. Then, we can construct a split chain for the sampled process $x_{mn}, n \in \mathbb{N}$. Note that this sampled chain has a transition kernel as P^m. We replace the discussion for the 1-small case with the sampled chain (also known as the m-skeleton of the original chain). If one can show that the sampled chain has an invariant measure π_m, then (see [233, Theorem 10.4.5]):

$$\pi(B) := \frac{1}{m}\sum_{k=0}^{m-1} \int \pi_m(dx) P^k(x,B) \tag{3.23}$$

is invariant for P. Furthermore, π is also invariant for the sampled chain with kernel P^m. Hence if P^m leads to a unique invariant probability measure, $\pi = \pi_m$.

From small to petite sets

Definition 3.2.8 *[233] A set $A \in \mathcal{B}(\mathbb{X})$ is $\nu_{\mathcal{R}}$-petite on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ if for some distribution $\mathcal{R}$ on $\mathbb{N}$, and some non-trivial positive measure $\nu_{\mathcal{R}}$,*

$$\sum_{n=0}^{\infty} P^n(x,B)\mathcal{R}(n) \geq \nu_{\mathcal{R}}(B), \qquad \forall x \in A, \text{ and } B \in \mathcal{B}(\mathbb{X}).$$

By [233, Proposition 5.5.6], if a Markov chain is ψ-irreducible and if a set C is ν-petite, then $\mathcal{R}$ can be taken to be a geometric distribution $a_\epsilon(i) = (1-\epsilon)\epsilon^i, \quad i \in \mathbb{Z}_+$ (with the randomly sampled chain also known as the resolvent kernel).

Another useful result to be utilized later is as follows.

Theorem 3.2.3 *[233] Let $\{x_t\}$ be μ-irreducible and let A be ν-petite. Then, there exists a sampling distribution such that A is ψ-petite where ψ is a maximal irreducibility measure. Furthermore, A is ψ-petite for a sampling distribution with finite mean.*

Definition 3.2.9 *A ψ-irreducible Markov chain is periodic with period d if there exists a partition of $\mathbb{X} = \cup_{i=1}^{d} \mathbb{X}_i \cup D$ so that $P(x, \mathbb{X}_{i+1}) = 1$ for all $x \in \mathbb{X}_i$ and $P(x, \mathbb{X}_1) = 1$ for all $x \in \mathbb{X}_d$, with $\psi(D) = 0$. If no such $d > 1$ exists, the chain is aperiodic.*

Another useful result is the following.

Theorem 3.2.4 *[233, Theorem 5.5.3] For an aperiodic and irreducible Markov chain $\{x_t\}$ every petite set is ν-small for some appropriate ν (but now ν may not be a maximal irreducibility mesure; compare with Theorem 3.2.3).*

The discussion up to (3.23) and the split chain argument applies also for an arbitrary sampling distribution $\mathcal{K}$ on $\mathbb{N}$. Suppose that we have

$$\int \pi_{\mathcal{K}}(dx) \Big(\sum_n \mathcal{K}(n) P^n(x, B) \Big) = \pi_{\mathcal{K}}(B), \quad B \in \mathcal{B}(\mathbb{X}) \tag{3.24}$$

Then,

$$\pi(B) := \int \sum_m \mathcal{K}(m) \sum_{k=0}^{m-1} \pi_{\mathcal{K}}(dx) P^k(x, B) \tag{3.25}$$

is an invariant measure for the original chain so that $\pi = \pi P$. By normalizing this measure, we obtain an invariant probability measure for the original chain, provided that $\sum_n n\mathcal{K}(n) < \infty$ (see Theorem 3.2.3).

Exercise 3.2.1 *Show that (3.25) is an invariant probability measure given that (3.24) holds.*

3.2.2 Existence of an invariant probability measure

We state the following very consequential results on the existence of invariant probability measures for Markov chains.

Theorem 3.2.5 *Consider an aperiodic and irreducible Markov chain $\{x_t\}$. If there exists a set A which is also an m-small set for some $m \in \mathbb{Z}_+$, and if the set satisfies*

$$\sup_{x \in A} E[\min(t > 0 : x_t \in A) | x_0 = x] < \infty,$$

then the Markov chain admits an invariant probability measure.

Note that for $m = 1$, we don't need irreducibility or aperiodicity, by directly following the splitting construction presented. In the following, we relax aperiodicity, in any case.

Theorem 3.2.6 *(Meyn-Tweedie) Consider a Harris recurrent Markov chain $\{x_t\}$. If there exists a μ-petite set A for some positive measure μ, and if the set satisfies*

$$\sup_{x \in A} E[\min(t > 0 : x_t \in A) | x_0 = x] < \infty,$$

then the Markov chain is positive Harris recurrent (and admits a unique invariant probability measure).

Remark 3.5. For the m-small case with $m > 1$, in view of the splitting construction, one question is whether

$$\sup_{x \in A} E[\min(mt > 0 : x_{mt} \in A) | x_0 = x] < \infty,$$

or in the petite case with a sampled chain with geometrically sampled times τ_k, whether

$$\sup_{x\in A} E[\min(\tau_k > 0 : x_{\tau_k} \in A)|x_0 = x] < \infty,$$

is implied by

$$\sup_{x\in A} E[\min(t > 0 : x_t \in A)|x_0 = x] < \infty.$$

For the small set case, under irreducibility and aperiodicity, the above holds. See Remark 4.4 on the positive Harris recurrence discussion for an m-skeleton and split chains: When a Markov chain has an invariant probability measure, the sampled chain (m-skeleton) also satisfies a *drift condition*, which then leads to the result that an *atom* constructed through an m-skeleton has a finite return property, which can be used to establish the existence of an invariant probability measure.

However, the discussion for the petite set case is more direct and can be arrived at via the properties of a geometrically sampled chain (following along the arguments in Exercise 3.5.4) In particular, an m-small set is 1-small for a geometrically sampled chain (since $\sum_{n=0}^{\infty} P^n(x, C)\mathcal{R}(n) \geq \mathcal{R}(m)P^m(x, C)$), which then allows for the arguments for existence to be applicable more directly (see also [109]). The utilization of petite sets, via a sampled chain, thus allows for relaxing the aperiodicity requirement and also with a more direct argument as discussed above.

In this case, the invariant measure satisfies the following, which is a generalization of Kac's Lemma [116]:

Theorem 3.2.7 *For a μ-irreducible Markov chain with invariant probability measure π, the following holds:*

$$\pi(A) = \int_C \pi(dx)E_x[\sum_{k=0}^{\tau_C-1} 1_{\{x_k\in A\}}], \quad \forall A \in \mathcal{B}(X), \mu(A) > 0, \pi(C) > 0$$

Observe that with the above, by taking $A = \mathbb{X}$ we have that for any Borel C,

$$\inf_{x\in C} E_x[\tau_C] \leq \frac{1}{\pi(C)} \leq \sup_{x\in C} E_x[\tau_C]$$

3.2.3 On small and petite sets: sufficient conditions (Optional)

Establishing the smallness or petiteness of a set may be difficult to directly verify. In the following, we present a few conditions that may be used to establish petiteness properties.

T-chains. By [233, p. 131], for a Markov chain with transition kernel P and $\mathcal{K}$ a probability measure on natural numbers, if there exists for every $E \in \mathcal{B}(\mathbb{X})$, a lower semi-continuous function $\mathcal{N}(\cdot, E)$ such that $\sum_{n=0}^{\infty} P^n(x, E)\mathcal{K}(n) \geq \mathcal{N}(x, E)$ for a sub-stochastic kernel $\mathcal{N}(\cdot, \cdot)$ with $\mathcal{N}(x, \mathbb{X}) > 0$ for all $x \in \mathbb{X}$, the chain is called a $T-$chain.

Theorem 3.2.8 *[233, Theorem 6.2.5] For a $T-$chain which is irreducible, every compact set S is petite.*

Proof Sketch. We will prove the result with the stronger assumption $P(x, A)$ is continuous in x for every Borel A (that is, P is strong Feller). Note that this implies that $\sum_{n=0}^{\infty} P^n(x, B)\mathcal{K}(n)$ is continuous for every sampling distribution $\mathcal{K}$, by the dominated convergence theorem. Due to irreducibility, by Theorem 3.2.9, a petite set B exists so that with a positive ν measure, for every Borel C, we have

$$\sum_{n=0}^{\infty} P^n(x, C)\mathcal{R}(n) \geq \nu(C), \quad \forall x \in B.$$

Now there exists $\mathcal{K}$ such that $\sum_{n=0}^{\infty} P^n(x, B)\mathcal{K}(n)$ puts a positive measure on B for every $x \in \mathbb{X}$, due to irreducibility and that B has positive measure under the irreducibility measure. Since

$$\sum_{n=0}^{\infty} P^n(x, B)\mathcal{K}(n)$$

is continuous, there exists $x^* \in S$ such that the minimum $\sum_{n=0}^{\infty} P^n(x,B)\mathcal{K}(n)$ over $x \in S$ is attained. The desired petiteness result then comes from bounding, for any Borel C, for an appropriate probability measure η:

$$\sum_r \eta(r)P^r(x,C) \geq \sum \mathcal{K}(n)\Big(\int_B P^n(x^*,dy)\sum_m \mathcal{R}(m)P^m(y,C)\Big)$$

for every $x \in S$. ◇

A reflection on the proof of the result above, via Lusin's theorem (see Theorem D.5.1), leads to the following: Small or petite sets exist for irreducible Markov chains.

Theorem 3.2.9 *[233, Thm 5.2.2] Let $\{x_t\}$ be μ-irreducible. Then, for every Borel B with $\mu(B) > 0$, there exists $m \geq 1$ and a ν_m-small set $C \subset B$ with $\mu(C) > 0$ and $\nu_m(C) > 0$.*

For a countable state space, under irreducibility, every finite set S is petite.

Tweedie's uniform countable additivity condition. Tweedie [310] considers the following. If S is such that the following *uniform countable additivity* condition

$$\lim_{n\to\infty} \sup_{x\in S} P(x,B_n) = 0, \tag{3.26}$$

is satisfied for $B_n \downarrow \emptyset$, then S is petite (and for example, (4.10) to be studied in *Chapter* 4 implies the existence of an invariant probability measure). In this case, there exists at most finitely many invariant probability measures. By [233, Proposition 5.5.5 (iii)], under irreducibility, the Harris recurrent component of the space can be expressed as a countable union of petite sets C_n with $\cup_{n=1}^{\infty} C_n$, with $\cup_m^{\infty} C_m \to \emptyset$ as $m \to \infty$. By Lemma 4 of Tweedie (2001), under uniform countable additivity, any set $\cup_{i=1}^{M} C_i$ is uniformly accessible from S. Therefore, if the Markov chain is irreducible, the condition (3.26) implies that the set S is petite. This may be easier to verify for a large class of applications. Under further conditions (such as if S is compact and V used in a drift criterion (4.10) has compact level sets), then the analysis will lead sufficient conditions leading to (3.26). In particular, [310, Lemma 1] notes that if S is bounded and V is continuous (and thus uniformly bounded on S), it suffices to test (3.26) only for B_n sets inside sets on which V is bounded (that is with B_1 such that $\sup_{x\in B_1} V(x) < \infty$.). In applications, this is often much easier to apply, see e.g. [350].

A further condition. We have the following complementary condition, where irreducibility can be relaxed, but the strong Feller property is imposed.

Proposition 3.6. *[16] Assume that*

(i) The transition kernel $\mathcal{T}$ is bounded from below by a conditional probability measure that admits a density with respect to some positive measure ϕ. In other words there exist a measurable $f\colon \mathbb{X}\times\mathbb{U}\times\mathbb{X}\to\mathbb{R}_+$, such that

$$P(x,D) \;\geq\; \int_D f(x,y)\phi(dy)$$

for every $D \in \mathcal{B}(\mathbb{X})$.

(ii) The function $f(x,y)$ is continuous in x for every fixed y.

(iii)It holds that

$$\int_{\mathbb{X}} (\inf_{x\in A}\; f(x,y))\phi(dy) \;>\; 0$$

for every nonempty compact set $A \subset \mathbb{X}$.

Then, every compact set is 1-small.

Proof. The measurable selection results in Appendix C (see [202, 283] and [169, Theorem 2]) show that, for any compact $A \subset \mathbb{X}$, there exist measurable functions g and F such that

$$\inf_{x\in A} f(x,y) \;=\; \min_{x\in A} f(x,y) \;=:\; F\big(g(y),y\big) \tag{3.27}$$

Thus, we have for all $x \in A$

$$\begin{aligned} P(x,D) \;&\geq\; \int_D \inf_{x\in A} f(x,y)\phi(\mathrm{d}y) \\ &=\; \int_D F(g(y),y)\phi(dy) \;=:\; \nu(D) \end{aligned} \tag{3.28}$$

for some finite (sub-probability) measure ν. Thus, every compact set is 1-small. ◇

3.2.4 Rates of convergence to equilibrium

We can extend Dobrushin's contraction result for the uncountable state space case. In this general setup, we define the Dobruhsin coefficient for a Markov chain with transition kernel P as

$$\delta(P) = \inf_{(x,y);\mathcal{A}_n} \sum_{i=1}^{n} \min\{P(x,A_i), P(y,A_i)\} \tag{3.29}$$

where the infimum is over all $x, y \in \mathbb{X}$ and all finite partitions $\mathcal{A}_n := \{A_i^n, i = 1, \cdots, n\}$ consisting of disjoint sets whose union is $\mathbb{X}$. Note that this definition holds for both continuous or countable $\mathbb{X}$. We then have for two probability measures π, π' (see Appendix D for a review on probability measures) [106]

$$\|\pi P - \pi' P\|_{TV} \leq (1 - \delta(P))\|\pi - \pi'\|_{TV}.$$

As such, if $\delta(P) > 0$, the iterations $\pi_t = \pi_{t-1}P$ converge to a unique fixed point geometrically fast. To better appreciate this coefficient, first note that by the property that $|a - b| = a + b - 2\min(a, b)$, the Dobrushin's coefficient in (3.13) can be written as (for the countable state space case):

$$\delta(P) = 1 - \frac{1}{2}\max_{i,k}\sum_j |(P(i,j) - P(k,j)|$$

For the continuous setup, in case $P(x, dy)$ is the transition kernel admitting a density for each x (that is $P(x, A) = \int_A p(x, y)dy$ with probability density function $p(x, \cdot)$), the expression

$$\delta(P) = 1 - \frac{1}{2}\sup_{x,z}\int_{\mathbb{R}} |p(x,y) - p(z,y)|dy,$$

is the Dobrushin's ergodic coefficient for $\mathbb{R}-$valued Markov processes.

The versatility of using Dobrushin's coefficient for establishing rates of convergence manifests itself in the following conditions (noted from [163, Theorem 3.2]).

Theorem 3.2.10 *Consider the following conditions.*

(i) There exists a state $x^* \in \mathbb{X}$ and a number $\beta > 0$ such that $P(\{x^*\}|x) \geq \beta$ for all $x \in \mathbb{X}$.

(ii) There exist $n \in \mathbb{N}$ and a non-trivial (positive)measure μ such that $P^n(\cdot|x) \geq \mu(\cdot)$ for all $x \in \mathbb{X}$.

(iii)There exist $n \in \mathbb{N}$ and a positive number $\beta < 1$ so that for all $x, x' \in \mathbb{X}$

$$\|P^n(\cdot|x) - P^n(\cdot|x')\|_{TV} \leq 2\beta.$$

(iv)There exist $c > 0, \beta \in (0, 1)$ such that there is a probability measure π with

$$\|\pi_0 P^n - \pi\| \leq c\beta^n, \qquad \pi_0 \in \mathcal{P}(\mathbb{X}), n \in \mathbb{N}$$

We have that

$$(i) \Rightarrow (ii) \Leftrightarrow (iii) \Rightarrow (iv)$$

Note that condition (ii) amounts to the entire state space being n-small. The results above can be established through an analysis based on Dobrushin's ergodic coefficient. In the next chapter, we will provide more relaxed conditions leading to rates of convergence, even though those conditions will not lead to a *uniform* (over $x \in \mathbb{X}$) rate of convergence.

3.3 Further Conditions on the Existence and Uniqueness of Invariant Probability Measures

3.3.1 Further conditions on existence of invariant probability measures

Markov chains with the Feller property

This section uses certain properties of spaces of probability measures, reviewed in Section D.

Definition 3.3.1 *(i) A Markov chain is weak Feller if $\int_{\mathbb{X}} P(x, dz)v(z)$ is continuous in x for every continuous and bounded v on $\mathbb{X}$.*

(ii) If the above holds (i.e., $\int_{\mathbb{X}} P(x, dz)v(z)$ is continuous in x) for every bounded measurable v, the Markov chain is called strong Feller.

Example 3.7. (i) Let $x_{t+1} = f(x_t) + w_t$, where $f : \mathbb{R} \to \mathbb{R}$ is continuous and $\{w_t\}$ is an i.i.d. real valued noise sequence. In this case $\{x_t\}$ is weak Feller, regardless of the random variable w_t.

(ii) The chain $\{x_t\}$ is strong Feller if $f : \mathbb{R} \to \mathbb{R}$ is continuous and w_t is an i.i.d. random sequence where w_t admits a continuous probability density function.

See Section 5.6 for further examples and discussions (while those examples involve controlled models, you can assume that the control term is a singleton in the context of the discussion in the current chapter).

Theorem 3.3.1 *Let $\{x_t\}$ be a weak Feller Markov process living in a compact subset of a complete, separable metric space. Then $\{x_t\}$ admits an invariant probability measure.*

Proof. Proof follows the observation that the space of probability measures on a compact set is tight (that is, it is weakly sequentially pre-compact), see *Appendix* D for a discussion on weak convergence. Consider a sequence

$$\mu_T = \frac{1}{T} \sum_{t=0}^{T-1} \mu_0 P^t, \qquad T \geq 1,$$

There exists a subsequence μ_{T_k} which converges weakly to some μ^*. It follows that for every continuous and bounded function f

$$\langle \mu_{T_k}, f \rangle := \int \mu_{T_k}(dx) f(x) \to \langle \mu^*, f \rangle$$

Likewise, since $Pf(x) = \int f(x_1) P(dx_1 | x_0 = x)$ is continuous in x (by the weak Feller condition), it follows that

$$\langle \mu_{T_k}, Pf \rangle := \int \mu_{T_k}(dx) \left(\int P(x, dy) f(y) \right) \to \langle \mu^*, Pf \rangle.$$

Now,

$$(\mu_{T_k} - \mu_{T_k}P)(f) = \frac{1}{T_k}E_{\mu_0}\left[\sum_{k=0}^{T_k-1} P_x^k f - \sum_{k=0}^{T_k-1} P_x^{k+1} f\right]$$
$$= \frac{1}{T_k}E_{\mu_0}\Big[f(x_0) - f(x_{T_k})\Big] \to 0. \tag{3.30}$$

Thus,

$$(\mu_{T_k} - \mu_{T_k}P)(f) = \langle \mu_{T_k}, f\rangle - \langle \mu_{T_k}P, f\rangle = \langle \mu_{T_k}, f\rangle - \langle \mu_{T_k}, Pf\rangle \to \langle \mu^* - \mu^* P, f\rangle = 0.$$

Now, if the relation $\langle \mu^* - \mu^* P, f\rangle = 0$ holds for every continuous and bounded function, it also holds for any measurable function f: This is because continuous functions are dense in measurable functions under the supremum norm (in other words, continuous and bounded functions form a *separating class* for the space of probability measures, see e.g. p. 13 in [42] or Theorem 3.4.5 in [120])). Thus, μ^* is an invariant probability measure. ◇

Remark 3.8. The theorem applies identically if instead of a compact set assumption, one assumes that the sequence μ_k takes values in a weakly compact set (e.g. via Prohorov's theorem [111]); that is, if the sequence admits a weakly converging subsequence.

Remark 3.9. Reference [213] gives the following example to emphasize the importance of the Feller property: Consider a Markov chain evolving in $[0, 1]$ given by: $P(x, x/2) = 1$ for all $x \neq 0$ and $P(0, 1) = 1$. This chain does not admit an invariant measure. This can be established by a continuity of probability argument for any invariant probability measure (if one existed) on the absorbing sets $(0, \delta)$ for any $\delta > 0$.

In the following, we generalize the above result to a case where the state space $\mathbb{X}$ is not compact, but is *locally compact*.

Theorem 3.3.2 *Let $\{x_t\}$ be a weak Feller Markov process taking values from a locally compact $\mathbb{X}$. Suppose further for some initial probability measure μ_0, with*

$$\mu_T = \frac{1}{T}\sum_{t=0}^{T-1} \mu_0 P^t, \qquad T \geq 1,$$

we have that for some compact B

$$\liminf_{T\to\infty} \mu_T(B) > 0.$$

Then, $\{x_t\}$ admits an invariant probability measure.

Proof. The proof builds on an application of the Banach-Alaoglu theorem; the space of signed measures under the total variation norm with finite total variation is the topological dual of the space of continuous functions which vanish at infinity, and the unit ball in this space is weak*-compact by the Banach-Alaoglu theorem. By an argument similar to the proof of Theorem 3.3.1, then there exists a subsequence μ_{T_k} which converges in the weak* sense to a limit μ^*. Since μ^* cannot be the trivial (all-zero) measure, μ^* must be invariant and positive. Normalizing this measure implies that there exists an invariant probability measure. ◇

Quasi-Feller chains

Often, one does not have the Feller property, but the set of discontinuity is appropriately negligibe.

Assumption 3.3.1 *For f bounded and continuous, $Pf(x) := E[f(X_{t+1})|X_t = x]$) is continuous on $\mathbb{X} \setminus D$ where D is a closed set with $P(X_{t+1} \in D|x) = 0$ for all x. Furthermore, with $D_\epsilon = \{z : d(z, D) < \epsilon\}$ for $\epsilon > 0$ and d the metric on $\mathbb{X}$, for some $K < \infty$, we have that for all x and $\epsilon > 0$*

$$P\Big(X_{t+1} \in D_\epsilon | x_t = x\Big) \leq K\epsilon.$$

Theorem 3.3.3 *Suppose that Assumption 3.3.1 holds. If the state space is compact, there exists an invariant probability measure for the Markov chain.*

Proof. The sequence of expected empirical probability measures

$$v_n(A) = E_x[\frac{1}{n}\sum_{k=0}^{n-1} 1_{\{X_k \in A\}}]$$

is tight, and thus there exists a weakly converging subsequence. Assumption 3.3.1 implies that every converging subsequence v_{n_k} of is such that for all $\epsilon > 0$

$$\limsup_{n_k \to \infty} v_{n_k}(D_\epsilon) \leq K\epsilon.$$

Note that with $v = \lim_{n_k \to \infty} v_{n_k}$, it follows from the Portmanteau theorem (see e.g. [111, Thm.11.1.1]) that

$$v(D_\epsilon) \leq K\epsilon.$$

Now, consider a weakly converging empirical occupation sequence v_{t_k} and let this sequence have an accumulation point v^*. We will show that v^* is invariant.

Observe that the transitioned probability measure $v_{t_k}P$ satisfies the following for every continuous and bounded f: Consider $\langle v_{t_k}, Pf\rangle = \langle v_{t_k}, g_f\rangle + \langle v_{t_k}, Pf - g_f\rangle$, where g_f is a continuous function which is equal to Pf outside an open neighborhood of D and is continuous with $\|g_f\|_\infty = \|\mathbb{P}f\|_\infty \leq \|f\|_\infty$. The existence of such a function follows from the Tietze-Urysohn extension theorem [111], where the closed set is given by $\mathbb{X} \setminus D_\epsilon$. It then follows from Assumption 3.3.1 that, for every $\epsilon > 0$ a corresponding g_f can be found so that $\langle v_{t_k}, Pf - g_f\rangle \leq K\|f\|_\infty\epsilon$, and since $\langle v_{t_k}, g_f\rangle \to \langle v^*, g_f\rangle$, it follows that

$$\begin{aligned}
&\limsup_{t_k \to \infty} |\langle v_{t_k}, Pf\rangle - \langle v^*, Pf\rangle| \\
&= \limsup_{t_k \to \infty} |\langle v_{t_k}, Pf - g_f\rangle - \langle v^*, Pf - g_f\rangle| \\
&\leq \limsup_{t_k \to \infty} |\langle v_{t_k}, Pf - g_f\rangle| + |\langle v^*, Pf - g_f\rangle| \\
&\leq 2K'\epsilon
\end{aligned} \tag{3.31}$$

Here, $K' = 2K\|f\|_\infty$ is fixed and ϵ may be made arbitrarily small. We conclude that v^* is invariant. ◇

Remark 3.10. In his definition for quasi-Feller chains, Lasserre assumes the state space to be locally compact. In the proof above [344] tightness is invoked directly with no use of convergence properties of the set of functions which decay to zero as is done in [167]; for a related result see Gersho [142].

Cases without the Feller condition

One can relax the weak Feller condition and instead consider spaces of probability measures which are setwise sequentially pre-compact. The proof of this result follows from a similar observation as (3.30) but with weak convergence replaced by *setwise* convergence (see Appendix D). Note that in this case, if $\mu_{T_k} \to \mu^*$ setwise, it follows that $\mu_{T_k}P(f) \to \mu^*P(f)$ and thus μ^* is invariant. It can be shown (as in the proof of Theorem 3.3.1) that a (sub)sequence of occupation measures which converges setwise, converges to an invariant probability measure. A sufficient condition for a sequence of probability measures to be setwise sequentially compact is that there exists a *finite* measure π such that $v_k \leq \pi$ for all $k \in \mathbb{N}$ [168].

As an example, consider a system of the form:

$$x_{t+1} = f(x_t) + w_t \tag{3.32}$$

where w_t admits a distribution with a bounded density function, which is positive everywhere and f is bounded. This system admits an invariant probability measure which is unique.

3.3.2 Uniqueness of an invariant probability measure

Unique ergodicity properties

For a Markov chain, the uniqueness of an invariant probability measure implies the ergodicity of the measure; such a Markov chain is often referred to as *uniquely ergodic*.

The following definition will be useful.

Definition 3.11. *Let π be a probability measure on $\mathbb{X}$ with metric d. The topological support of π is defined with*

$$\operatorname{supp} \pi := \{x : \pi(B_r(x)) > 0\}, \quad \forall r > 0,$$

where $B_r(x) = \{y \in \mathbb{X} : d(x, y) < r\}$.

Theorem 3.3.4 *Let $\{x_t\}$ be a ψ-irreducible Markov chain which admits an invariant probability measure. The invariant measure is unique.*

Proof. Let there be two invariant probability measures μ_1 and μ_2. Then, there exists two *mutually singular* invariant probability measures ν_1 and ν_2, that is $\nu_1(B_1) = 1$ and $\nu_2(B_2) = 1$, $B_1 \cap B_2 = \emptyset$ and that $P^n(x, B_1^C) = 0$ for all $x \in B_1$ and $n \in \mathbb{Z}_+$ and likewise $P^n(z, B_1^C) = 0$ for all $z \in B_1$ and $n \in \mathbb{Z}_+$. This then implies that the irreducibility measure has zero support on B_1^C and zero support on B_2^C and thus on $\mathbb{X}$, leading to a contradiction. ◇

A further result on uniqueness is given next.

Definition 3.12. *For a Markov chain with transition kernel P, a point x is accessible (or reachable) if for every y and every open neighborhood O of x, there exists $k > 0$ such that $P^k(y, O) > 0$.*

One can show that if a point is accessible, it belongs to the (topological) support of every invariant measure (see, e.g., Lemma 2.2 in [154]). The support (or spectrum) of a probability measure is defined to be the set of all points x for which every open neighbourhood of x has positive measure. A Markov chain V_t is said to have the strong Feller property at x if $E[f(X_{t+1})|X_t = x)$ is continuous at x for every measurable and bounded f.

Theorem 3.3.5 *[154] [259] If a Markov chain has the strong Feller property at an accessible point, then the chain can have at most one invariant probability measure.*

Proof. Let there be two invariant probability measures μ_1 and μ_2. Then, as earlier, there exists two mutually singular invariant probability measures ν_1 and ν_2, that is $\nu_1(B_1) = 1$ and $\nu_2(B_2) = 1$, $B_1 \cap B_2 = \emptyset$ and that $P^n(x, B_1^C) = 0$ for all $x \in B_1$ and $n \in \mathbb{Z}_+$ and likewise $P^n(z, B_1^C) = 1$ for all $z \in B_2$ and $n \in \mathbb{Z}_+$. Now, every point x in S is so that one can approach x through two sequences y_n, z_n, one in B_1 and one in B_2 whose evaluations of $P^n(\cdot, B_1^C)$ are 1 apart from each other as y_n, z_n converge to one another (through x). This violates strong continuity. ◇

Another useful result is the following. Let us first recall the following: A family of functions F mapping a metric space $\mathbb{S}$ to $\mathbb{R}$ is said to be *equi-continuous at a point* $x_0 \in \mathbb{S}$ if, for every $\epsilon > 0$, there exists a $\delta > 0$ such that $d(x, x_0) \leq \delta \implies |f(x) - f(x_0)| \leq \epsilon$ for all $f \in F$. The family F is said to be *equicontinuous* if it is equicontinuous at each $x \in \mathbb{S}$.

Definition 3.13. *[233, Chapter 6] A Markov chain with transition kernel P is called an e-chain if for each continuous function f with compact support, the sequence of functions $\{\int P^n(x, dy)f(y), n \in \mathbb{Z}_+\}$ is equi-continuous on compact sets.*

Theorem 3.3.6 *[233, Prop. 18.4.2] If a Markov chain is an e-chain, $\mathbb{X}$ is compact, and a reachable state x^* exists, then there exists a unique invariant probability measure.*

In the following, we present a more concise argument compared with [233, Prop. 18.4.2].

Proof. By compactness, by Theorem 3.3.1 we know that there exists at least one invariant probability measure. Let there be two different probability measures ν^1, ν^2. We may assume ν^1 and ν^2 to be ergodic[2], via an ergodic decomposition argument of invariant measures on compact subsets [309, Theorem 6.1]. Now, similar to the proof of Theorem 3.3.5, since x^* must belong to the support of any two distinct probability measures (recall that $B_r(x^*)$ is visited under either of the probability measures in finite time for any given $r > 0$) we have that there exists two sequences y_n, z_n which converge to one another (through x^*) where y_n, z_n belong to the support sets of these two distinct probability measures ν^1, ν^2.

Now, by equi-continuity and the Arzela-Ascoli theorem [111], we have that

$$P^{(N)}(f)(x) := \frac{1}{N} \sum_{k=0}^{N-1} \int P^n(x, dy) f(y) \tag{3.33}$$

has a subsequence which converges (in the sup norm) to a limit $F_f^* : \mathbb{X} \to \mathbb{R}$, where is F_f^* continuous.

The above imply that, for every continuous and bounded f, the term

$$\lim_{n\to\infty} | \lim_{N\to\infty} P^{(N)}(f)(y_n) - P^{(N)}(f)(z_n)| = 0.$$

Suppose not; there would be an $\epsilon > 0$ and a subsequence n_k for which the difference

$$| \lim_{N\to\infty} P^{(N)}(f)(y_{n_k}) - P^{(N)}(f)(z_{n_k})| > \epsilon.$$

However, for each fixed n_k, we have that

$$\lim_{N\to\infty} P^{(N)}(f)(y_{n_k})$$

converges by the ergodicity of ν^1 to $\langle \nu^1, f \rangle$ and the limit, by the Arzela-Ascoli theorem, will be equal to $F_f^*(y_{n_k})$ (as every converging subsequence would have to converge to the limit; which also implies that the subsequential convergence in (3.33) is in fact a sequential convergence). The same argument applies for $P^{(N)}(f)(z_{n_k}) \to \langle \nu^2, f \rangle = F_f^*(z_{n_k})$.

The above would then imply that $|F_f^*(y_{n_k}) - F_f^*(z_{n_k})| \geq \epsilon$ for every (y_{n_k}, z_{n_k}). This would be a contradiction due to the continuity of F_f^*.

Therefore, the time averages of f under ν^1 and ν^2 will be arbitrarily close to each other. However, since continuous functions *separate* probability measures (e.g. via the metric given in (D.3), see also [120, Theorem 3.4.5]), this implies that the probability measures ν^1 and ν^2 must be equal. ◇

There exist further refinements on unique ergodicity via such equi-continuity conditions on transition kernels [195] and with equi-continuity replaced with uniform Lipschitz regularity [1].

3.4 Ergodic Theorems for Markov Chains

3.4.1 Ergodic theorems for positive Harris recurrent chains

Let $c \in L_1(\mu) := \{f : \int |f(x)|\mu(dx) < \infty\}$. Suppose that μ is an invariant ergodic probability measure for a Markov chain (a sufficient condition being that μ is the unique invariant probability measure [168, Prop. 2.4.3]). Then (see e.g. [168, Chapter 2]) it follows that for μ almost everywhere $x \in \mathbb{X}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} c(x_t) = \int c(x)\mu(dx),$$

[2] we say that an invariant measure μ measure is ergodic if for every absorbing set S, $\mu(S) \in \{0, 1\}$ [168, Definition 2.4.1]

P_x almost surely (that is conditioned on $x_0 = x$, with probability one, the above holds); see also Theorem 3.4.2. Furthermore, again with $c \in L_1(\mu)$, for μ almost everywhere $x \in \mathbb{X}$

$$\lim_{T\to\infty} \frac{1}{T} E_x\Big[\sum_{t=1}^{T} c(x_t)\Big] = \int c(x)\mu(dx),$$

On the other hand, the positive Harris recurrence property allows the almost sure convergence to take place for **every** initial condition: If μ is the invariant probability measure for a *positive Harris recurrent* Markov chain, it follows that for all $x \in \mathbb{X}$ and for every $c \in L_1(\mu)$ [233, Theorem 17.1.7] or [168, Theorem 4.2.13]

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} c(x_t) = \int c(x)\mu(dx), \tag{3.34}$$

almost surely. However, for every $c \in L_1(\mu)$, while (7.48) holds for all $x \in \mathbb{X}$, it is not generally true that

$$\lim_{T\to\infty} \frac{1}{T} E_x[\sum_{t=1}^{T} c(x_t)] = \int c(x)\mu(dx).$$

Thus, we can not in general relax the boundedness condition for the convergence of the expected costs. However, with c bounded, forall $x \in \mathbb{X}$

$$\lim_{T\to\infty} \frac{1}{T} E_x\Big[\sum_{t=1}^{T} c(x_t)\Big] = \int c(x)\mu(dx) \tag{3.35}$$

This follows as a consequence of Fatou's lemma and (7.48). Further refinements are possible via return properties to small sets and f-regularity of cost functions [13, 233]; e.g., this convergence holds if Theorem 4.2.4 holds for the given f and for some Lyapunov function V with $X_0 = x \in \{z : V(z) < \infty\}$; see the proof of Theorem 4.2.4 and [233, Theorem 14.0.1] for further related results. We refer the reader to [233, Chapters 14 and 17] or [168, Chapters 2 and 4] for additional discussions. See Exercise 3.5.8.

3.4.2 Further ergodic theorems for Markov chains

Although beyond the scope of this course, for completeness, we state the following. When an invariant probability measure is known to exist for a Markov chain, we state the following ergodicity results.

Theorem 3.4.1 *[167, Theorems 2.3.4-2.3.5] Let $\bar{P}$ be an invariant probability measure for a Markov process.*

(i) [Individual ergodic theorem] Let $X_0 = x$. For every $f \in L_1(\bar{P})$

$$\frac{1}{N} E_x[\sum_{n=0}^{N-1} f(X_n)] \to f^*(x),$$

for all $x \in B_f$ where $\bar{P}(B_f) = 1$ (where B_f denotes that the set of convergence may depend on f) for some f^*.

(ii) [Mean ergodic theorem] Furthermore, the convergence $\frac{1}{N} E_x[\sum_{n=0}^{N-1} f(X_n)] \to f^*(x)$ is in $L_1(\bar{P})$.

Theorem 3.4.2 *[167, Theorem 2.5.1] Let $\bar{P}$ be an invariant probability measure for a Markov process. With $X_0 = x$, for every $f \in L_1(\bar{P})$*

$$\frac{1}{N} \sum_{n=0}^{N-1} f(X_n) \to f^*(x),$$

for all $x \in B_f$ where $\bar{P}(B_f) = 1$ for some $f^*(x)$ with

$$\int \bar{P}(dx) f^*(x) = \int \bar{P}(dx) f(x)$$

One may state further refinements; see [167] for the locally compact case and [336] for the Polish state space case.

Theorem 3.4.3 *[336, Prop. 5.4] or [168, Theorem 3.1(g)] Let $\bar{P}$ be an invariant probability measure for a Markov process.*

(i) [Ergodic decomposition and weak convergence] For x, $\bar{P}$ a.s., $\frac{1}{N} E_x[\sum_{t=0}^{N-1} 1_{\{x_n \in \cdot\}}] \to P_x(\cdot)$ weakly and $\bar{P}$ is invariant for $P_x(\cdot)$ in the sense that

$$\bar{P}(B) = \int P_x(B) \bar{P}(dx)$$

(ii) [Convergence in total variation] For all $\mu \in \mathcal{P}(\mathbb{X})$ which satisfies that $\mu \ll \bar{P}$ (that is, μ is absolutely continuous with respect to $\bar{P}$), there exists an invariant v^* such that

$$\| E_\mu[\frac{1}{N} \sum_{t=0}^{N-1} 1_{\{T^n X \in \cdot\}}] - v^*(\cdot) \|_{TV} \to 0.$$

3.5 Exercises

Exercise 3.5.1 *For a countable state space Markov chain, prove that if $\{x_t\}$ is irreducible, then all states have the same period.*

Exercise 3.5.2 *Prove that*

$$P_x(\tau_A = 1) = P(x, A),$$

and for $n \geq 1$,

$$P_x(\tau_A = n) = \sum_{i \notin A} P(x, i) P_i(\tau_A = n - 1)$$

Exercise 3.5.3 *Let $\{x_t\}$ be a Markov chain defined on state space $\{0, 1, 2\}$. Let the one-stage probability transition matrix be given by:*

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 2/3 & 1/3 \end{bmatrix}$$

Compute $E[\min(t \geq 0 : x_t = 2) | x_0 = 0]$, that is the expected minimum number of stages for the state to move from 0 to 2.

Hint: Building on the previous exercise, one way to solve this problem is as follows: Note that if the expected minimum time to go to state 2 is from state 1 is t_1 and the expected minimum time to go to state 2, from state 0 is t_0, then the expected minimum time to go to state 2 from state 0 will be $t_0 = 1 + P(0,2)t_2 + P(0,1)t_1 + P(0,0)t_0$, where $t_2 = 0$. You can follow this line of reasoning to obtain the result.

Exercise 3.5.4 *Let $(\Omega, \mathcal{F}, P)$ be a probability space on which a Markov chain is defined: Let $\mathbb{X}$ be a finite set and X_n be the $\mathbb{X}$-valued Markov chain. Let $\alpha \in \mathbb{X}$ with $E_\alpha[\tau_\alpha] = 5$, where*

$$\tau_\alpha := \min\{k > 0 : x_k = \alpha\}$$

As we know from our class, this Markov chain admits an invariant probability measure, call it π. Suppose that X_n is irreducible so that the invariant probability measure is unique.

Now, let Y_n be an i.i.d. $\{0,1\}$-valued Bernoulli process (defined on the same probability space) with

$$P(Y_n = 1) = \eta \in (0,1).$$

Let

$$\tau_{(\alpha,1)} := \min\{k > 0 : (X_k, Y_k) = (\alpha, 1)\}.$$

a) Is $Z_k := (X_k, Y_k)$ a Markov chain? Prove your answer.
b) Find $E[\tau_{(\alpha,1)}|(X_0, Y_0) = (\alpha, 1)]$.

Exercise 3.5.5 *Show that irreducibility of a Markov chain in a finite state space implies that every set A and every x satisfies $U(x, A) = \infty$.*

Exercise 3.5.6 *Show that for an irreducible Markov chain, either the entire chain is transient, or recurrent.*

Exercise 3.5.7 *Show that for $\alpha_t \in (0,1)$,*

$$\prod_{t=0}^{\infty}(1-\alpha_t) > 0$$

if and only if $\sum_t \alpha_t < \infty$.

Hint. For one direction, use $\log(1-x) < -x$ for small $x \in (0,1)$. For the other direction, use $\lim_{x\to 0} \frac{\log(1-x)}{x} = -1$ and that as a result for small enough $x > 0 : \log(1-x) > -2x$ and that $\sum_t \alpha_t < \infty$ implies that $\alpha_t \to 0$.

Exercise 3.5.8 *In view of Exercise 3.5.7, let us revise the example given in Remark 3.2: Let $\mathbb{X} = \mathbb{N}$, $P(1,1) = 1$ and for $x > 1$: $P(x, x+1) = 1 - 1/x$ and $P(x, 1) = 1/x$. This chain is then Positive Harris Recurrent with invariant measure δ_1 and irreducibility measure also δ_1. Show that with $f(x) = x - 1$:*

$$\lim_{N\to\infty} \frac{1}{N} E_x[\sum_{k=0}^{N-1} f(x_k)] \neq E_{\delta_1}[f(X)] = 0, \qquad x \neq 1$$

Thus, expected empirical summations do not necessarily converge to the summation under the invariant measure when the function is not bounded. Note that this would be the case if the functions are bounded. Observe also that sample path convergence here does not imply the convergence of expected averages.

Exercise 3.5.9 *For a Markov chain with a countable space $\mathbb{X}$, and $a \in \mathbb{X}$, show that if $P_a(\tau_a < \infty) < 1$ then $E_a[\sum_{k=1}^{\infty} 1_{\{x_k = a\}}] < \infty$.*

Exercise 3.5.10 *For a Markov chain with a countable space $\mathbb{X}$, and $a \in \mathbb{X}$, show that if $P_a(\tau_a < \infty) = 1$ then $P_a(\sum_{k=1}^{\infty} 1_{\{x_k = a\}} = \infty) = 1$.*

Exercise 3.5.11 *Consider a Markov chain with state space $[0,1]$ and transition kernel given as follows:*

$$P(X_1 = \frac{x}{4} | X_0 = x) = 1, \qquad x \in [0,1].$$

Does there exist an invariant probability measure π for this Markov chain? If so, what is one such measure? Is this a unique invariant probability measure? If there is no invariant probability measure, precisely explain why this is the case.

Exercise 3.5.12 (Gambler's Ruin) *Consider an asymmetric random walk defined as follows: $P(x_{t+1} = x+1 | x_t = x) = p$ and $P(x_{t+1} = x-1 | x_t = x) = 1-p$ for any integer x. Suppose that $x_0 = x$ is an integer between 0 and N. Let $\tau = \min(k > 0 : x_k \notin [1, N-1])$. Compute $P_x(x_\tau = N)$ (you may use Matlab for your solution).*

Hint: Observe that one can obtain a recursion as $P_x(x_\tau = N) = pP_{x+1}(x_\tau = N) + (1-p)P_{x-1}(x_\tau = N)$ for $1 \le x \le N-1$ with boundary value conditions $P_N(x_\tau = N) = 1$ and $P_0(x_\tau = N) = 0$. One observes that

$$P_{x+1}(x_\tau = N) - P_x(x_\tau = N) = \frac{1-p}{p}\Big(P_x(x_\tau = N) - P_{x-1}(x_\tau = N)\Big)$$

and in particular

$$P_N(x_\tau = N) - P_{N-1}(x_\tau = N) = (\frac{1-p}{p})^{N-1}\Big(P_1(x_\tau = N) - P_0(x_\tau = N)\Big)$$

Exercise 3.5.13 *Let $x_{t+1} = f(x_t) + w_t$, where $f : \mathbb{R} \to \mathbb{R}$ is continuous and $\{w_t\}$ is an i.i.d. real valued noise sequence.*

a) Show that $\{x_t\}$ is weak Feller, regardless of the random variable w_t.

b) Show that $\{x_t\}$ is strong Feller, if w_t is a Gaussian random variable with a positive variance.

Exercise 3.5.14 *a) Consider a Markov chain defined on $\mathbb{Z}_+$ with the transition kernel*

$$P(x_1 = x+1|x_0 = x) = 1 - \frac{1}{x+1}, \qquad x \neq 0, x \in \mathbb{Z}_+,$$
$$P(x_1 = 0|x_0 = x) = \frac{1}{x+1}, \qquad x \neq 0, x \in \mathbb{Z}_+,$$

with

$$P(x_1 = 1|x_0 = 0) = 1.$$

Does there exist an invariant probability measure π for this Markov chain? If so, what is one such measure?

b) Consider a Markov chain defined on $\mathbb{Z}_+$ with the transition kernel

$$P(x_1 = x+1|x_0 = x) = \frac{1}{x+1}, \qquad x \neq 0, x \in \mathbb{Z}_+,$$
$$P(x_1 = 0|x_0 = x) = 1 - \frac{1}{x+1}, \qquad x \neq 0, x \in \mathbb{Z}_+,$$

with

$$P(x_1 = 1|x_0 = 0) = 1.$$

Does there exist an invariant probability measure π for this Markov chain? If so, what is one such measure?

b) Consider a Markov chain defined on $[0, 1]$ with the transition kernel:

$$P(x_1 = \frac{x}{2}|x_0 = x) = 1, \qquad x \neq 0, x \in [0, 1],$$
$$P(x_1 = 1|x_0 = 0) = 1.$$

Does there exist an invariant probability measure π for this Markov chain? If so, what is one such measure?

Exercise 3.5.15 *Consider a square and join opposite corners of this square by straight lines meeting at the point C. Consider the symmetric random walk performed by a particle on these 5 vertices, starting at some vertex A. Find*

(a) the expected time to return to A,

(b) the expected number of visits to C before returning to A,

(c) the expected time to return to A given that there is no prior visit to C.

4

Martingales, Foster-Lyapunov Criteria for Stability of Markov Chains, and Stochastic Iterative Dynamics

In this chapter, we will study martingales, which constitute a critical class of stochastic processes for our understanding of stochastic dynamics. We will arrive at stochastic stability of Markov chains through martingale methods and Foster-Lyapunov type stability criteria; this will be followed by an analysis on stochastic iterative dynamics.

4.1 Martingales

In this section, we introduce martingales and discuss a number of important martingale theorems. Only a few of these will be critical within the scope of our coverage, some others are presented for completeness.

These are very important for us to understand stabilization of controlled stochastic systems. These also will pave the way to optimization of dynamical systems as well as the supporting theory for stochastic learning, reinforcement learning, and approximation algorithms to be studied later. The second half of this chapter is on the stability of Markov chains or the stabilization of controlled Markov Chains via martingale and Lyapunov methods.

4.1.1 More on expectations and conditional probability

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{G}$ be a subset of $\mathcal{F}$ which is itself a σ-field (such a collection is said to be a sub-σ-field of $\mathcal{F}$). Let X be an $\mathbb{R}-$valued random variable measurable with respect to $(\Omega, \mathcal{F})$ with a finite absolute expectation that is

$$E[|X|] = \int_{\Omega} |X(\omega)| P(d\omega) < \infty,$$

where $\omega \in \Omega$. We call such random variables *integrable*.

We say that Ξ is the conditional expectation random variable (and is also called a version of the conditional expectation) of X given $\mathcal{G}$, denoted with,

$$E[X|\mathcal{G}],$$

if

1. Ξ is $\mathcal{G}$-measurable.
2. For every $A \in \mathcal{G}$,

$$E[1_A \Xi] = E[1_A X],$$

where we use the notation

$$E[1_A \Xi] = \int_{\Omega} \Xi(\omega) 1_{\{\omega \in A\}} P(d\omega) = \int_A \Xi(\omega) P(d\omega)$$

For example, if the information that we know about a process is whether an event $A \in \mathcal{F}$ happened or not -that is, the sub-σ-field is $\sigma(\{A\}) = \{\emptyset, \Omega, A, \Omega \setminus A\}$-, then the conditional expectation of X given the sigma-field generated by A is the random variable:

$$\eta(\omega) = E[X|\sigma(\{A\})](\omega) = X_A 1_{\{\omega \in A\}} + X_{A^C} 1_{\{\omega \notin A\}}.$$

Observe that this random variable is $\sigma(\{A\})$-measurable. We then have that for $w \in A$:

$$\eta(\omega) =: X_A =: E[X|A] = \frac{1}{P(A)} \int_A P(d\omega) X(\omega).$$

This follows from the fact that $\int_A E[X|A](\omega) P(d\omega) = E[X|A] \int_A P(d\omega)$ since $\eta(\omega) = E[X|A]$ cannot distinguish between any $\omega \in A$: this is a consequence of the fact that the conditional expectation is $\sigma(\{A\})$-measurable. Thus, we can simply write $\eta(\omega) = E[X|A]$ for the conditional expectation rather that $\eta(\omega)$. If the information we have is that A did not take place; then for $\omega \notin A$:

$$\eta(\omega) =: X_{A^C} := E[X|A^C] = \frac{1}{P(\Omega \setminus A)} \int_{\Omega \setminus A} P(d\omega) X(\omega).$$

Note that conditional probability can be expressed as

$$P(X \in B|\mathcal{G}) = E[1_{\{X \in B\}}|\mathcal{G}],$$

hence, conditional probability is a special case of conditional expectation.

It is a useful exercise (see Exercise 1.6.6) to consider the σ-field generated by an observation variable, and what a conditional expectation means in this case.

Theorem 4.1.1 *Let X be an $\mathbb{X}$ valued random variable, where $\mathbb{X}$ is a complete, separable, metric space and Y be another $\mathbb{Y}-$valued random variable, Then, X is $\mathcal{F}_Y$ or $\sigma(Y)$ (the $\sigma-$field generated by Y) measurable if and only if there exists a measurable function $f : \mathbb{Y} \to \mathbb{X}$ such that $X(\omega) = f(Y(\omega))$.*

With the above, the expectation $E[X|\sigma(Y)](\omega : Y(\omega) = y_0) = E[X|Y = y_0]$ can be defined as a measurable function on $\sigma(Y)$, and this expectation can be expressed as a measurable function of Y.

The notion of conditional expectation is key for the development of stochastic processes which evolve according to a transition kernel. This is useful for optimal decision making when only partial information is available with regard to a random variable.

The following discussion is optional until the next subsection.

Existence of Conditional Expectation.

Theorem 4.1.2 (Radon-Nikodym) *Let μ and ν be two σ-finite positive measures on $(\Omega, \mathcal{F})$ such that $\nu(A) = 0$ implies that $\mu(A) = 0$ (that is μ is absolutely continuous with respect to ν). Then, there exists a measurable function $f : \Omega \to \mathbb{R}_+$, (integrable under ν if μ is a finite measure), such that for every A:*

$$\mu(A) = \int_A f(\omega) \nu(d\omega)$$

The representation above is unique, up to sets of measure zero. With the above discussion, the conditional expectation $X = E[X|\mathcal{F}']$ exists for any sub-σ-field $\mathcal{F}' \subset \mathcal{F}$, as the following discussion shows. Let X be an integrable non-negative random variable and observe that for any Borel $A \in \mathcal{F}'$

$$\int_A \Big(E[X|\mathcal{F}'](\omega) \Big) P(d\omega) = \int_A X(\omega) P(d\omega).$$

We may view $\zeta(A) := \int_A X(\omega)P(d\omega)$ as a measure (defined on the measurable space $(\Omega, \mathcal{F}')$) which is absolutely continuous with respect to P, and thus, $\Big(E[X|\mathcal{F}'](\omega)\Big)$, is the Radon-Nikodym derivative of this measure with respect to P (This discussion extends to arbitrary integrable variables by considering the negative valued portion of the variable separately).

In case X is a real random variable which is of second-order (i.e., with finite second moment), another way to establish existence is through a Hilbert theoretic approach, by viewing the conditional expectation as the projection of X onto a subspace consisting of the set of all functions measurable on $\mathcal{F}'$. We will revisit this later in the notes while deriving the Kalman Filter in *Chapter 6*. However, for this we would require X to be square-integrable (i.e., with a finite second moment).

4.1.2 Some properties of conditional expectation:

One very important property is given by the following.

Iterated expectations:

Theorem 4.1.3 *For three σ-fields over a given set, if $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, and X is $\mathcal{F}-$measurable and integrable, it follows that:*

$$E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{H}]$$

Proof. Proof follows by taking a set $A \in \mathcal{H}$, which is also in $\mathcal{G}$ and $\mathcal{F}$. Let η be the conditional expectation variable with respect to $\mathcal{H}$. Then it follows that

$$E[1_A\eta] = E[1_A X]$$

Now let $E[X|\mathcal{G}]$ be η'. Then, it must be that $E[1_A\eta'] = E[1_A X]$ for all $A \in \mathcal{G}$ and hence for all $A \in \mathcal{H}$. Thus, we have that for all $A \in \mathcal{H}$

$$E[1_A\eta'] = E[1_A\eta],$$

and as η is $\mathcal{H}$-measurable, $E[\eta'|\mathcal{H}] = \eta$. ◇

Theorem 4.1.4 *Let $\mathcal{G} \subset \mathcal{F}$, and Y be $\mathcal{G}-$measurable. Let X be $\mathcal{F}-$measurable and XY be integrable. Then, P almost surely*

$$E[XY|\mathcal{G}] = YE[X|\mathcal{G}]$$

Proof. First assume that $Y = Y_n$ is a simple function (a simple random variable of the form: $Y_n(\omega) = \sum_{i=1}^n a_i 1_{\{\omega\in A_i\}}$ with $A_i \in \mathcal{G}$). Let us call $E[X|\mathcal{G}] = \eta$ and call $E[XY|\mathcal{G}] = \zeta$.

Then, for all $A \in \mathcal{G}$

$$\begin{aligned}
&\int_A Y_n\eta(\omega)P(d\omega) = \int_A \sum_{i=1}^n a_i 1_{\{\omega\in A_i\}}\eta(\omega)P(d\omega) \\
&= \sum_{i=1}^n a_i \int_{A\cap A_i} \eta(\omega)P(d\omega) = \sum_{i=1}^n a_i \int_{A\cap A_i} X(\omega)P(d\omega) \qquad (4.1)\\
&= \int_A \sum_{i=1}^n a_i 1_{\{\omega\in A_i\}} X(\omega)P(d\omega) = \int_A Y_n X P(d\omega)
\end{aligned}$$

Here, (4.1) holds since $A \cap A_i \in \mathcal{G}$ and $E[X|\mathcal{G}] = \eta$. On the other hand,

$$\int_A \zeta(\omega)P(d\omega) = \int_A X(\omega)Y_n(\omega)P(d\omega)$$

$$= \int_A \sum_{i=1}^{n} a_i 1_{\{\omega \in A_i\}} X(\omega) P(d\omega) = \sum_{i=1}^{n} a_i \int_{A \cap A_i} X(\omega) P(d\omega) = \int_A Y_n X P(d\omega)$$

Thus, for all $A \in \mathcal{G}$

$$\int_A Y_n X P(d\omega) = \int_A E[XY_n|\mathcal{G}](\omega) P(d\omega) = \int_A Y_n E[X|\mathcal{G}](\omega) P(d\omega), \tag{4.2}$$

and the two conditional expectations $E[XY_n|\mathcal{G}]$ and $Y_n E[X|\mathcal{G}]$ are equal. Now, the proof is complete by noting that any integrable Y can be approached from below monotonically by a sequence of simple functions measurable on $\mathcal{G}$. The monotone convergence theorem leads to the desired result. ◇

4.1.3 Discrete-time martingales

Let $(\Omega, \mathcal{F}, P)$ be a probability space. An increasing family $\{\mathcal{F}_n\}$ of sub-$\sigma-$fields of $\mathcal{F}$ is called a *filtration*.

A sequence of random variables defined on $(\Omega, \mathcal{F}, P)$ is said to be adapted to $\mathcal{F}_n$ if X_n is $\mathcal{F}_n$-measurable, that is $X_n^{-1}(D) = \{w \in \Omega : X_n(w) \in D\} \in \mathcal{F}_n$ for all Borel D. This holds for example if $\mathcal{F}_n = \sigma(X_m, m \leq n), n \geq 0$; in this case we call the filtration, the *natural filtration*.

Given a filtration $\mathcal{F}_n$ and a sequence of real random variables adapted to it, $(X_n, \mathcal{F}_n)$ is said to be a **martingale** if

$$E[|X_n|] < \infty$$

and

$$E[X_{n+1}|\mathcal{F}_n] = X_n.$$

We will often take the sigma-fields to be the natural filtration $\mathcal{F}_n = \sigma(X_1, X_2, \ldots, X_n)$.

Let $n > m \in \mathbb{Z}_+$. Since $\mathcal{F}_m \subset \mathcal{F}_n$, it must be that $A \in \mathcal{F}_m$ should also be in $\mathcal{F}_n$. Thus, if X_n is a martingale sequence, we have for $A \in \mathcal{F}_m$

$$E[1_A X_n] = E[1_A X_{n-1}] = \cdots = E[1_A X_m],$$

and thus $E[X_n|\mathcal{F}_m] = X_m$.

If we have that

$$E[X_n|\mathcal{F}_m] \geq X_m$$

then $\{X_n\}$ is called a **submartingale**.

And, if

$$E[X_n|\mathcal{F}_m] \leq X_m$$

then $\{X_n\}$ is called a **supermartingale**.

A useful concept related to filtrations is that of a stopping time, which we discussed while studying Markov chains. A stopping time τ is a random time, whose occurrence at any given time is causally measurable with respect to the filtration in the sense that for each $n \in \mathbb{N}$, $\{\tau \leq n\} \in \mathcal{F}_n$.

Definition 4.1.1 (Filtration up to a stopping time) *Let $\mathcal{F}_t$ denote a filtration and τ be a stopping time with respect to this filtration so that for every k, $\{\tau \leq k\} \in \mathcal{F}_k$. Then, the σ-field of events up to τ, $\mathcal{F}_\tau$, is the collection of all events $A \in \mathcal{F}$ that satisfies:*

$$A \cap \{\tau \leq t\} \in \mathcal{F}_t, \quad \forall t \in \mathbb{Z}_+.$$

Intuitively, then, the natural filtration up to a stopping time is all the information generated by a stochastic process up to the stopping time.

Exercise 4.1.1 *Let τ be a stopping time with $\tau \geq m$ for some $m \in \mathbb{Z}_+$. Show that $\mathcal{F}_m \subset \mathcal{F}_\tau$.*

4.1.4 Doob's optional sampling theorem

Theorem 4.1.5 *Suppose $(X_n, \mathcal{F}_n)$ is a martingale sequence, and $\rho, \tau < n$ (for some fixed $n \in \mathbb{N}$) are (uniformly) bounded stopping times with $\rho \leq \tau$. Then,*

$$E[X_\tau | \mathcal{F}_\rho] = X_\rho$$

Proof. We observe that

$$\begin{aligned}
E[X_\tau - X_\rho | \mathcal{F}_\rho] &= E[\sum_{k=\rho}^{\tau-1} X_{k+1} - X_k | \mathcal{F}_\rho] \\
&= E[\sum_{k=\rho}^{\infty} 1_{\{\tau > k\}} (X_{k+1} - X_k) | \mathcal{F}_\rho] \\
&= E[\sum_{k=\rho}^{n} 1_{\{\tau > k\}} (X_{k+1} - X_k) | \mathcal{F}_\rho] \\
&= \sum_{k=\rho}^{n} E[1_{\{\tau > k\}} (X_{k+1} - X_k) | \mathcal{F}_\rho] \qquad (4.3) \\
&= \sum_{k=\rho}^{n} E[E[1_{\{\tau > k\}} (X_{k+1} - X_k) | \mathcal{F}_k] | \mathcal{F}_\rho] \\
&= E[\sum_{k=\rho}^{n} 1_{\{\tau > k\}} E[(X_{k+1} - X_k) | \mathcal{F}_k] | \mathcal{F}_\rho] \\
&= E[\sum_{k=\rho}^{\tau-1} 0 | \mathcal{F}_\rho] = 0 \qquad (4.4)
\end{aligned}$$

Here, we invoke Theorem 4.1.3 and Theorem 4.1.4, since $1_{\{\tau > k\}}$ is $\mathcal{F}_k$-measurable. ◇

The statement of the theorem leads to inequalities for supermartingales or submartingales with the appropriate inequality signs.

In the above, the main properties we used were (i) the fact that the sub-fields are nested, (ii) n is bounded so that the expectation of the sum can be written as the sum of expectations in (4.3).

Let us try to see why boundedness of the stopping times is important: Consider the following game. Suppose that one draws a fair coin; with equal probabilities of heads and tails. If we have a tail, we win a dollar, and a head will cause us to lose a dollar. Suppose we have 0 dollars at time 0 and we decide to stop when we have 5 dollars, that is at time $\tau = \min(n > 0 : X_n = 5)$. In this case, clearly $E[X_\tau] = 5$, as we will stop when we have 5 dollars. But $E[X_\tau] \neq X_0$!

Remark 4.1. For this example, $X_n = X_{n-1} + W_n$, where W_n is either -1 or 1 with equal probabilities and X_n is the amount of money we have. Clearly X_n is a martingale sequence. The problem is that one might have to wait for an arbitrarily long period of amount of time to be able to have the 5 dollars, the sequence $E[|X_n|]$ is not uniformly bounded, and $\sum_{k=\rho}^{n} 1_{\{\tau > k\}} (X_{k+1} - X_k) = X_{\min(\tau, n)} - X_\rho$ is not a (uniformly) integrable sequence, and the proof method adopted in Theorem 4.1.5 will not be applicable. Note that if we were able to claim that

$$\begin{aligned}
E[X_\tau - X_\rho | \mathcal{F}_\rho] &= E[\lim_{n \to \infty} X_{\min(\tau, n)} - X_\rho | \mathcal{F}_\rho] \\
&= E\left[\lim_{n \to \infty} \left(\sum_{k=\rho}^{n} 1_{\{\tau > k\}} (X_{k+1} - X_k) \right) \middle| \mathcal{F}_\rho \right] = \lim_{n \to \infty} E[\sum_{k=\rho}^{n} 1_{\{\tau > k\}} (X_{k+1} - X_k) | \mathcal{F}_\rho], \qquad (4.5)
\end{aligned}$$

then the result would have been applicable even if we didn't have a finite upper bound on the stopping times. This requires in particular:

(i) the almost sure finiteness of τ so that $\lim_{n\to\infty} X_{\min(\tau,n)} = X_\tau$, and

(ii) a dominated convergence result for the sequence $X_{\min(\tau,n)} - X_\rho = \sum_{k=\rho}^{n} 1_{\{\tau>k\}}(X_{k+1} - X_k)$ (e.g., the presence of an integrable random variable $G(\omega)$ with $X_{\min(\tau,n)}(\omega) \leq G(\omega)$).

If these hold, then we can indeed apply the argument above when the stopping times are not bounded. More on this will be discussed below in Theorem 4.1.14, in the context of uniform integrability.

4.1.5 Doob's maximal inequality (optional)

Theorem 4.1.6 *For a non-negative supermartingale M_n, for all $\lambda > 0$,*

$$P(\sup_{0\leq n<\infty} M_n \geq \lambda) \leq \frac{M_0}{\lambda}$$

Proof. Let $\tau^N = \min\left\{\min\{n \geq 0 : M_n \geq \lambda\}, N\right\}$ for some $N \in \mathbb{N}$. Then,

$$P(\max_{0\leq n<N} M_n \geq \lambda) = P(M_{\tau^N} \geq \lambda) \leq \frac{E[M_{\tau^N}]}{\lambda} \leq \frac{M_0}{\lambda},$$

where the first inequality follows from Markov's inequality and the last from Doob's optional sampling theorem. The relation above applies for all $N \in \mathbb{N}$, and since the left hand side is non-decreasing in N, the limit of it as $N \to \infty$ is well-defined. Furthermore, by an application of continuity in probability $\lim_{N\to\infty} P(\max_{0\leq n<N} M_n \geq \lambda) = P(\sup_{0\leq n<\infty} M_n \geq \lambda)$, and the result follows. ◇

An important generalization is known as Doob's L_p-maximal inequality.

Theorem 4.1.7 *For a submartingale M_n, let $M_N^* = \sup_{0\leq n<N} |M_n|$. Then, for every $p > 1$:*

$$E[|M_N^*|^p] \leq (\frac{p}{p-1})^p E[|M_N|^p]$$

Proof Sketch. Write $E[|M_N^*|^p = \int_0^\infty P(|M_N^*|^p > t)dt = p\int_0^\infty P(|M_N^*|^p > s^p)s^{p-1}ds$, where we apply a change of variables with $t = s^p$. Let $\tau_s^N := \min\{\min\{n \geq 0 : M_n \geq s\}, N\}$. Then, we have that for every $s > 0$,

$$sP(|M_N^*| > s) \leq E[|M_{\tau_s^N}|1_{\{|M_N^*|>s\}}] \leq E[|M_N|1_{\{|M_N^*|>s\}}],$$

by the submartingale property and the optional sampling theorem (as $E[|M_N||\mathcal{F}_{\tau_s^N}] \geq |E[M_N|\mathcal{F}_{\tau_s^N}]| \geq |M_{\tau_s^N}|$). Therefore,

$$\begin{aligned} E[|M_N^*|^p &\leq p\int_0^\infty P(|M_N^*| > s)s^{p-1}ds \leq p\int_0^\infty E[|M_N|1_{|M_N^*|>s}]s^{p-2}ds \\ &= E[|M_N|\int_0^\infty p1_{\{|M_N^*|>s\}}s^{p-2}ds] = E[|M_N|\int_0^{|M_N^*|} ps^{p-2}ds] \\ &= E[|M_N|\frac{p}{p-1}|M_N^*|^{p-1}] \leq \frac{p}{p-1}(E[|M_N|^p])^{\frac{1}{p}}(E[|M_N^*|^p])^{1-\frac{1}{p}}, \end{aligned} \tag{4.6}$$

where we use Hölder's inequality in the final step. The result then follows by rearranging the terms. ◇

This inequality is very useful in the approximation theory for controlled diffusions, as it relates the maximal deviations to L_p deviations.

4.1.6 An important martingale convergence theorem

We first discuss **Doob's upcrossing lemma**. Let (a, b) be a non-empty interval. Let $X_0 \in (a, b)$. Define a sequence of stopping times

$$T_1 = \min\{N; \min(n : 0 \le n \le N, X_n \le a)\} \qquad T_2 = \min\{N; \min(n : T_1 \le n \le N, X_n \ge b)\}$$

$$T_3 = \min\{N; \min(n : T_2 \le n \le N, X_n \le a)\} \qquad T_4 = \min\{N; \min(n : T_3 \le n \le N, X_n \ge b)\}$$

and for $m > 2$:

$$T_{2m-1} = \min\{N; \min(n : T_{2m-2} \le n \le N, X_n \le a)\} \qquad T_{2m} = \min\{N; \min(n : T_{2m-1} \le n \le N, X_n \ge b)\}$$

The number of upcrossings of (a, b) up to time N is the random variable $\zeta_N(a, b) =$ the number of times between 0 and N, $\{X_n\}$ crosses the strip (a, b) from below a to above b.

Note that if the sequence is a supermartingale, $X_{T_2} - X_{T_1}$ has a negative expectation.

Theorem 4.1.8 *Let X_T be a supermartingale sequence. Then,*

$$E[\zeta_N(a, b)] \le \frac{E[\max(0, a - X_N)]}{b - a} \le \frac{E[|X_N|] + |a|}{b - a}.$$

Proof.
There are three possibilities that might take place: The process can end, at time N while the process is below a, between a and b, or above b. If it crosses above b, then we have completed an upcrossing. In view of this, we may proceed as follows: Let $\beta_N := \min(m : T_{2m} = N \quad \text{or} \quad T_{2m-1} = N)$ (note that if $T_{2m-1} = N$, $T_{2m} = N$ as well). By the supermartingale property

$$\begin{aligned}
0 \ &\ge E[\sum_{i=1}^{\beta_N} X_{T_{2i}} - X_{T_{2i-1}}] \\
&= E[\Big(\sum_{i=1}^{\beta_N} X_{T_{2i}} - X_{T_{2i-1}}\Big) 1_{\{T_{2\beta_N - 1} \ne N\}} 1_{\{T_{2\beta_N} = N\}}] + E[\Big(\sum_{i=1}^{\beta_N} X_{T_{2i}} - X_{T_{2i-1}}\Big) 1_{\{T_{2\beta_N - 1} = N\}} 1_{\{T_{2\beta_N} = N\}}] \\
&= E\Big(\sum_{i=1}^{\beta_N - 1} X_{T_{2i}} - X_{T_{2i-1}}\Big) + E[(X_N - X_{T_{2\beta_N - 1}}) 1_{\{T_{2\beta_N - 1} \ne N\}} 1_{\{T_{2\beta_N} = N\}}] \qquad (4.7)
\end{aligned}$$

Thus,

$$\begin{aligned}
&E\Big(\sum_{i=1}^{\beta_N - 1} X_{T_{2i}} - X_{T_{2i-1}}\Big) \le -E[(X_N - X_{T_{2\beta_N - 1}}) 1_{\{T_{2\beta_N - 1} \ne N\}} 1_{\{T_{2\beta_N} = N\}}] \\
&= E[(X_{T_{2\beta_N - 1}} - X_N) 1_{\{T_{2\beta_N - 1} \ne N\}} 1_{\{T_{2\beta_N} = N\}}] \le E[\max(0, a - X_N) 1_{\{T_{2\beta_N} = N\}}] \le E[\max(0, a - X_N)] \qquad (4.8)
\end{aligned}$$

Since, $E\Big(\sum_{i=1}^{\beta_N - 1} X_{T_{2i}} - X_{T_{2i-1}}\Big) \ge E[\beta_N - 1](b - a)$, it follows that $\zeta_N(a, b) = (\beta_N - 1)$ satisfies:

$$E[\zeta_N(a, b)](b - a) \le E[\max(0, a - X_N)] \le |a| + E[|X_N|],$$

and the result follows. ◇

Recall that a sequence of random variables X_n defined on a probability space $(\Omega, \mathcal{F}, P)$ converges to X almost surely (a. s.) if

$$P\Big(w : \lim_{n \to \infty} X_n(w) = X(w)\Big) = 1.$$

Theorem 4.1.9 *Suppose X_n is a supermartingale and $\sup_{n\geq 0} E[\max(0,-X_n)] < \infty$. Then $\lim_{n\to\infty} X_n = X$ exists (almost surely). The same result applies for submartingales, by regarding $-X_n$ as a supermartingale and the condition $\sup_{n\geq 0} E[\max(0,X_n)] < \infty$. A sufficient condition for both cases is that*

$$\sup_{n\geq 0} E[|X_n|] < \infty.$$

Proof. The proof follows from Doob's upcrossing lemma. Now, for any fixed a, b (independent of ω) with $a < b$, by the upcrossing lemma we have that

$$E[\zeta_N(a,b)] \leq E[\max(0, a - X_N)] \leq \frac{E[|X_N|] + |a|}{(b-a)},$$

which is uniformly bounded. The above holds for every N. Since $\zeta_N(a,b)$ is a monotonically increasing sequence in N, by the monotone convergence theorem it follows that

$$\lim_{N\to\infty} E[\zeta_N(a,b)] = E[\lim_{N\to\infty} \zeta_N(a,b)] \leq \sup_N \frac{E[|X_N|] + |a|}{(b-a)} < \infty.$$

Thus, for every fixed $a < b$, the number of up-crossings has to be finite almost surely. Hence, the limsup cannot be above b and the liminf cannot be below a, for otherwise the number of up-crossings would be infinite. It then follows that

$$P(\omega : |\limsup X_n(\omega) - \liminf X_n(\omega)| > (b-a)) = 0,$$

since this probability can be expressed also as

$$P\bigg(\cup_{r\in\mathbb{Q}} \bigg\{ \omega : \limsup X_n(\omega) > (b-a+r), \liminf X_n(\omega)| < r \bigg\} \bigg),$$

and by a union bound argument, the probability is upper bounded by the probability of a countable union of zero probability events which is zero. Finally, a continuity of probability argument (for taking $b - a \to 0$) then leads to

$$P\bigg(\omega : |\limsup X_n(\omega) - \liminf X_n(\omega)| > 0 \bigg) = 0.$$

◇

We can also show that the limit variable has finite absolute expectation.

Theorem 4.1.10 (Submartingale Convergence Theorem) *Suppose X_n is a submartingale and $\sup_{n\geq 0} E[|X_n|] < \infty$. Then $X := \lim_{n\to\infty} X_n$ exists (almost surely) and $E[|X|] < \infty$.*

Proof. Note that, $\sup_{n\geq 0} E[|X_n|] < \infty$, is a sufficient condition both for a submartingale and a supermartingale in Theorem 4.1.9. Hence $X_n \to X$ almost surely. For finiteness, suppose $E[|X|] = \infty$. By Fatou's lemma,

$$\limsup_{n\to\infty} E[|X_n|] \geq \liminf_{n\to\infty} E[|X_n|] \geq E[\liminf_{n\to\infty} |X_n|] = E[\lim_{n\to\infty} |X_n|] = \infty.$$

But this is a contradiction as we had assumed that $\sup_n E[|X_n|] < \infty$. ◇

4.1.7 The ergodic theorem[Optional]

See Exercise 4.5.11.

4.1.8 Further martingale theorems [Optional]

This section is optional. If you wish not to study it, please proceed to the discussion on stabilization of Markov Chains.

Theorem 4.1.11 *Let X_n be a martingale such that X_n converges to some integrable X in L_1 that is $E[|X_n - X|] \to 0$. Then,*

$$X_n = E[X|\mathcal{F}_n], \quad n \in \mathbb{N}$$

We will use the following while studying the convex analytic method, as well as on the stabilization of Markov chains while extending the optional sampling theorem to situations where the sampling (stopping) time is not bounded from above by a finite number. Let us define uniform integrability:

Definition 4.1.2 *: A sequence of random variables $\{X_n, n \in \mathbb{N}\}$ is uniformly integrable if*

$$\lim_{K\to\infty} \sup_{n\in\mathbb{N}} \int_{|X_n|\geq K} |X_n| P(dX_n) = 0$$

This implies that

$$\sup_n E[|X_n|] < \infty$$

Let for some $\epsilon > 0$,

$$\sup_n E[|X_n|^{1+\epsilon}] < \infty.$$

This implies that the sequence is uniformly integrable as

$$\sup_n \int_{|X_n|\geq K} |X_n| P(dX_n) \leq \sup_n \int_{|X_n|\geq K} (\frac{|X_n|}{K})^{\epsilon} |X_n| P(dX_n) \leq \sup_n \frac{1}{K^{\epsilon}} E[|X_n|^{1+\epsilon}] \to 0,$$

as $K \to \infty$. The following result is important in many applications:

Theorem 4.1.12 *If X_n is a uniformly integrable martingale, then $X = \lim_{n\to\infty} X_n$ exists almost surely (for all sequences with probability 1) and in L_1 (i.e. $E[|X - X_n|] \to 0$), and $X_n = E[X|\mathcal{F}_n]$.*

Theorem 4.1.13 *Let X be integrable and $\mathcal{F}_n$ be a filtration (not necessarily the natural filtration). Then, $M_n = E[X|\mathcal{F}_n]$ is uniformly integrable.*

Proof. First note that by Jensen's inequality $|E[X|\mathcal{F}_n]| \leq E[|X||\mathcal{F}_n]$ (since $|\cdot|$ is a convex function). Therefore, by Markov's inequality, for any $K \in \mathbb{R}_+$:

$$P(|E[X|\mathcal{F}_n]| > K] \leq E[|E[X|\mathcal{F}_n]|]/K \leq E[E[|X||\mathcal{F}_n]]/K = \frac{E[|X|]}{K},$$

which decays to zero as $K \to \infty$. Now, consider the measure defined with $|X(\omega)|dP(\omega)$: For any set sequence A_m with $P(A_m) \to 0$, we have that

$$\lim_{m\to\infty} E[1_{A_m}|X|] = 0 \tag{4.9}$$

This follows from a contradiction argument: suppose this is not true, then there exists a subsequence A_{m_k} with $E[1_{A_{m_k}}|X|] \geq \epsilon$ some fixed $\epsilon > 0$ and a further subsequence $A_{m'_k}$ with a finite $\sum_{m'_k} P(A_{m'_k})$. Then a monotone convergence theorem violation can be established so that with $B_n = \cup_{m'_k \geq n} A_{m'_k}$, $E[1_{B_n}|X|] \not\to 0$ where B_n is a monotone decreasing sequence whose measure vanishes. Therefore

$$E\left[|E[X|\mathcal{F}_n]|1_{\{|E[X|\mathcal{F}_n]|>K\}}\right] \leq E\left[E[|X||\mathcal{F}_n]1_{\{|E[X|\mathcal{F}_n]|>K\}}\right]$$

$$= E\Big[E[|X|1_{\{|E[X|\mathcal{F}_n]|>K\}}|\mathcal{F}_n]\Big] = E[|X|1_{\{|E[X|\mathcal{F}_n]|>K\}}]$$

where we use Theorem 4.1.3 (iterated expectations) and thus combining the above

$$\lim_{K\to\infty}\sup_n E\Big[|E[X|\mathcal{F}_n]|1_{|E[X|\mathcal{F}_n]|>K}\Big] \leq \lim_{K\to\infty}\sup_n E\Big[|X|1_{|E[X|\mathcal{F}_n]|>K}\Big] = 0,$$

where the last step follows from (4.9). ◇

Optional Sampling Theorem For Uniformly Integrable Martingales

The following builds on Remark 4.1.

Theorem 4.1.14 *Let $(X_n, \mathcal{F}_n)$ be a uniformly integrable martingale sequence, and ρ, τ are finite stopping times with $\rho \leq \tau$. Then,*

$$E[X_\tau|\mathcal{F}_\rho] = X_\rho$$

Proof. See the discussion following (4.5) for an explicit analysis and derivation. ◇

Azuma-Hoeffding inequality for martingales with bounded increments

The following is an important concentration result:

Theorem 4.1.15 *Let X_t be a martingale sequence such that $|X_t - X_{t-1}| \leq c$ for every t, almost surely. Then for any $x > 0$,*

$$P(\frac{X_t - X_0}{t} \geq x) \leq 2e^{-\frac{tx^2}{2c}}$$

As a result, $\frac{X_t}{t} \to 0$ almost surely.

Backwards (reverse) martingales and decreasing information

An important class of martingales is known as backward martingales. A sequence of increasing σ-fields with a negative time index,

$$\cdots \subset \mathcal{F}_{-n} \subset \mathcal{F}_{-n+1} \subset \cdots \subset \mathcal{F}_{-2} \subset \mathcal{F}_{-1} \subset \mathcal{F}_0,$$

is called a reverse filtration. Note that here information is decreasing as $n \to -\infty$. M_n is called a backwards martingale with respect to the reverse filtration if (i) $E[|M_{-n}|] < \infty$, (ii) $E[M_{-n+1}|\mathcal{F}_{-n}] = M_{-n}$ for all $n \in \mathbb{Z}$.

Using similar arguments as those in the proof of the martingale convergence theorem studied earlier, we can arrive at the following:

Theorem 4.1.16 *Let $(M_{-n}, \mathcal{F}_{-n})$ be a backwards martingale sequence. Then, $\lim_{n\to-\infty} M_n =: M_{-\infty} = E[M_0| \cap_{n=0}^{\infty} \mathcal{F}_{-n}]$ almost surely and also in L_1.*

4.2 Stability of Markov Chains: Foster-Lyapunov Techniques

Via martingale theory, a Markov chain's stability can be characterized by drift conditions, as we discuss below in detail.

4.2.1 Criterion for invariance (existence of invariant probability measures) and positive Harris recurrence

Theorem 4.2.1 *[Foster-Lyapunov for Positive Harris Recurrence] [233] Let S be a petite set, $b \in \mathbb{R}$, $\epsilon > 0$, and $V : \mathbb{X} \to \mathbb{R}_+$. If the following is satisfied for all $x \in \mathbb{X}$:*

$$E[V(x_{t+1})|x_t = x] = \int_{\mathbb{X}} P(x, dy)V(y) \leq V(x) - \epsilon + b1_{\{x \in S\}}, \tag{4.10}$$

then the chain is positive Harris recurrent (and thus a unique invariant probability measure π exists).

Proof. We will first assume that S is such that $\sup_{x \in S} V(x) < \infty$. Define $\bar{M}_0 := V(x_0)$, and for $t \geq 1$

$$\bar{M}_t := V(x_t) - \sum_{i=0}^{t-1}(-\epsilon + b1_{\{x_i \in S\}})$$

We have that

$$E[\bar{M}_{(t+1)}|x_s, s \leq t] \leq \bar{M}_t, \quad \forall t \geq 0.$$

It follows from (4.10) that $E_x[|\bar{M}_t|] \leq \infty$ for all t (by an application of the monotone convergence theorem applied inductively: suppose that $E[V(x_t)] < \infty$; then first show that $E[E[\min(N_1, V(x_{t+1}))|x_t]] \leq E[V(x_t)] + b$, then the take the limit as $N_1 \to \infty$ to conclude that $E[V(x_{t+1})] < \infty$ as well) and thus, $\{\bar{M}_t\}$ is a supermartingale sequence with respect to the natural filtration $\mathcal{F}_t = \sigma(x_0, \cdots, x_t)$. Now, define a stopping time: $\tau^N := \min(\tau, N)$, where $\tau = \min\{i > 0 : x_i \in S\}$. Note that the stopping time τ^N is bounded. Hence, we have, by the martingale optional sampling theorem

$$E[\bar{M}_{\tau^N}|x_0] \leq M_0.$$

Thus, we obtain

$$\epsilon E_{x_0}[\sum_{i=0}^{\tau^N - 1} 1] \leq V(x_0) + bE_{x_0}[\sum_{i=0}^{\tau^N - 1} 1_{\{x_i \in S\}}]$$

Thus,

$$\epsilon E_{x_0}[\tau^N - 1 + 1] \leq V(x_0) + b,$$

and by the monotone convergence theorem (and that $P(\tau < \infty) = 1$ by the uniform bound on the expectation below),

$$\lim_{N \to \infty} E_{x_0}[\tau^N] = E_{x_0}[\tau] \leq \frac{V(x_0) + b}{\epsilon}. \tag{4.11}$$

Note that, the first equality above is a consequence of the drift criterion:

$$\frac{V(x_0) + b}{\epsilon} \geq \lim_{N \to \infty} E_{x_0}[\tau^N] \geq \limsup_{N \to \infty}(NP_{x_0}(\tau \geq N) + E_{x_0}[\tau 1_{\{N > \tau\}}]) \geq \limsup_{N \to \infty} NP_{x_0}(\tau \geq N),$$

implying that $P_{x_0}(\tau \geq N) \to 0$ as $N \to \infty$ and that $P_{x_0}(\tau_S < \infty) = 1$.) Now, if we had that

$$\sup_{x \in S} V(x) < \infty, \tag{4.12}$$

the proof would essentially be complete in view of Theorem 3.2.6. The fact that $E_x[\tau] \leq \frac{V(x)+b}{\epsilon} < \infty$ for any $x \in \mathbb{X}$ leads to the Harris recurrence of the chain since this implies that $P_x(\tau < \infty) = 1$ for every x and petiteness implies that the chain would be positive Harris recurrent [233, Proposition 9.1.7] (see also [90, Theorem 3.1]).

Typically, condition (4.12) is satisfied. However, the theorem statement does not impose this condition. Then, we proceed with constructing another petite set on which (4.12) holds. Following [233, Chapter 11], define for some $l \in \mathbb{Z}_+$

$$V_S(l) = \{x \in S : V(x) \leq l\}$$

We will show that $B := V_S(l)$ is itself a petite set which is recurrent and satisfies the uniform finite-mean-return property. It can be shown that, without any loss S is petite for some measure ν with $\nu(S) > 0$, [1] and thus by a continuity of probability argument, for sufficiently large l, we also have that $\nu(B) > 0$. Again, since S is petite for measure ν, we have that

$$K_a(x, B) \geq 1_{\{x \in S\}} \nu(B), \quad x \in \mathbb{X},$$

where $K_a(x, B) = \sum_{i \in \mathbb{N}} a(i) P^i(x, B)$, and hence

$$1_{\{x \in S\}} \leq \frac{1}{\nu(B)} K_a(x, B)$$

Now, for $x \in B$,

$$E_x[\tau_B] \leq V(x) + bE_x[\sum_{k=0}^{\tau_B - 1} 1_{\{x_k \in S\}}] \leq V(x) + bE_x[\sum_{k=0}^{\tau_B - 1} \frac{1}{\nu(B)} K_a(x_k, B)] \tag{4.13}$$

$$= V(x) + b\frac{1}{\nu(B)} E_x[\sum_{k=0}^{\tau_B - 1} K_a(x_k, B)] = V(x) + b\frac{1}{\nu(B)} E_x[\sum_{k=0}^{\tau_B - 1} \sum_i a(i) P^i(x_k, B)] \tag{4.14}$$

$$= V(x) + b\frac{1}{\nu(B)} \sum_i a(i) E_x[\sum_{k=0}^{\tau_B - 1} 1_{\{x_{k+i} \in B\}}]$$

$$\leq V(x) + b\frac{1}{\nu(B)} \sum_i a(i)(1 + i), \tag{4.15}$$

where (4.15) follows since at most once the process can hit B between 0 and $\tau_B - 1$. Now, the petiteness measure can be adjusted such that $\sum_i a_i i < \infty$ (by Theorem 3.2.3 or [233, Proposition 5.5.6]), leading to the result that

$$\sup_{x \in B} E_x[\tau_B] \leq \sup_{x \in B} V(x) + b\frac{1}{\nu(B)} \sum_i a(i)(1 + i) < \infty.$$

Finally, since S is petite, so is B and it can be shown that $P_x(\tau_B < \infty) = 1$ for all $x \in \mathbb{X}$. This concludes the proof. $\diamond$

Remark 4.2. Note that irreducibility of the Markov chain is not imposed a priori building on [233, Proposition 9.1.8] or [90, Theorem 3.1], the drift criterion and the small/petite nature of the set leads to an irreducible Markov chain (possibly defined on a proper subset of $\mathbb{X}$).

Remark 4.3. Meyn and Tweedie [233, Theorem 13.0.1] show that under the hypotheses of Theorem 4.2.1, together with aperiodicity, it also follows that for any initial state $x \in \mathbb{X}$,

$$\lim_{n \to \infty} \sup_{B \in \mathcal{B}(\mathbb{X})} |P^n(x, B) - \pi(B)| = 0,$$

that is $P^n(x, \cdots)$ converges to π in *total variation*, for every $x \in \mathbb{X}$. This follows from a coupling argument, to be discussed further in the chapter.

Exercise 4.2.1 *Consider a queuing system with*

$$Q_{t+1} = \max(Q_t + A_t - N1_{Q_t \geq N}, 0)$$

where A_t is an i.i.d. Poisson arrival process with rate λ so that

[1] To see this, first take some set B with $\nu(B) > 0$ and also with $V(x) < M$ for all $x \in B$ for some M large enough (such a set B exists since the set $\{x : V(x) < \infty\}$ is absorbing and then by a continuity of probability argument) and then consider the transitions from B to S using the uniform probability of transitions by some time m large enough, by Markov's inequality given (4.11), which then provides a positive lower bound on transitions under a sampling distribution on $\mathbb{Z}_+$ from any $x \in S$ to S via first visiting B and then coming back to S.

$$P(A_t = m) = e^{-\lambda}\frac{\lambda^m}{m!}, \quad m \in \mathbb{Z}_+$$

Suppose that $N > \lambda$. Show that Q_t is positive Harris recurrent.

Remark 4.4. We note that if x_t is aperiodic and irreducible and such that for some small set A we have $\sup_{x\in A} E[\min(t > 0 : x_t \in A)|x_0 = x] < \infty$, then the sampled chain $\{x_{km}\}$ is such that $\sup_{x\in A} E[\min(km > 0 : x_{km} \in A)|x_0 = x] < \infty$, and the split chain discussion in Section 3.2.1 applies (See [233, Theorem 11.3.14]). The argument for this builds on the fact that, with $\sigma_C = \min(k \geq 0 : x_k \in C)$, $V(x) := 1+E_x[\sigma_C]$, it follows that $E[V(x_{t+1})|x_t = x] \leq V(x)-1+b1_{\{x\in C\}}$ and iterating the expectation m times we obtain that

$$E[V(x_{t+m})|x_t = x] \leq V(x) - m + bE_x[\sum_{k=0}^{m-1} 1_{\{x_k\in C\}}]. \tag{4.16}$$

By [233], it follows that $E_x[\sum_{k=0}^{m-1} 1_{\{x_k\in C\}}] \leq m1_{\{x\in C_\epsilon\}} + m\epsilon$ for some petite set C_ϵ and $\epsilon > 0$ (this follows from the observation that $\{x : P^k(x, C) \geq \epsilon\}$ will be included in the petite set for at least one k with $1 \leq k \leq m-1$ and the complement of these sets $\{x : P^k(x, C) < \epsilon\}$ will contribute to an upper bound of $m\epsilon$ in (4.16)). This set is petite also for the sampled chain (see Lemma 4.2.1). As a result, we have a drift condition for the m-skeleton, the return time for an artificial atom constructed through the split chain is finite and hence an invariant probability measure for the m-skeleton, and thus by (3.23), an invariant probability measure for the original chain exists. ◇

In the following, we relax the existence of a petite set or irreducibility, but impose that the space is locally compact (and not just Polish or standard Borel). This builds on [233, Theorem 12.3.4] or [168, Theorem 7.2.4].

Theorem 4.2.2 *If the Markov chain is weak Feller, the space is locally compact, and S is compact; under (4.10), there exists an invariant probability measure.*

Proof. Iterating (4.10) we obtain that, with

$$P^{(n)}(x, S) := \frac{1}{n}E_x[\sum_{k=0}^{n-1} 1_{\{x_k\in S\}}],$$

we arrive at

$$\liminf_{n\to\infty} P^{(n)}(x, S) \geq \frac{\epsilon}{b}.$$

The result then follows from Theorem 3.3.2. ◇

There are other versions of Foster-Lyapunov criteria, as we discuss in the following.

4.2.2 Criterion for finite expectations

Theorem 4.2.3 *[Comparison Theorem] [233, Theorem 14.2.2] Let $V : \mathbb{X} \to \mathbb{R}_+$, $f, g : \mathbb{X} \to \mathbb{R}_+$. Let $\{x_n\}$ be a Markov chain on $\mathbb{X}$. If the following is satisfied:*

$$\int_{\mathbb{X}} P(x, dy)V(y) \leq V(x) - f(x) + g(x), \qquad x \in \mathbb{X},$$

then, for any stopping time τ with $P(\tau < \infty) = 1$, it follows that

$$E[\sum_{t=0}^{\tau-1} f(x_t)] \leq V(x_0) + E[\sum_{t=0}^{\tau-1} g(x_t)]$$

Proof. As in Theorem 4.2.1, define $\bar{M}_0 := V(x_0)$, and for $t \geq 1$

$$\bar{M}_t := V(x_t) + \sum_{i=0}^{t-1}(f(x_i) - g(x_i)).$$

It follows that

$$E[\bar{M}_{(t+1)}|x_s, s \le t] \le \bar{M}_t, \quad \forall t \ge 0.$$

Now, define a stopping time: $\tau^N = \min(\tau, \min(k > 0 : k + V(x_k) + \sum_{i=0}^{k-1} f(x_k) + g(x_k) \ge N))$. Note that the stopping time τ^N is bounded. It then follows that (through defining a supermartingale: $M_t := \bar{M}_{\min(t,\tau^N)}$), and by the martingale optional sampling theorem:

$$E[M_{\tau^N}|x_0] \le M_0 = V(x_0).$$

Hence, we obtain

$$E\left[V(x_{\tau^N}) + \sum_{i=0}^{\tau^N-1}(f(x_i) - g(x_i))\Big|x_0\right] \le \bar{M}_0 = V(x_0),$$

and thus by the fact that the terms inside the expectations are separately integrable, we have that

$$E[\sum_{i=0}^{\tau^N-1} f(x_i)|x_0] \le \bar{M}_0 = V(x_0) + E[\sum_{i=0}^{\tau^N-1} g(x_i)|x_0] - E[V(x_{\tau^N})|x_0].$$

Now, since each of the terms in the expectations is positive, and that $E[V(x_{\tau^N})|x_0] \ge 0$, the monotone convergence theorem implies the desired result. ◇

Theorem 4.2.3 above also allows for the computation of useful bounds. For example if $g(x) = b1_{\{x\in A\}}$, then one obtains that $E[\sum_{t=0}^{\tau-1} f(x_t)] \le V(x_0) + b$. In view of the invariant measure properties, if $f(x) \ge 1$, this provides a bound on $\int \pi(dx)f(x)$, as we note next.

Theorem 4.2.4 *[Criterion for finite expectations] [233] Let S be a petite set, $b \in \mathbb{R}_+$ and $V : \mathbb{X} \to \mathbb{R}_+$, $f : \mathbb{X} \to [\epsilon, \infty)$ for some $\epsilon > 0$. Let $\{x_n\}$ be a Markov chain on $\mathbb{X}$.*

(i) If the following is satisfied:

$$\int_{\mathbb{X}} P(x, dy)V(y) \le V(x) - f(x) + b1_{\{x\in S\}}, \qquad x \in \mathbb{X}, \tag{4.17}$$

then for every $x_0 = z \in \mathbb{X}$,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) = \int \mu(dx)f(x) \le b, \tag{4.18}$$

almost surely, where μ is the invariant probability measure on $\mathbb{X}$.

(ii) If $\{x_t\}$ is positive Harris recurrent, even if $f : \mathbb{X} \to \mathbb{R}_+$ (and not necessarily $f : \mathbb{X} \to [\epsilon, \infty)$ for some $\epsilon > 0$) and $S = \mathbb{X}$ itself (that is, with no indicator function), (4.17) implies (4.18).

That under Theorem 4.2.4, the process is a positive Harris recurrent Markov chain is a consequence of Theorem 4.2.1. The proof of Theorem 4.2.4 will then build on the following result and the ergodicity of a positive Harris recurrent Markov chain.

Theorem 4.2.5 *Let (4.17) hold (but with not necessarily an irreducibility assumption), or the following more relaxed form hold:*

$$\int_{\mathbb{X}} P(x, dy)V(y) \le V(x) - f(x) + b, \qquad x \in \mathbb{X} \tag{4.19}$$

Under every invariant probability measure π, $\int \pi(dx)f(x) \le b$.

Proof. By Theorem 4.2.3, with taking T to be a deterministic stopping time, for any initial condition $x_0 = z$

$$\limsup_{T\to\infty} \frac{1}{T} E_z[\sum_{k=0}^{T-1} f(x_k)] \leq \limsup_{T\to\infty} \frac{1}{T}\Big(V(z) + bT\Big) = b. \tag{4.20}$$

Now, suppose that π is any invariant probability measure. Fix $N < \infty$, let $f_N = \min(N, f)$, and apply Fatou's Lemma as follows, where we use the notation $\pi(f) = \int \pi(dx) f(x)$,

$$\begin{aligned}\pi(f_N) &= \limsup_{n\to\infty} \pi\Big(\frac{1}{n}\sum_{t=0}^{n-1} P^t f_N\Big) \\ &\leq \pi\Big(\limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} P^t f_N\Big) \leq b\,.\end{aligned}$$

Fatou's Lemma is justified to obtain the first inequality, because f_N is bounded. The monotone convergence theorem, with taking $N \to \infty$, then gives $\pi(f) \leq b$. ◇

Remark 4.5. We note that where the system starts from, or what the initial distribution is on x_0 is, affects the convergence properties of

$$\frac{1}{T} E[\sum_{k=0}^{T-1} f(x_k)].$$

See Section 3.4.1 for a detailed discussion. In particular, it is not necessarily the case that for every initial measure convergence of expected normalized values to $\int \pi(dx) f(x)$ holds. Furthermore, sample paths and expectations have slightly different convergence characteristics for unbounded f.

4.2.3 Criterion for recurrence

Theorem 4.2.6 (Foster-Lyapunov for Recurrence) *Let S be a compact set, $b < \infty$, and V be an inf-compact functional from $\mathbb{X} \to \mathbb{R}_+$ such that for all $\alpha \in \mathbb{R}_+$ $\{x : V(x) \leq \alpha\}$ is compact (note: this implies that $\lim_{||x||\to\infty} V(x) = \infty$ if $\mathbb{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$). Let the following be satisfied for the Markov chain $\{x_k\}$:*

$$\int_{\mathbb{X}} P(x, dy) V(y) \leq V(x) + b1_{\{x\in S\}}, \quad \forall x \in \mathbb{X}\,, \tag{4.21}$$

Furthermore, with $\tau_S = \min(t > 0 : x_t \in S)$, and $\tau_{B_N} = \min(t > 0 : x_t \in B_N)$ where $B_N = \{z : V(z) \geq N\}$, if we have that $P_x(\min(\tau_S, \tau_{B_N}) = \infty) = 0$ for every $x \in \mathbb{X}$ and $N \in \mathbb{N}$, it must be that

$$P_x(\tau_S < \infty) = 1$$

for all $x \in \mathbb{X}$

Proof. Define two stopping times: Let $\tau_S = \min(t > 0 : x_t \in S)$ and $\tau_{B_N} = \min(t > 0 : x_t \in B_N)$ where $B_N = \{z : V(z) \geq N\}$ with $N \geq V(x)$ where $x_0 = x$. Note that $V(x_t)$ is bounded until $\tau^N := \min(\tau_S, \tau_{B_N})$ and until this time $E[V(x_{t+1})|\mathcal{F}_t] \leq V(x_t)$. By assumption $\tau^N = \min(\tau_S, \tau_{B_N}) < \infty$ with probability 1. Define $M_t = V(x_{\min(t,\tau^N)})$, which is a supermartingale sequence uniformly bounded (see Exercise 4.5.3). It follows then that a variation of the optional sampling theorem (see Theorem 4.1.14) applies so that

$$E_x[M_{\tau^N}] = E_x[V(x_{\min(\tau_S,\tau_{B_N})})] \leq V(x)$$

Without any loss take $x \notin S$ (for otherwise, in the next time stage x_1 we can make the argument replacing x_0 with x_1) and there exists N large enough so that $x \notin (S \cup B_N)$. Now, for $x \notin (S \cup B_N)$, since when exiting into B_N the minimum value of the Lyapunov function is N:

$$V(x) \geq E_x[V(x_{\min(\tau_S,\tau_{B_N})})] \geq P_x(\tau_{B_N} < \tau_S)N + P_x(\tau_{B_N} \geq \tau_S)M,$$

for some finite non-negative $M := \inf_{x\in S} V(x)$.

Hence,

$$P_x(\tau_{B_N} < \tau_S) \leq \frac{V(x)}{N}.$$

We also have that $P(\min(\tau_S, \tau_{B_N}) = \infty) = 0$. As a consequence, we have that

$$P_x(\tau_S = \infty) \leq P(\tau_{B_N} < \tau_S) \leq V(x)/N$$

and taking the limit as $N \to \infty$, $P_x(\tau_S = \infty) = 0$. ◇

Remark 4.6. If S is further petite, then once the petite set is visited, any other set with a positive measure (under an irreducibility measure, since the petiteness measure can be used to construct an irreducibility measure) is visited with probability 1 infinitely often and hence the chain is Harris recurrent. ◇

Exercise 4.2.2 *Show that the random walk on $\mathbb{Z}$ is Harris recurrent.*

4.2.4 Criterion for transience

Criteria for transience is somewhat more difficult to establish. One convenient way is to construct a stopping time sequence and show that the state does not come back to some set infinitely often. We state the following.

Theorem 4.2.7 (Criterion for Transience) *[233], [154] Let $V : \mathbb{X} \to \mathbb{R}_+$. If there exists a set A such that $E[V(x_{t+1})|x_t = x] \leq V(x)$ for all $x \notin A$ and $\exists \bar{x} \notin A$ such that $V(\bar{x}) < \inf_{z\in A} V(z)$, then $\{x_t\}$ is not recurrent, in the sense that $P_{\bar{x}}(\tau_A < \infty) < 1$.*

Proof. Let $x = \bar{x}$. Proof follows from observing that

$$V(x) \geq \int_y V(y)P(x, dy) \geq (\inf_{z\in A} V(z))P(x, A) + \int_{y\notin A} V(y)P(x, dy) \geq (\inf_{z\in A} V(z))P(x, A)$$

It thus follows that

$$P(\tau_A < 2) = P(x, A) \leq \frac{V(x)}{(\inf_{z\in A} V(z))}$$

Likewise,

$$\begin{aligned}
V(\bar{x}) &\geq \int_{\mathbb{X}} V(y)P(\bar{x}, dy) \\
&\geq (\inf_{z\in A} V(z))P(\bar{x}, A) + \int_{y\notin A} (\int_{\mathbb{X}} V(s)P(y, ds))P(\bar{x}, dy) \\
&\geq (\inf_{z\in A} V(z))P(\bar{x}, A) + \int_{y\notin A} P(\bar{x}, dy)\Big((\inf_{s\in A} V(s))P(y, A) + \int_{s\notin A} V(s)P(y, ds)\Big) \\
&\geq (\inf_{z\in A} V(z))P(\bar{x}, A) + \int_{y\notin A} P(\bar{x}, dy)((\inf_{s\in A} V(s))P(y, A)) \\
&= (\inf_{z\in A} V(z))\Big(P(\bar{x}, A) + \int_{y\notin A} P(\bar{x}, dy)P(y, A)\Big). \qquad (4.22)
\end{aligned}$$

Thus, noting that $P(\{\omega : \tau_A(\omega) < 3\}) = \int_A P(\bar{x}, dy) + \int_{y\notin A} P(\bar{x}, dy)P(y, A)$, we observe:

$$P_{\bar{x}}(\tau_A < 3) \leq \frac{V(\bar{x})}{(\inf_{z\in A} V(z))}.$$

Thus, this follows for any n: $P_{\bar{x}}(\tau_A < n) \leq \frac{V(\bar{x})}{(\inf_{z\in A} V(z))} < 1$. Continuity of probability measures (by defining: $B_n = \{\omega : \tau_A < n\}$ and observing $B_n \subset B_{n+1}$ and that $\lim_n P(\tau_A < n) = P(\cup_n B_n) = P(\tau_A < \infty) < 1$) now leads to $P_{\bar{x}}(\tau_A < \infty) < 1$. $\diamond$

Observe the striking difference with the inf-compactness condition leading to recurrence and the condition above, leading to non-recurrence.

4.2.5 Criterion for almost sure convergence to equilibrium

The following build on stochastic stability theorems due to Khasminskii [193] and Kushner [204].

Theorem 4.2.8 *(i) Let x_n be Markov so that for some $V : \mathbb{X} \to \mathbb{R}_+$ and $k : \mathbb{X} \to \mathbb{R}_+$, we have that*

$$E[V(x_{n+1})|x_n = x] \leq V(x) - k(x), \quad x \in \mathbb{X}.$$

Then, $k(x_n) \to 0$ with probability 1.

(ii) Let $S_\lambda := \{x : V(x) \leq \lambda\}$, and suppose that

$$E[V(x_{n+1})|x_n = x] \leq V(x) - k(x), \qquad x \in S_\lambda.$$

If $x_0 \in S_\lambda$, then,

$$P_{x_0}\left(\sup_{0\leq n<\infty} V(x_n) \geq \lambda\right) \leq V(x_0)/\lambda. \tag{4.23}$$

Hence, the paths remain in Q_λ with probability at least $1-\frac{V(x_0)}{\lambda}$. Furthermore, for paths that remain in Q_λ, $k(x_n) \to 0$ with probability 1.

(iii) Suppose that for each $\gamma > 0$, there exists $\delta > 0$ so that $k(x) \geq \delta$ for $|x| \geq \gamma$ and that $k(0) = 0$. Then, the origin is globally asymptotically stable with probability 1, that is, $\lim_{n\to\infty} x_n = 0$ almost surely.

(iv) Suppose that for some increasing function $c : \mathbb{X} \to \mathbb{R}_+$ with $c(0) = 0$ and $c(x) > 0$ for all $x \neq 0$, we have that $c(|x|) \leq V(x)$ and for some $\alpha > 0$

$$E[V(x_{n+1})|x_n = x] \leq V(x) - \alpha V(x), \quad x \in \mathbb{X}.$$

Then, the system is exponentially asymptotically stable in the sense that:

$$P_{x_0}\{\sup_{N\leq n<\infty} V(x_n) \geq \lambda\} \leq \frac{V(x_0)(1-\alpha)^N}{\lambda}.$$

Proof.

(i) By arguments presented earlier, it follows that $0 \leq E_{x_0}[V(x_n)] \leq V(x_0) - E_{x_0}[\sum_{m=0}^{n-1} k(x_m)]$. Thus, $E_{x_0}[\sum_{m=0}^{\infty} k(x_m)] < \infty$. But then, $\sum_{m=0}^{\infty} k(x_m) < \infty$ almost surely and thus $k(x_m) \to 0$ almost surely.

(ii) Define $M_0 = V(x_0)$ and for $n > 0 : M_n = V(x_n) - \sum_{m=0}^{n-1} k(x_m)$, which is a supermartingale sequence with respect to the natural filtration. Let us stop the process x_n on first leaving S_λ where $S_\lambda^c = \mathbb{X} \setminus S_\lambda$. Then, the stopped process $M_{\min(t,\tau_{S_\lambda^c})}$ is also a supermartingale process, where the drift equation holds with $k(x) = 0$ for $x \notin S_\lambda$ and if $\tau_{S_\lambda^c} = \infty$, we have that $k(x_m) \to 0$.

On the other hand, the bound in (4.23) builds essentially on the proof of Doob's maximal inequality Theorem 4.1.7, which notes that for a non-negative supermartingale R_n, for all $\lambda > 0$,

$$P(\max_{0\le n<\infty} R_n \ge \lambda) \le \frac{R_0}{\lambda}$$

With M_n the super-martingale sequence defined as before, we have that

$$P_{x_0}(V(x_{\tau_{S^c_\lambda}}) \ge \lambda) \le P_{x_0}\left(V(x_{\tau_{S^c_\lambda}}) - \sum_{m=0}^{\tau_{S^c_\lambda}-1} k(x_m) \ge \lambda\right) \le \frac{M_0}{\lambda}$$

(iii)By (i), we have that $k(x_n) \to 0$; the hypothesis then implies that $x_n \to 0$.

(iv)Observe first that $M_n := \frac{V(x_n)}{(1-\alpha)^n}$ is also a supermartingale. Apply Doob's maximal inequality (Theorem 4.1.7) as follows:

$$P(\sup_{N\le n<\infty} \frac{V(x_n)(1-\alpha)^n}{(1-\alpha)^n} \ge \lambda) \le P(\sup_{N\le n<\infty} \frac{V(x_n)}{(1-\alpha)^n} \ge \frac{\lambda}{(1-\alpha)^N}) \le \frac{E[M_N](1-\alpha)^N}{\lambda} \le \frac{M_0(1-\alpha)^N}{\lambda}.$$

◇

4.2.6 State dependent drift criteria: Deterministic and random-time

In many applications, a drift term (e.g. by a controller) can be applied on a system only intermittently.

Theorem 4.2.9 *[350] Suppose that $\{x_t\}$ is a φ-irreducible and aperiodic Markov chain. Suppose moreover that there are functions $V\colon \mathbb{X} \to (0,\infty)$, $\delta\colon \mathbb{X} \to [1,\infty)$, $f\colon \mathbb{X} \to [1,\infty)$, a small set C on which V is bounded, and a constant $b \in \mathbb{R}$, such that*

$$\begin{aligned} E[V(x_{\tau_{i+1}}) \mid \mathcal{F}_{\tau_i}] &\le V(x_{\tau_i}) - \delta(x_{\tau_i}) + b1_C(x_{\tau_i}) \\ E\Big[\sum_{k=\tau_i}^{\tau_{i+1}-1} f(x_k) \,\Big|\, \mathcal{F}_{\tau_i}\Big] &\le \delta(x_{\tau_i}), \qquad\qquad i \ge 0. \end{aligned} \tag{4.24}$$

Then the following hold:

(i) $\{x_t\}$ *is positive Harris recurrent, with unique invariant distribution* π

(ii) $\pi(f) := \int f(x)\,\pi(dx) < \infty$.

(iii) *For any function g that is bounded by f, in the sense that* $\sup_x |g(x)|/f(x) < \infty$*, we have convergence of moments in the mean, and the strong law of large numbers holds:*

$$\begin{aligned} \lim_{t\to\infty} E_x[g(x_t)] &= \pi(g) \\ \lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} g(x_t) &= \pi(g) \qquad a.s.,\ x \in \mathbb{X} \end{aligned}$$

By taking $f(x) = 1$ for all $x \in \mathbb{X}$, we obtain the following corollary to Theorem 4.2.9.

Corollary 4.2.1 *[350] Suppose that $\boldsymbol{X}$ is a φ-irreducible Markov chain. Suppose moreover that there is a function $V : \mathbb{X} \to (0,\infty)$, a petite set C on which V is bounded,, and a constant $b \in \mathbb{R}$, such that the following hold:*

$$\begin{aligned} E[V(x_{\tau_{z+1}}) \mid \mathcal{F}_{\tau_z}] &\le V(x_{\tau_z}) - 1 + b1_{\{x_{\tau_z}\in C\}} \\ \sup_{z\ge 0} E[\tau_{z+1} - \tau_z \mid \mathcal{F}_{\tau_z}] &< \infty. \end{aligned} \tag{4.25}$$

Then $\boldsymbol{X}$ is positive Harris recurrent. ◇

The above extend the deterministic state-dependent results presented in [233], [234]: Let $\tau_z, z \geq 0$ be a sequence of stopping times, measurable on a filtration, possible generated by the state process.

Without the irreducibility condition, if the chain is weak Feller, if (4.10) holds with S compact, then there exists at least one invariant probability measure as discussed in Section 3.3.1.

Theorem 4.2.10 *[350] Suppose that $\boldsymbol{X}$ is a Feller Markov chain, not necessarily φ-irreducible. If* (4.24) *holds with C compact then there exists at least one invariant probability measure. Moreover, there exists $c < \infty$ such that, under any invariant probability measure π,*

$$E_\pi[f(x)] = \int_{\mathbb{X}} \pi(dx) f(x) \leq c. \tag{4.26}$$

Petite sets and sampling

Unfortunately the techniques we reviewed earlier that rely on petite sets become unavailable in the random time drift setting considered in Section 4.2.6 as a petite set C for $\{x_n\}$ is not necessarily petite for $\{x_{\tau_n}\}$. Some of the discussion in this section is due to Zurkowski et. al. [354]

Lemma 4.2.1 *[354] Suppose $\{x_t\}$ is an aperiodic and irreducible Markov chain. If there exists sequence of stopping times $\{\tau_n\}$ independent of $\{x_t\}$, then any C that is small for $\{x_t\}$ is petite for $\{x_{\tau_n}\}$.*

Proof. Since C is petite, it is small by Theorem 3.2.4 for some m. Let C be (m, δ, ν)-small for $\{x_t\}$.

$$P^{\tau_1}(x, \cdot) = \sum_{k=1}^{\infty} P(\tau_1 = k) P^k(x, \cdot) \geq \sum_{k=m}^{\infty} P(\tau_1 = k) \int P^m(x, dy) P^{k-m}(y, \cdot)$$
$$\geq \sum_{k=m}^{\infty} P(\tau_1 = k) \int 1_C(x) \delta \nu(dy) P^{k-m}(y, \cdot) \tag{4.27}$$

which is a well defined measure. Therefore defining $\kappa(\cdot) = \int \nu(dy) \sum_{k=m}^{\infty} P(\tau_1 = k) P^{k-m}(y, \cdot)$, we have that C is $(1, \delta, \kappa)$-small for $\{x_{\tau_n}\}$. ◇

Thus, one can relax the condition that V is bounded on C in Theorem 4.2.9, if the sampling times are deterministic. Another condition is when the sampling instances are hitting times to a set which contains C [354].

4.2.7 Convergence Rates to Equilibrium

In addition to obtaining bounds on the rate of convergence through Dobrushin's coefficient studied earlier, a more relaxed and often more general approach is via Foster-Lyapunov drift conditions and an associated coupling analysis.

Regularity and ergodicity are concepts closely related through the work of Meyn and Tweedie [233], [237] and Tuominen and Tweedie [308].

Definition 4.2.1 *A set $A \in \mathcal{B}(\mathcal{X})$ is called (f, r)-regular if*

$$\sup_{x \in A} E_x[\sum_{k=0}^{\tau_B - 1} r(k) f(x_k)] < \infty$$

for all $B \in \mathcal{B}^+(\mathcal{X})$. A finite measure ν on $\mathcal{B}(\mathcal{X})$ is called (f, r)-regular if

$$E_\nu[\sum_{k=0}^{\tau_B-1} r(k)f(x_k)] < \infty$$

for all $B \in \mathcal{B}^+(\mathcal{X})$, and a point x is called (f,r)-regular if the measure δ_x is (f,r)-regular.

This leads to a lemma relating regular distributions to regular atoms.

Lemma 4.2.2 *If a Markov chain $\{x_t\}$ has an atom $\alpha \in \mathcal{B}^+(\mathcal{X})$ and an (f,r)-regular distribution λ, then α is an (f,r)-regular set.*

Definition 4.2.2 (f-norm) *For a function $f : \mathbb{X} \to [1,\infty)$ the f-norm of a measure μ defined on $(\mathbb{X}, \mathcal{B}(\mathcal{X}))$ is given by*

$$\|\mu\|_f = \sup_{g \le f} |\int \mu(dx)g(x)|.$$

The total variation norm is the f-norm when $f = 1$, denoted by $\|.\|_{TV}$.

Definition 4.2.3 *A Markov chain $\{x_t\}$ with invariant distribution π is (f,r)-ergodic if*

$$r(n)\|P^n(x,\,\cdot\,) - \pi(\,\cdot\,)\|_f \to 0 \quad \text{as } n \to \infty \text{ for all } x \in \mathbb{X}. \tag{4.28}$$

If (4.28) is satisfied for a geometric r (so that $r(n) = M\zeta^n$ for some $\zeta > 1$, $M < \infty$) and $f = 1$ then the Markov chain $\{x_t\}$ is called geometrically ergodic.

Coupling inequality and moments of return times to a small set The main idea behind the coupling inequality is to bound the total variation distance between the distributions of two random variables by the probability they are different Let X and Y be two jointly distributed random variables on a space $\mathbb{X}$ with distributions μ_x, μ_y respectively. Then we can bound the total variation between the distributions by the probability the two variables are not equal.

$$\begin{aligned}
\|\mu_x - \mu_y\|_{TV} =& \sup_A |\mu_x(A) - \mu_y(A)| \\
=& \sup_A |P(X \in A, X = Y) + P(X \in A, X \neq Y) \\
& \qquad - P(Y \in A, X = Y) - P(Y \in A, X \neq Y)| \\
\le& \sup_A |P(X \in A, X \neq Y) - P(Y \in A, X \neq Y)| \\
\le& P(X \neq Y)
\end{aligned}$$

The coupling inequality is useful in discussions of ergodicity when used in conjunction with parallel Markov chains. Later, we will see that the coupling inequality is also useful to establish the existence of optimal solutions to average cost optimization problems.

One creates two Markov chains having the same one-step transition probabilities. Let $\{x_n\}$ and $\{x'_n\}$ be two Markov chains that have probability transition kernel $P(x,\cdot)$, and let C be an (m,δ,ν)-small set. We use the coupling construction provided by Roberts and Rosenthal [264], building on the splitting technique presented in Section 3.2.1.

Let $x_0 = x$ and $x'_0 \sim \pi$ where π is the invariant probability measure for both Markov chains.

(1) If $x_n = x'_n$ then $x_{n+1} = x'_{n+1} \sim P(x_n,\cdot)$

(2) Else, if $(x_n, x'_n) \in C \times C$ then with probability δ, $x_{n+m} = x'_{n+m} \sim \nu(\cdot)$ with probability $1-\delta$ then

independently

$$x_{n+m} \sim \tfrac{1}{1-\delta}(P^m(x_n,\cdot) - \delta\nu(\cdot))$$

$$x'_{n+m} \sim \tfrac{1}{1-\delta}(P^m(x'_n,\cdot) - \delta\nu(\cdot))$$

(3) Else, independently $x_{n+m} \sim P^m(x_n,\cdot)$ and $x'_{n+m} \sim P^m(x'_n,\cdot)$.

The in-between states $x_{n+1},...x_{n+m-1}, x'_{n+1},...x'_{n+m-1}$ are distributed conditionally given x_n, x_{n+m},x'_n, x'_{n+m}.

By the Coupling Inequality and the previous discussion with Nummelin's Splitting technique in Section 3.2.1 we have $\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq P(x_n \neq x'_n)$.

Remark 4.7. Through the coupling inequality one can show that $\pi_0 P^n \rightarrow \pi$ in total variation. Furthermore, if Theorem 4.2.4 holds, one can also show that with some further analysis if the initial condition is a fixed deterministic state $\int \Big(P^n(x,dz) - \pi(dz)\Big)f(z) \rightarrow 0$, where f is not necessarily bounded. This does not imply, however, $\int \Big((\pi_0 P^n)(dz) - \pi(dz)\Big)f(z) \rightarrow 0$ for a *random* initial condition. A sufficient condition for the latter to occur is that $\int \pi_0(dz)V(z) < \infty$ provided that Theorem 4.2.4 holds (see Theorem 14.3.5 in [233]).

Rates of convergence: Geometric ergodicity

In this section, following [233] and [264], we review results stating that a strong type of ergodicity, geometric ergodicity, follows from a simple drift condition. An irreducible Markov chain is said to satisfy the *univariate drift condition* if there are constants $\lambda \in (0,1)$ and $b < \infty$, along with a function $V : \mathbb{X} \rightarrow [1,\infty)$, and a small set C such that

$$PV \leq \lambda V + b1_C. \tag{4.29}$$

Theorem 4.2.11 *[264, Theorem 9] Suppose $\{x_t\}$ is an aperiodic, irreducible Markov chain with invariant distribution π. Suppose C is a $(1,\epsilon,\nu)$-small set and $V : \mathbb{X} \rightarrow [1,\infty)$ satisfies the univariate drift condition with constants $\lambda \in (0,1)$ and $b < \infty$. Then $\{x_t\}$ is geometrically ergodic.*

That geometric ergodicity follows from the univariate drift condition with a small set C is proven by Roberts and Rosenthal by using the coupling inequality to bound the TV-norm, but an alternate proof is given by Meyn and Tweedie [233] resulting in the following theorem.

Theorem 4.2.12 *[233, Theorem 15.0.1] Suppose $\{x_t\}$ is an aperiodic and irreducible Markov chain. Then the following are equivalent:*

(i) $E_x[\tau_B] < \infty$ for all $x \in \mathbb{X}$, $B \in \mathcal{B}^+(\mathbb{X})$, the invariant distribution π of $\{x_t\}$ exists and there exists a petite set C, constants $\gamma < 1$, $M > 0$ such that for all $x \in C$

$$|P^n(x,C) - \pi(C)| < M\gamma^n.$$

(ii) For a petite set C and for some $\kappa > 1$

$$\sup_{x\in C} E_x[\kappa^{\tau_C}] < \infty.$$

(iii) For a petite set C, constants $b > 0$ $\lambda \in (0,1)$, and a function $V : \mathbb{X} \rightarrow [1,\infty]$ (finite for some x) such that

$$PV \leq \lambda V + b1_C.$$

Any of the conditions imply that there exists $r > 1$, $R < \infty$ such that for any x

$$\sum_{n=0}^{\infty} r^n \|P^n(x,\,\cdot\,) - \pi(\,\cdot\,)\|_V \le RV(x).$$

We note that if (iii) above holds, (ii) holds for for all $\kappa \in (1, \lambda^{-1})$.

We now show that under (4.29), Theorem 4.2.12 (ii) holds. If (4.29) holds, the sequence $\{M_n\}$ is supermartingale (with respect to the natural filtration), where

$$M_n = \lambda^{-n} V(x_n) - \sum_{k=0}^{n-1} b 1_C(x_k) \lambda^{-(k+1)},$$

with $M_0 = V(x_0)$. Then, with (4.29), defining $\tau_B^N = \min\{N, \tau_B\}$ for $B \in \mathcal{B}^+(\mathbb{X})$ gives, by Doob's optional sampling theorem,

$$E_x\left[\lambda^{-\tau_B^N} V(x_{\tau_B^N})\right] \le V(x) + E_x\left[\sum_{n=0}^{\tau_B^N - 1} b 1_C(x_n) \lambda^{-(n+1)}\right] \tag{4.30}$$

for any $B \in \mathcal{B}^+(\mathbb{X})$, and $N \in \mathbb{Z}_+$.

Since V is bounded above on C, we have that $C \subset \{V \le L_1\}$ for some L_1 and thus,

$$\sup_{x \in C} E_x\left[\lambda^{-\tau_C^N} V(x_{\tau_C^N})\right] \le L_1 + \lambda^{-1} b.$$

and by the monotone convergence theorem, and the fact that V is bounded from below by 1 everywhere and bounded from above on C,

$$\sup_{x \in C} E_x\left[\lambda^{-\tau_C}\right] \le L_1(L_1 + \lambda^{-1} b).$$

Using the coupling inequality, Roberts and Rosenthal [264] prove that geometric ergodicity follows from the univariate drift condition. They show that under mild conditions [264, Prop. 11], the univariate drift condition implies a drift condition for the pair of Markov chains who will be coupled in the small set $C \times C$:

Proposition 4.2.1 *[264, Proposition 11] Suppose the univariate drift condition (4.29) is satisfied for $V : \mathbb{X} \to [1, \infty)$ and constants $\lambda \in (0,1)$ $b < \infty$ and small set C. Letting $d = \inf_{x \in C^C} V(x)$, if $d > \frac{b}{1-\lambda} - 1$, then the bivariate drift condition is satisfied for $h(x,y) = \frac{1}{2}(V(x) + V(y))$ and $\alpha^{-1} = \lambda + b/(d+1) < 1$; that is, we have the following condition.*

$$\begin{aligned} \bar{P}h(x,y) \le & \frac{h(x,y)}{\alpha} \quad (x,y) \notin C \times C \\ \bar{P}h(x,y) < & \infty \quad (x,y) \in C \times C \end{aligned}$$

where

$$\bar{P}h(x,y) = \int_{\mathbb{X}} \int_{\mathbb{X}} h(z,w) P(x,dz) P(y,dw)$$

But now if one applies Theorem 4.2.12 (ii), the desired coupling condition and hence the convergence rate result will follow.

We also note that the univariate drift condition allows us to assume that V is bounded on C without any loss (see Lemma 14 of [264]).

Subgeometric ergodicity

Here, we review the class of subgeometric rate functions (see [154, Sec. 4], [89, Sec. 5], [233], [108], [308]).

Let Λ_0 be the family of functions $r : \mathbb{N} \to \mathbb{R}_{>0}$ such that

$$r \text{ is non-decreasing}, \quad r(1) \geq 2$$

and

$$\frac{\log r(n)}{n} \downarrow 0 \quad \text{as } n \to \infty$$

The second condition implies that for all $r \in \Lambda_0$ if $n > m > 0$ then

$$n \log r(n+m) \leq n \log r(n) + m \log r(n) \leq n \log r(n) + n \log r(m)$$

so that

$$r(m+n) \leq r(m) r(n) \quad \text{for all } m, n \in \mathbb{N}. \tag{4.31}$$

The class of subgeometric rate functions Λ defined in [308] is the class of sequences r for which there exists a sequence $r_0 \in \Lambda_0$ such that

$$0 < \liminf_{n\to\infty} \frac{r(n)}{r_0(n)} \leq \limsup_{n\to\infty} \frac{r(n)}{r_0(n)} < \infty.$$

The main theorem we cite on subgeometric rates of convergence is due to Tuominen and Tweedie [308].

Theorem 4.2.13 *[308, Theorem 2.1] Suppose that $\{x_t\}_{t\in\mathbb{N}}$ is an irreducible and aperiodic Markov chain on state space $\mathbb{X}$ with stationary transition probabilities given by P. Let $f : \mathbb{X} \to [1, \infty)$ and $r \in \Lambda$ be given. The following are equivalent:*

(i) there exists a petite set $C \in \mathcal{B}(\mathbb{X})$ such that

$$\sup_{x\in C} E_x[\sum_{k=0}^{\tau_C - 1} r(k) f(x_k)] < \infty$$

(ii) there exists a sequence (V_n) of functions $V_n : \mathbb{X} \to [0, \infty]$, a petite set $C \in \mathcal{B}(\mathbb{X})$ and $b \in \mathbb{R}_+$ such that V_0 is bounded on C,

$$V_0(x) = \infty \Rightarrow V_1(x) = \infty,$$

and

$$PV_{n+1} \leq V_n - r(n) f + b r(n) 1_C, \quad n \in \mathbb{N}$$

(iii) there exists an (f, r)-regular set $A \in \mathcal{B}^+(\mathbb{X})$.

(iv) there exists a full absorbing set S which can be covered by a countable number of (f, r)-regular sets.

Theorem 4.2.14 *[308] If a Markov chain $\{x_t\}$ satisfies Theorem 4.2.13 for (f, r) then $r(n)\|P^n(x_0, \cdot) - \pi(\cdot)\|_f \to 0$ as n increases*

The conditions of Theorem 4.2.13 may be hard to check, especially (ii), comparing a sequence of Lyapunov functions $\{V_k\}$ at each time step. We briefly discuss the methods of Douc et al. [108] (see also Hairer [154]) that extend the subgeometric ergodicity results and show how to construct subgeometric rates of ergodicity from a simpler drift condition. [108] assumes that there exists a function $V : \mathbb{X} \to [1, \infty]$, a concave monotone nondecreasing differentiable function $\phi : [1, \infty] \to (0, \infty]$, a set $C \in \mathcal{B}(\mathbb{X})$ and a constant $b \in \mathbb{R}$ such that

$$PV + \phi \circ V \leq V + b 1_C. \tag{4.32}$$

If an aperiodic and irreducible Markov chain $\{x_t\}$ satisfies the above with a petite set C, and if $V(x_0) < \infty$, then it can be shown that $\{x_t\}$ satisfies Theorem 4.2.13(ii). Therefore $\{x_t\}$ has invariant distribution π and is $(\phi \circ V, 1)$-ergodic so that $\lim_{n\to\infty} \|P^n(x, \cdot) - \pi(\cdot)\|_{\phi\circ V} = 0$ for all x in the set $\{x : V(x) < \infty\}$ of π-measure 1. The results by Douc et al. build then on trading off $(\phi \circ V, 1)$ ergodicity for $(1, r_\phi)$-ergodicity for some rate function r_ϕ, by carefully constructing the function utilizing concavity; see Propositions 2.1 and 2.5 of [108] and Theorem 4.1(3) of [154].

To achieve ergodicity with a nontrivial rate and norm one can invoke a result involving the class of *pairs of ultimately non decreasing functions*, defined in [108]. The class $\mathcal{Y}$ of pairs of ultimately non decreasing functions consists of pairs $\Psi_1, \Psi_2 : \mathbb{X} \to [1, \infty)$ such that $\Psi_1(x)\Psi_2(y) \leq x + y$ and $\Psi_i(x) \to \infty$ for one of $i = 1, 2$.

Proposition 4.2.2 *Suppose $\{x_t\}$ is an aperiodic and irreducible Markov chain that is both $(1, r)$-ergodic and $(f, 1)$-ergodic for some $r \in \Lambda$ and $f : \mathbb{X} \to [1, \infty)$. Suppose $\Psi_1, \Psi_2 : \mathbb{X} \to [1, \infty)$ are a pair of ultimately non decreasing functions. Then $\{x_t\}$ is $(\Psi_1 \circ f, \Psi_2 \circ r)$-ergodic.*

Therefore we can show that if $(\Psi_1, \Psi_2) \in \mathcal{Y}$ and a Markov chain satisfies the condition (4.32), then it is $(\Psi_1 \circ \phi \circ V, \Psi_2 \circ r_\phi)$-ergodic.

Thus, we observe that the hitting times to a small set is an important random variable in characterizing not only the existence of an invariant probability measure, but also how fast a Markov chain converges to equilibrium. Further results exist in the literature to obtain more *computable* criteria for subgeometric rates of convergence, see e.g. [108].

Rates of convergence under random-time state-dependent drift criteria

The following result builds on and generalizes Theorem 2.1 in [350].

Theorem 4.2.15 *[354] Let $\{x_t\}$ be an aperiodic and irreducible Markov chain with a small set C. Suppose there are functions $V : \mathbb{X} \to (0, \infty)$ with V bounded on C, $f : \mathbb{X} \to [1, \infty), \delta : \mathbb{X} \to [1, \infty)$, a constant $b \in \mathbb{R}$, and $r \in \Lambda$ such that for a sequence of stopping times $\{\tau_n\}$*

$$E[V(x_{\tau_{n+1}}) \mid x_{\tau_n}] \leq V(x_{\tau_n}) - \delta(x_{\tau_n}) + b1_C(x_{\tau_n})$$
$$E\Big[\sum_{k=\tau_n}^{\tau_{n+1}-1} f(x_k)r(k) \,\Big|\, \mathcal{F}_{\tau_n}\Big] \leq \delta(x_{\tau_n}). \tag{4.33}$$

Then $\{x_t\}$ satisfies Theorem 4.2.13 and is (f, r)-ergodic.

Further conditions and examples are available in [354].

4.3 Applications to Stochastic Learning Algorithms and Iterative Dynamics

In this section, we present a number of applications of martingale theory to the analysis of stochastic dynamics, which will have applications to stochastic learning and reinforcement learning results to be studied later in the book.

We start with a convergence theorem useful in stochastic approximation]

Theorem 4.3.1 *[243, p. 33, Exercise II-4] Let X_k, β_k, Y_k be three sequences of non-negative random variables defined on a common probability space and $\mathcal{F}_k$ be a filtration so that all three random sequences are adapted to it. Suppose that*

$$E[X_{k+1}|\mathcal{F}_k] \leq (1 + \beta_k)X_k + Y_k, \quad k \in \mathbb{N}.$$

Then, limit $\lim_{n\to\infty} X_n$ exists and is finite with probability one conditioned on the event that $\sum_{k\in\mathbb{N}} \beta_n < \infty$ and $\sum_{k\in\mathbb{N}} Y_n < \infty$.

Proof sketch. (i) Define

$$M_n = X'_n - \sum_{m=1}^{n-1} Y'_m,$$

where $X'_n = \frac{X_n}{\prod_{s=1}^{n-1}(1+\beta_s)}$ and $Y'_n = \frac{Y_n}{\prod_{s=1}^{n}(1+\beta_s)}$. Define the stopping time:

$$\tau_a = \min(n : \sum_{m=1}^{n-1} Y'_m > a).$$

(ii) Show first that $a + M_{\min(\tau_a,n)}, \quad n \in \mathbb{N}$ is a positive supermartingale (iii) Thus, for any fixed a, $a + M_{\min(\tau_a,n)}, n \in \mathbb{N}$ converges to a limit. For a given sample path (almost surely), taking a sufficiently large a, by the boundedness of $\sum_{k\in\mathbb{N}} Y_n$ show that $\tau_a = \infty$ for sufficiently large a for this given sample path. (iv) Then invoke the supermartingale convergence theorem 4.1.9. Finally, using the fact that $\sum_{k\in\mathbb{N}} Y_k < \infty$ and the implication (from $\sum_{k\in\mathbb{N}} \beta_n < \infty$) that $\prod_n (1+\beta_n) < \infty$ under the stated conditions, complete the proof. ◇

This theorem is important for a large class of optimization problems (such as the convergence of stochastic gradient descent algorithms) as well as stochastic approximation algorithms. For further reading on stochastic approximation methods, see [209] and [34] and for a recent review [319]. We will use this result to establish the convergence of the celebrated Q-learning algorithm in Theorem 9.1.1. A slight generalization of this result appears in [263].

Theorem 4.3.2 *[Another convergence theorem useful in stochastic approximation and Q-Learning convergence analysis] Let X_k, Y_k, Z_k be three sequences of non-negative random variables defined on a common probability space and $\mathcal{F}_k$ be a filtration so that all three random sequences are adapted to it. Suppose that*

$$E[Y_{k+1}|\mathcal{F}_k] \leq Y_k - X_k + Z_k \tag{4.34}$$

and $\sum_k Z_k < \infty$. Then, $\sum_k X_k < \infty$ and Y_k converges to some random variable Y almost surely.

Proof sketch [341].: Apply Theorem 4.3.1 by noting first that $E[Y_{k+1}|\mathcal{F}_k] \leq Y_k + Z_k$ with $\beta_k = 0$. This implies that Y_k converges. Now write $M_t = Y_t + \sum_{m=1}^{t-1} X_m$ leading to $E[M_{t+1}|\mathcal{F}_t] \leq M_t + Z_t$. Applying Theorem 4.3.1 again, it follows that M_t converges and since Y_t converges, so does $\sum_t X_t$. ◇

We now apply the above to an explicit iterative stochastic dynamics:

Theorem 4.3.3 *[38, Corollary 4.1] Consider the following: Let r_t be a scalar and*

$$r_{t+1} = (1 - \alpha_t)r_t + \alpha_t w_t,$$

where $\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty$, and the noise w_t is so that $E[w_t|\mathcal{F}_{t-1}] = 0$ with

$$E[w_t^2|\mathcal{F}_t] \leq A_t,$$

where A_t is possibly a random variable (thus, sample path dependent). If A_t is bounded with probability 1 (that is, $\sup_{t\in\mathbb{Z}_+} |A_t(\omega)| < \infty$ almost surely), then $r_t \to 0$ almost surely.

We remark that if the bounded random variable sequence A_t above was instead a fixed number, the proof would have been slightly more direct.

Proof. Take $Y_t = r_t^2$ and apply Theorem 4.3.2. ◇

We end the section, with a final application[2]:

[2]Thanks to Prof. Jerome Le Ny (of Polytechnique Montreal).

Theorem 4.3.4 *[Application in stochastic optimization: Stochastic gradient descent] Consider a convex function $f : \mathbb{R}^n \to \mathbb{R}$ and denote the set of minima of f by X^*. We know from convex analysis that X^* contains, if non-empty, either a single point, or is a convex set. Denote the subdifferential [70] of f at x, that is the set of subgradients of f at x, by $\partial f(x)$ and let d_t be a random variable which is a noisy version of a sub gradient of f at x_t at time t. A stochastic subgradient algorithm is one with the form:*

$$x_{k+1} = x_k - \gamma_k d_{k+1}, \quad x_0 \in \mathbb{R}^n, \tag{4.35}$$

where γ_k is a sequence of non-negative step sizes. We have the following theorem:

Suppose that the set of minima X^* is non-empty and that the stochastic subgradients satisfy that

$$\sup_k E[||d_{k+1}||^2|\mathcal{F}_k] < K < \infty.$$

where $\mathcal{F}_k = \sigma(x_0, d_s, s \leq k)$ with the condition that

$$g_{k+1} = E[d_{k+1}|\mathcal{F}_k] \in \partial f(x).$$

Moreover, $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$. Then the sequence of iterates (4.35) converges almost surely to some element $x^* \in X^*$.

Proof. $y \in \mathbb{R}^n$, we have that due to the definition of a subgradient,

$$f(y) \geq f(x) + g_{k+1}^T(y - x_k). \tag{4.36}$$

Thus,

$$\begin{aligned}
&E[||x_{k+1} - y||^2|\mathcal{F}_k] = E[||x_k - \gamma_k d_{k+1} - y||^2|\mathcal{F}_k] \\
&= E[||x_k - y||^2 - 2\gamma_k(x_k - y)^T d_{k+1} + \gamma_k^2||d_{k+1}||^2|\mathcal{F}_k] \\
&= E[||x_k - y||^2|\mathcal{F}_k] - 2\gamma_k(x_k - y)^T E[d_{k+1}|\mathcal{F}_k] + E\gamma_k^2||d_{k+1}||^2|\mathcal{F}_k] \\
&\leq E[||x_k - y||^2|\mathcal{F}_k] - 2\gamma_k(f(x_k) - f(y)) + \gamma_k^2 K
\end{aligned}$$

where in the inequality we use (4.35). Now, let in the above $y = \bar{x}^* \in X^*$ for some element in X^*. Then one obtains through the comparison theorem (Theorem 4.2.3) that

$$E[\sum_k \gamma_k(f(x_k) - f(y))] \leq ||x_0 - y||^2 + \sum_k \gamma_k^2 K.$$

In particular, since $f(x_k) - f(y) \geq 0$, through the convergence theorem from the preceding exercises we have that almost surely

$$\sum_k \gamma_k(f(x_k) - f(y)) < \infty.$$

Thus, $f(x_k) \to f(y)$. We now show that indeed $x_k \to$ some particular element in X^* (and does not wander in the set). By the convergence result in (4.34) we know that for any $x^* \in X^*$, $||x_k - x^*||$ converges almost surely. This implies that x_k is bounded almost surely. Now, consider a countable dense subset $\{x^{1,*}, \cdots, x^{n,*}, \cdots\}$ of X^*. It must be that $||x_k - x^{i,*}||$ converges for all i through the convergence theorem. On the other hand, since $||x_k||$ is bounded, there exists a converging subsequence for x_{k_n}. But the limit of each such subsequence must be identical for otherwise $||x_{k_n} - x^{i,*}||$ would have different limits. Thus, x_{k_n} must converge to one element in X^*. ◇

4.4 Conclusion

This concludes our discussion on martingales, their applications to controlled Markov chains, ergodic theorems, as well as stochastic iterative dynamics. We will revisit one more application of martingales while discussing the convex analytic

approach to controlled Markov problems. Notably, we have observed that drift criteria are very powerful tools to establish various forms of stochastic stability and instability.

4.5 Exercises

Exercise 4.5.1 *Let X be an integrable random variable defined on $(\Omega, \mathcal{F}, P)$. Let $\mathcal{G} = \{\Omega, \emptyset\}$. Show that $E[X|\mathcal{G}] = E[X]$, and if $\mathcal{G} = \sigma(X)$ then $E[X|\mathcal{G}] = X$.*

Exercise 4.5.2 *Consider (4.5) and through this relation establish a sufficient condition on the martingale sequence X_n so that the optimal sampling theorem would be applicable even if the stopping times in Theorem 4.1.5 would not necessarily be bounded from above by a deterministic constant.*

Exercise 4.5.3 *A useful property of martingales is that a stopped martingale is a martingale. This is very useful for proving stability results when one lives in a bounded set since the stopped martingale sequence will typically be uniformly bounded (and hence the optional sampling theorem will be applicable without requiring a stopping time to be uniformly bounded). Let τ be a stopping time that is finite almost surely. Let $X_t, \mathcal{F}_t$ be a martingale sequence. Define $M_t = X_{\min(t,\tau)}$. Show that $(M_t, \mathcal{F}_t)$ is a martingale sequence:*

$$E[M_{n+1}|\mathcal{F}_n] = M_n$$

Hint: Write $M_{n+1} = M_n + 1_{\{\tau>n\}}(M_{n+1} - M_n)$. Then, show that $E[1_{\{\tau>n\}}(M_{n+1} - M_n)|\mathcal{F}_n] = 1_{\{\tau>n\}}E[M_{n+1} - M_n|\mathcal{F}_n] = 0$.

Exercise 4.5.4 *a) Consider a Controlled Markov Chain with the following dynamics:*

$$x_{t+1} = ax_t + bu_t + w_t,$$

where w_t is a zero-mean Gaussian noise with a finite variance, $a, b \in \mathbb{R}$, $b \neq 0$, are the system dynamics coefficients. One controller policy which is admissible (that is, the policy at time t is measurable with respect to $\sigma(x_0, x_1, \ldots, x_t)$ and is a mapping to $\mathbb{R}$) is the following:

$$u_t = -\frac{a + 0.5}{b} x_t.$$

Show that $\{x_t\}$, under this policy, has a unique invariant probability measure.

b) Consider a similar setup to the one earlier, with $b = 1$:

$$x_{t+1} = ax_t + u_t + w_t,$$

where w_t is a zero-mean Gaussian noise with a finite variance, and $a \in \mathbb{R}$ is a known number.

This time, suppose, we would like to find a control policy such that there exists an invariant probability measure π for $\{x_t\}$ and under this invariant probability measure

$$E_\pi[x^2] < \infty$$

Further, suppose we restrict the set of control policies to be linear, time-invariant; that is of the form $u(x_t) = kx_t$ for some $k \in \mathbb{R}$.

Find the set of all k values for which there exists an invariant probability measure that has a finite second moment.

Hint: Use Foster-Lyapunov criteria.

Exercise 4.5.5 *Suppose that some price process $\{x_t, t \in \mathbb{Z}_+\}$ is given by the following dynamics:*

$$x_{t+1} = \max(x_t + w_t, 0), \quad t \in \mathbb{Z}_+,$$

where $\{w_t\}$ is a sequence of independent and identically distributed $\{-1,1\}$-valued random variables with mean $\bar{w} > 0$. Furthermore, $x_0 \in \mathbb{Z}_+, x_0 > 0$ is a given initial condition for the process.

Is the price process recurrent in the sense that, $P_{x_0}(\tau_0 < \infty) = 1$, where $\tau_0 = \min(l > 0 : x_l = 0)$?

Exercise 4.5.6 *Consider a queuing process, with i.i.d. Poisson arrivals and departures, with arrival mean μ and service mean λ and suppose the process is such that when a customer leaves the queue, with probability p (independent of time) it comes back to the queue. That is, the dynamics of the system satisfies:*

$$L_{t+1} = \max(L_t + A_t - N_t + p_t N_t, 0), \quad t \in \mathbb{N}.$$

where $E[A_t] = \lambda, E[N_t] = \mu$ and $E[p_t] = p$.

For what values of μ, λ is such a system stochastically stable? Prove your statement.

Exercise 4.5.7 *Consider a two server-station network; where a router routes the incoming traffic, as is depicted in Figure 5.1.*

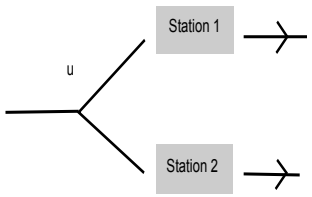

Fig. 4.1

Let L_t^1, L_t^2 denote the number of customers in stations 1 and 2 at time t. Let the dynamics be given by the following:

$$L_{t+1}^1 = \max(L_t^1 + u_t A_t - N_t^1, 0), \quad t \in \mathbb{N}.$$

$$L_{t+1}^2 = \max(L_t^2 + (1-u_t) A_t - N_t^2, 0), \quad t \in \mathbb{N}.$$

Customers arrive according to an independent Bernoulli process, A_t, with mean λ. That is, $P(A_t = 1) = \lambda$ and $P(A_t = 0) = 1 - \lambda$. Here $u_t \in [0,1]$ is the router action.

Station 1 has a Bernoulli service process N_t^1 with mean n_1, and Station 2 with n_2.

Suppose that a router decides to follow the following algorithm to decide on u_t: If a customer arrives, the router simply sends the incoming customer to the shortest queue.

Find sufficient conditions (on λ, n_1, n_2) for this algorithm to lead to a stochastically stable system with invariant measure π which satisfies $E_\pi[L^1 + L^2] < \infty$.

Note: For this problem, we acknowledge the lecture notes of Prof. Bruce Hajek: ECE567 Communication Network Analysis, University of Illinois at Urbana-Champaign [157].

Exercise 4.5.8 *Consider the following two-server system:*

$$\begin{aligned} x_{t+1}^1 &= \max(x_t^1 + A_t^1 - u_t^1, 0) \\ x_{t+1}^2 &= \max(x_t^2 + A_t^2 + u_t^1 1_{(u_t^1 \le x_t^1 + A_t^1)} - u_t^2, 0), \end{aligned} \tag{4.37}$$

where $1_{(.)}$ denotes the indicator function and A_t^1, A_t^2 are independent and identically distributed (i.i.d.) random variables with geometric distributions, that is, for $i = 1, 2$,

$$P(A_t^i = k) = p_i(1-p_i)^k \quad k \in \{0,1,2,\dots,\},$$

for some scalars p_1, p_2 such that $E[A_t^1] = 1.5$ and $E[A_t^2] = 1$.

Suppose the control actions u_t^1, u_t^2 are such that $u_t^1 + u_t^2 \leq 5$ for all $t \in \mathbb{Z}_+$ and $u_t^1, u_t^2 \in \mathbb{Z}_+$. At any given time t, the controller has to decide on u_t^1 and u_t^2 with knowing $\{x_s^1, x_s^2, s \leq t\}$ but not knowing A_t^1, A_t^2.

Is this server system stochastically stabilizable by some policy, that is, does there exist an invariant probability measure under some control policy?

If your answer is positive, provide a control policy and show that there exists a unique invariant distribution.

Exercise 4.5.9 *Let there be a single server, serving two queues; where the server serves the two queues adaptively in the following sense. The dynamics of the two queues is expressed as follows:*

$$L_{t+1}^i = \max(L_t^i + A_t^i - N_t^i, 0), \quad i = 1,2; \quad t \in \mathbb{Z}_+$$

where L_t^i is the total number of arrivals which are still in the queue at time t and A_t^i is the number of customers that have just arrived at time t.

We assume, for $i = 1,2$, $\{A_t^i\}$ has an independent and identical distribution (i.i.d.) which is Bernoulli so that $P(A_t^i = 1) = \lambda_i = 1 - P(A_t^i = 0)$.

Suppose that the service process is given by:

$$N_t^1 = 1_{\{L_t^1 \geq L_t^2\}} \qquad N_t^2 = 1_{\{L_t^2 > L_t^1\}}$$

For what values of λ_1, λ_2 is the system stochastically stable, in the sense of the existence of an invariant probability measure.

Exercise 4.5.10 *Let X be a real random variable with $E[|X|] < \infty$. Let $Y_0, Y_1, Y_2, \cdots$ be a sequence of random variables. Let $\mathcal{F}_n$ be the σ-field generated by $Y_0, Y_1, \dots, Y_n$. a) Is it the case that*

$$\lim_{n\to\infty} E[X|\mathcal{F}_n]$$

exists? b) Is it the case that

$$\lim_{n\to\infty} E[X|\mathcal{F}_n] = E[X|\mathcal{F}_\infty],$$

where $F_\infty := \sigma(Y_1, Y_2, \cdots)$

Exercise 4.5.11 *Prove the Ergodic Theorem for a finite state space Markov chain; that is the result that for an irreducible Markov chain $\{x_t\}$ living in a finite space $\mathbb{X}$, which has a unique invariant probability measure μ, the following applies almost surely:*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(x_t) = \sum_i f(i)\mu(i),$$

for every $f : \mathbb{X} \to \mathbb{R}$.

Hint: You may proceed as follows. Define a sequence of empirical occupation measures for $T \in \mathbb{N}$, $A \in \mathcal{B}(\mathbb{X})$:

$$v_T(A) = \frac{1}{T} \sum_{t=0}^{T-1} 1_{\{X_t \in A\}}, \quad \forall A \in \mathcal{B}(\mathbb{X}).$$

Now, define:

$$F_t(A) = \left(\sum_{s=1}^{t} 1_{\{X_s \in A\}} - t \sum_{\mathbb{X}} P(A|x) v_t(x) \right)$$

$$= \left(\sum_{s=1}^{t} 1_{\{X_s \in A\}} - \sum_{s=0}^{t-1} \sum_{\mathbb{X}} P(A|x) 1_{\{X_s = x\}} \right) \tag{4.38}$$

Let $\mathcal{F}_t = \sigma(X_0, \cdots, X_t)$. Verify that, for $t \geq 2$,

$$\begin{aligned}
& E[F_t(A)|\mathcal{F}_{t-1}] \\
&= E\left[\left(\sum_{s=1}^{t} 1_{\{X_s \in A\}} - \sum_{s=0}^{t-1} \sum_{\mathbb{X}} P(A|x) 1_{\{X_s = x\}} \right) \middle| \mathcal{F}_{t-1} \right] \\
&= E\left[\left(1_{\{X_t \in A\}} - \sum_{\mathbb{X}} P(A|x) 1_{\{X_{t-1} = x\}} \right) \middle| \mathcal{F}_{t-1} \right] \\
&\quad + \left(\sum_{s=1}^{t-1} 1_{\{X_s \in A\}} - \sum_{s=0}^{t-2} \sum_{\mathbb{X}} P(A|x) 1_{\{X_s = x\}} \right) \\
&= 0 + \left(\sum_{s=1}^{t-1} 1_{\{X_s \in A\}} - \sum_{s=0}^{t-2} \sum_{\mathbb{X}} P(A|x) 1_{\{X_s = x\}} \right) \middle| \mathcal{F}_{t-1} \Bigg] \qquad (4.39) \\
&= F_{t-1}(A), \qquad (4.40)
\end{aligned}$$

where the last equality follows from the fact that $E[1_{\{X_t \in A\}}|\mathcal{F}_{t-1}] = P(X_t \in A|\mathcal{F}_{t-1})$. Furthermore,

$$|F_t(A) - F_{t-1}(A)| \leq 1.$$

Now, we have a sequence which is a martingale sequence. We will invoke a martingale convergence theorem; which is applicable for **martingales with bounded increments***. By a version of the martingale stability theorem, it follows that*

$$\lim_{t \to \infty} \frac{1}{t} F_t(A) = 0.$$

You need to now complete the remaining steps.

Hint: You can use the Azuma-Hoeffding inequality (Theorem 4.1.15) [98] and the Borel-Cantelli Lemma to complete the steps.

We note that a similar argument could also be made for countably infinite $\mathbb{X}$ or uncountable $\mathbb{X}$ under additional conditions.

Exercise 4.5.12 *Let τ be a stopping time with respect to the filtration $\mathcal{F}_t$. Let X_n be a (discrete-time) sequence of random variables so that each X_n is $\mathcal{F}_n$-measurable. Show that X_τ is $\mathcal{F}_\tau$-measurable.*
Hint: We need to show that for every real a: $\{X_\tau \leq a\} \cap \{\tau \leq k\} \in \mathcal{F}_k$. Observe that $\{X_\tau \leq a\} \cap \{\tau \leq k\} = \cup_{m=0}^{k} \{X_\tau \leq a\} \cap \{\tau = m\}$ and that for each m, $\{X_\tau \leq a\} \cap \{\tau = m\} \in \mathcal{F}_m \subset \mathcal{F}_k$.

Exercise 4.5.13 *To appreciate that the condition of measurability on control or estimation policies is not a superfluous one, read the paper [324]: G. L. Wise. A note on a common misconception in estimation. Systems & Control letters, 1985: 355-356.*

5

Optimal Stochastic Control with Finite and Discounted Infinite Horizons and Dynamic Programming

In this chapter, we introduce the method of *dynamic programming* for controlled stochastic systems, and consider optimal stochastic control problems under finite horizon and discounted infinite horizon expected cost criteria.

Recall that a fully observed Markov control model is a five-tuple

$$(\mathbb{X}, \mathbb{U}, \{\mathbb{U}(x), x \in \mathbb{X}\}, \mathcal{T}, c)$$

such that $\mathbb{X}$ is the (standard Borel) state space, $\mathbb{U}$ is the action space, $\mathbb{U}(x) \subset \mathbb{U}$ is the control action set when the state is x, so that

$$\mathbb{K} = \{(x, u) : x \in \mathbb{X}, u \in \mathbb{U}(x)\} \subset \mathbb{X} \times \mathbb{U},$$

is the set of feasible state-action pairs. $\mathcal{T}$ is a stochastic kernel on $\mathbb{X}$ given $\mathbb{K}$. Finally $c : \mathbb{K} \to \mathbb{R}$ is the cost function.

One also can have a dependence of the cost function on the time variable so that c_t can be the cost at time t or the action set $\mathbb{U}(x)$ can also depend on time. In this case one can add the time variable t, as a further component, to the state variable x, with a deterministic evolution for the time variable. Conceptually, such a generalization does not introduce any further obstacles for finite horizon problems. Often, $c_t \equiv c$, that is c does not depend on time (however, there may be a terminal cost different from c, to be considered).

Let, as in Section 2.2.1, Γ_A denote the set of all admissible policies. Let $\gamma = \{\gamma_t, 0 \leq t \leq N-1\} \in \Gamma_A$ be a policy. Consider the following expected cost:

$$J_{\mathcal{N}}(x, \gamma) := E_x^{\gamma}[\sum_{t=0}^{N-1} c(x_t, u_t) + c_N(x_N)], \tag{5.1}$$

where $c_N(.)$ is the terminal cost function. Define

$$J_{\mathcal{N}}^*(x) := \inf_{\gamma \in \Gamma_A} J_{\mathcal{N}}(x, \gamma)$$

5.1 Dynamic Programming, Optimality of Markov Policies and Bellman's Principle of Optimality

The goal is to find, if there exists one, an admissible policy such that $J_{\mathcal{N}}^*(x)$ is attained; this will be an optimal policy. We note that the infimum value, in general, may not be attained by some policy. In the following, we will also present conditions which will ensure the existence of optimal policies.

5.1.1 Backwards Induction

Let $h_t = \{x_{[0,t]}, u_{[0,t-1]}\}$ denote the history process for $t \in \mathbb{N}$. By Theorem 4.1.3, provided that the cost is integrable under the induced probability measure given a policy, we note that the cost can be expressed as:

$$\begin{aligned}
J_{\mathcal{N}}(x,\gamma) = E_x^{\gamma}\Big[&c(x_0,u_0)\\
&+E^{\gamma}\Big[c(x_1,u_1)\\
&\quad+E^{\gamma}\Big[c(x_2,u_2)\\
&\quad+\dots\\
&\qquad+E^{\gamma}[c(x_{N-1},u_{N-1})+c_N(x_N)|h_{N-1}]\Big|h_{N-2}\Big]\dots\Big|h_1\Big]\Big|x_0=x\Big],\\
= E_x^{\gamma_0,\cdots,\gamma_{N-1}}\Big[&c(x_0,u_0)\\
&+E^{\gamma_1,\cdots,\gamma_{N-1}}\Big[c(x_1,u_1)\\
&\quad+E^{\gamma_2,\cdots,\gamma_{N-1}}\Big[c(x_2,u_2)\\
&\quad+\dots\\
&\qquad+E^{\gamma_{N-1}}[c(x_{N-1},u_{N-1})+c_N(x_N)|h_{N-1}]\Big|h_{N-2}\Big]\dots\Big|h_1\Big]\Big|x_0=x\Big],\\
= E_x^{\gamma_0}\Big[&c(x_0,u_0)\\
&+E^{\gamma_1}\Big[c(x_1,u_1)\\
&\quad+E^{\gamma_2}\Big[c(x_2,u_2)\\
&\quad+\dots\\
&\qquad+E^{\gamma_{N-1}}[c(x_{N-1},u_{N-1})+c_N(x_N)|h_{N-1}]\Big|h_{N-2}\Big]\dots\Big|h_1\Big]\Big|x_0=x\Big],
\end{aligned}$$

Thus, by the equalities above, we obtain:

$$\begin{aligned}
\inf_{\gamma\in\Gamma_A} J_{\mathcal{N}}(x,\gamma) = \inf_{\gamma_0} E_x^{\gamma_0}\Big[&c(x_0,u_0)\\
&+\inf_{\gamma_1} E^{\gamma_1}\Big[c(x_1,u_1)\\
&\quad+\inf_{\gamma_2} E^{\gamma_2}\Big[c(x_2,u_2)\\
&\quad+\dots\\
&\qquad+\inf_{\gamma_{N-1}} E^{\gamma_{N-1}}[c(x_{N-1},u_{N-1})+c_N(x_N)|h_{N-1}]\Big|h_{N-2}\Big]\dots\Big|h_1\Big]\Big|x_0=x\Big],
\end{aligned}\tag{5.2}$$

The discussion above reveals that we can start with the final time stage, obtain a solution for γ_{N-1} and move backwards for $t \leq N-2$. To this end, we present a critical supporting result in the following.

5.1.2 Optimality of Deterministic Markov Policies

We will observe that when there is an optimal solution, the optimal solution can be taken to be Markov. Even when an optimal policy may not exist, any measurable policy can be replaced with one which is Markov, under fairly general conditions, as we discuss below. In the following, first, we will follow David Blackwell's [48] and Hans Witsenhausen's [330] ideas to obtain a very interesting result.

Theorem 5.1.1 (Blackwell's Theorem on Redundancy of Information beyond the State) *Let $\mathbb{X}, \mathbb{Y}, \mathbb{U}$ be complete, separable, metric spaces, and let P be a probability measure on $\mathcal{B}(\mathbb{X} \times \mathbb{Y})$, and let $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ be a Borel measurable and bounded cost function. Then, for any Borel measurable function $\gamma : \mathbb{X} \times \mathbb{Y} \to \mathbb{U}$, there exists another Borel measurable function $\gamma^* : \mathbb{X} \to \mathbb{U}$ such that*

$$\int_{\mathbb{X}} c(x, \gamma^*(x)) P_{\mathbb{X}}(dx) \leq \int_{\mathbb{X}\times\mathbb{Y}} c(x, \gamma(x, y)) P(dx, dy)$$

where $P_{\mathbb{X}}$ is the marginal of P on $\mathbb{X}$. Thus, policies based only on x almost surely, are optimal.

Proof. We will construct a γ^* given γ. Let $u = \gamma(x, y)$. To emphasize the random nature of the variables considered, let us again denote with capital letters X, Y, U the random variables whose realizations are x, y and u, respectively. Given γ, we write for any Borel $D \subset \mathbb{U}$ and $x \in \mathbb{X}$,

$$P^{\gamma}(U \in D|x) = P(\gamma(X, Y) \in D|X = x) = \int_{\mathbb{Y}} 1_{\{\gamma(x,y)\in D\}} P(Y \in dy|X = x).$$

We then have

$$\int_{\mathbb{X}\times\mathbb{Y}} c(x, \gamma(x, y)) P(dx, dy) = \int_{\mathbb{X}} \left(\int_{\mathbb{U}} c(x, u) P^{\gamma}(du|x) \right) P(dx),$$

Consider

$$h^{\gamma}(x) := \int_{\mathbb{U}} c(x, u) P^{\gamma}(du|x) \tag{5.3}$$

- Suppose the space $\mathbb{U}$ is countable. In this case, let us enumerate the elements in $\mathbb{U}$ as $\{u^i, i = 1, 2, \dots\}$. Then, we could define:
$$D_i = \{x \in \mathbb{X} : c(x, u^i) \leq h^{\gamma}(x)\}, i = 1, 2, \dots.$$
We note that $\mathbb{X} = \bigcup_i D_i$: Suppose not, then $\exists x \in \mathbb{X}$ with $c(x, u^i) > h^{\gamma}(x)$ for all $i \in \mathbb{N}$, and thus for this x:
$$h^{\gamma}(x) = \left(\sum_{\mathbb{U}} c(x, u) P^{\gamma}(du|x) \right) > h^{\gamma}(x), \tag{5.4}$$
leading to a contradiction. Now define,
$$\gamma^*(x) = u^k \quad \text{if} \quad x \in D_k \setminus (\cup_{i=1}^{k-1} D_i), k = 1, 2, \dots,$$
Such a function is measurable, by construction and performs at least as good as γ.
- We now provide a proof for the actual statement. Let $D = \{(x, u) \in \mathbb{X} \times \mathbb{U} : c(x, u) \leq h^{\gamma}(x)\}$. D is a Borel set since $c(x, u) - h^{\gamma}(x)$ is Borel. Define $D_x = \{u \in \mathbb{U} : (x, u) \in D\}$ for all $x \in \mathbb{X}$. Now, for every element x we can pick a member u which is in D; this defines a map from $\mathbb{X}$ to $\mathbb{U}$. The question now is whether the constructed map is Borel measurable. Now, for every x, $\int 1_{\{\gamma(x,y)\in D\}} P(dy|x) > 0$ by the relation (5.3), since otherwise we would arrive at a contradiction via (5.4). Then, by a measurable selection theorem of Blackwell and Ryll-Nardzewski [51] (see also p. 255 of [117]), there exists a Borel-measurable function $\gamma^* : \mathbb{X} \to \mathbb{U}$ such that its *graph* is contained in D, that is, $\{(x, \gamma^*(x)) \in D\}$.

$\diamond$

The following culminate into the following critical result.

Theorem 5.1.2 *Let $\{(x_t, u_t)\}$ be a controlled Markov chain. Consider (5.1), that is, the minimization of $E^{\gamma}[\sum_{t=0}^{N-1} c(x_t, u_t) + c_N(x_N)]$, over all admissible control policies. Any such policy can be replaced with one which is (deterministic) Markov and which is at least as good as the original policy. In particular, if an optimal control policy exists, there is no loss in restricting policies to be Markov.*

Proof. The proof follows from a sequential application of Theorem 5.1.1, starting with the final time stage. For any admissible policy, the cost

$$E[c(x_{N-1}, \gamma_{N-1}(h_{N-1})) + \int_{\mathbb{X}} c_N(z)\mathcal{T}(dz|x_{N-1}, \gamma_{N-1}(h_{N-1}))],$$

can be replaced with a measurable policy γ^*_{N-1}

$$E[c(x_{N-1}, \gamma^*_{N-1}(x_{N-1})) + \int_{\mathbb{X}} c_N(z)\mathcal{T}(dz|x_{N-1}, \gamma^*_{N-1}(x_{N-1}))],$$

which leads to a cost that is at least as good as one obtained with γ_{N-1}.

We define

$$J_{N-1}(x_{N-1}) := E\left[c(x_{N-1}, \gamma^*_{N-1}(x_{N-1})) + \int_{\mathbb{X}} c_N(z)\mathcal{T}(dz|x_{N-1}, \gamma^*_{N-1}(x_{N-1}))\right],$$

and consider then, in view of (5.2),

$$E\left[c(x_{N-2}, \gamma_{N-2}(h_{N-2})) + \int_{\mathbb{X}} J_{N-1}(z)\mathcal{T}(dz|x_{N-2}, \gamma_{N-2}(h_{N-2}))\right].$$

This expression can also be lower bounded by a measurable Markov policy γ^*_{N-2} so that the expected cost

$$J_{N-2}(x_{N-2}) := E\left[c(x_{N-2}, \gamma^*_{N-2}(x_{N-2})) + \int_{\mathbb{X}} J_{N-1}(z)\mathcal{T}(dz|x_{N-2}, \gamma^*_{N-2}(x_{N-2}))\right].$$

is lower than that achieved by the admissible policy. By induction, for all time stages, one can replace the policies with a deterministic Markov policy which leads to a cost which is at least as desirable as the cost achieved by the admissible policy. ◇

Given this result, we have the following important optimality principle.

5.1.3 Bellman's principle of optimality and Dynamic Programming

Consider 5.1. Let $\{J_t(x_t)\}$ be a sequence of functions on $\mathbb{X}$ defined by

$$J_N(x) = c_N(x)$$

and for $0 \leq t \leq N-1$

$$J_t(x) = \min_{u \in \mathbb{U}_t(x)} \{c(x, u) + \int_{\mathbb{X}} J_{t+1}(z)\mathcal{T}(dz|x, u)\}.$$

Let there be minimizing *measurable functions* which are deterministic, denoted by $\{f_t(x)\}$, so that

$$J_t(x) = c(x, f_t(x_t)) + \int_{\mathbb{X}} J_{t+1}(z)\mathcal{T}(dz|x, f_t(x))\}$$

Then we have the following:

Theorem 5.1.3 *The policy $\gamma^* = \{f_0, f_1, \ldots, f_{N-1}\}$ is optimal and the optimal expected cost function (also called the value function) is equal to*

$$J^*_{\mathcal{N}}(x) = J_0(x)$$

Proof. We compare the expected cost generated by the above policy, with respect to the cost obtained by any other policy, which can be taken to be deterministic Markov in view of Theorem 5.1.2.

We provide the proof by a backwards induction method in view of (5.2). Consider the time stage $t = N-1$. For this stage, the optimal cost (or, value) is equal to

$$J_{N-1}(x) = \min_u \{c(x,u) + \int_{\mathbb{X}} c_N(z)\mathcal{T}(dz|x_{N-1} = x, u_{N-1} = u)\}$$

Suppose there is a cost $C^*_{N-1}(x)$, achieved by some policy $\eta = \{\eta_k, k \in \{0,1,\cdots,N-1\}\}$, which we take to be deterministic Markov (without loss). Since,

$$\begin{aligned}
&C^*_{N-1}(x)\\
&= c(x,\eta_{N-1}(x)) + \int_{\mathbb{X}} c_N(z)\mathcal{T}(dz|x_{N-1} = x, u_{N-1} = \eta_{N-1}(x_{N-1}))\\
&\geq J_{N-1}(x)\\
&= \min_u \{c(x,u) + \int_{\mathbb{X}} c_N(z)\mathcal{T}(dz|x_{N-1} = x, u_{N-1} = u)\},
\end{aligned} \tag{5.5}$$

it must be that $C^*_{N-1}(x) \geq J_{N-1}(x)$. Now, we move to time stage $N-2$. In this case, the cost is given by

$$\begin{aligned}
C^*_{N-2}(x) \;&=\; c(x,\eta(x_{N-2})) + \int_{\mathbb{X}} C^*_{N-1}(z)\mathcal{T}(dz|x_{N-2} = x, u_{N-2} = \eta(x_{N-2}))\\
&\geq \min_u \{c(x,u) + \int_{\mathbb{X}} J_{N-1}(z)\mathcal{T}(dz|x_{N-2} = x, u_{N-2} = u)\}\\
&=: J_{N-2}(x)
\end{aligned}$$

where the inequality is due to the fact that $C^*_{N-1}(x) \geq J_{N-1}(x)$ and the minimization. We can, by induction, show that the recursion holds for all $0 \leq t \leq N-2$.

◇

5.1.4 Examples

Example 5.1 (Dynamic Programming and Investment). [165, Section 3.6] A investor's wealth dynamics is given by the following:

$$x_{t+1} = u_t w_t,$$

where $\{w_t\}$ is an i.i.d. $\mathbb{R}_+$-valued stochastic process with $E[w_t] = \bar{w}$. The investor has access to the past and current wealth information and his actions. The goal is to maximize, for some $b > 0$,

$$J(x_0, \gamma) = E^{\gamma}_{x_0}[\sum_{t=0}^{T-1} b(x_t - u_t)].$$

The investor's action set for any given x is: $\mathbb{U}(x) = [0,x]$. We will find an optimal admissible policy.

For this problem, the state space is $\mathbb{R}_+$, the control action space at state x is $[0,x]$, the information at the controller is $I_t = \{x_{[0,t]}, u_{[0,t-1]}\}$. The kernel is described by the relation $x_{t+1} = u_t w_t$. Using *Dynamic Programming*

$$\begin{aligned}
J_{T-1}(x) &= \max_{u\in[0,x_{T-1}]} E[b(x_{T-1} - u_{T-1})|x_{T-1} = x, u_{T-1} = u]\\
&= \max_{u\in[0,x]} b(x-u) = b(x).
\end{aligned} \tag{5.6}$$

Since there is no more *future*, the investor needs to collect the wealth at time $T-1$, that is $u_{T-1}=0$. For $t=T-2$

$$\begin{aligned}
&J_{T-2}(x) = \max_{u\in[0,x]} E[b(x-u)+J_{T-1}(x_{T-1})|x_{t-2}=x, u_{T-2}=u] \\
&= \max_{u\in[0,x]} E[b(x-u)+bx_{T-1}|x_{t-2}=x, u_{T-2}=u] \\
&= \max_{u\in[0,x]} \Big(b(x-u)+bE[w_{T-2}]u\Big) \\
&= \max_{u\in[0,x]} \Big(bx+b(\bar{w}-1)u\Big)
\end{aligned}$$

It follows then that if $\bar{w}>1$, $u_{T-2}=x_{T-2}$ (that is, investment is favourable), otherwise $u_{T-2}=0$. Recursively, one concludes that if $\bar{w}>1$, $u_t=x_t$ is optimal until $t=T-1$, at $t=T-1$, $u_{T-1}=0$, leading to $J_0(x_0)=b\bar{w}^{T-1}x_0$.

If $\bar{w}<1$, it is optimal to collect at time 0, that is $u_0=0$, leading to $J_0(x_0)=bx_0$. If $\bar{w}=1$, both of these policies lead to the same reward.

Example 5.2 (Linear Quadratic Systems). Consider the following Linear Quadratic (LQ) problem with $q>0, r>0, p_T>0$:

$$\inf_{\gamma} E_x^{\gamma}[\sum_{t=0}^{T-1} qx_t^2+r_t^2+p_Tx_T^2]$$

for a linear system:

$$x_{t+1}=ax_t+u_t+w_t,$$

where w_t is a zero-mean random variable with variance $\sigma_w^2<\infty$. We can show, by the method of completing the squares, that:

$$J_t(x_t)=P_tx_t^2+\sum_{k=t}^{T-1}P_{t+1}\sigma_w^2$$

where

$$P_t=q+P_{t+1}a^2-\frac{P_{t+1}^2a^2}{P_{t+1}+r}$$

and the optimal control policy is

$$u_t=\frac{-P_{t+1}a}{P_{t+1}+r}x_t.$$

Note that, the optimal control policy is Markov (as it uses only the current state). For a more general treatment for such LQ problems, see Section 5.3. A typical setup is the case where w_t is Gaussian; in this case the problem above is often referred to as the *Linear Quadratic Gaussian (LQG)* optimal control problem.

5.2 Existence of Minimizing Selectors and Measurability

The above dynamic programming arguments hold when there exist minimizing control policies (selectors measurable with respect to the Borel σ-field on $\mathbb{X}$). The following results build on [169, Theorem 2], [283], [282] and [202] (see Appendix C). We also refer the reader to [165] for a comprehensive analysis and detailed literature review and [126, Theorem 2.1].

Measurable Selection Hypothesis: Given a sequence of functions $J_t:\mathbb{X}\to\mathbb{R}$, there exists

$$J_t(x)=\min_{u_t\in\mathbb{U}_t(x)}(c(x_t,u_t)+\int_{\mathbb{X}}J_{t+1}(y)\mathcal{T}(dy|x,u)),$$

for all $x\in\mathbb{X}$, for $t\in\{0,1,2\ldots,N-1\}$ with

$$J_N(x_N)=c_N(x_N).$$

Furthermore, there exist measurable functions f_t such that

$$J_t(x) = c(x_t, f_t(x_t)) + \int_{\mathbb{X}} J_{t+1}(y)\mathcal{T}(dy|x, f_t(x_t)),$$

$\diamond$

Recall that a set in a normed linear space is (sequentially) compact if every sequence in the set has a converging subsequence.

Assumption 5.2.1 (Condition WF) *(i) For every continuous and bounded v on $\mathbb{X}$ (that is, $v \in C_b(\mathbb{X})$), $\int_{\mathbb{X}} \mathcal{T}(dy|x,u)v(y)$ is a continuous function on $\mathbb{X} \times \mathbb{U}$ (in this case, we call $\mathcal{T}$ a weakly continuous transition kernel).*

(ii) The cost function to be minimized $c(x,u)$ is bounded and continuous on both $\mathbb{U}$ and $\mathbb{X}$.

(iii)If applicable, c_N is continuous and bounded.

(iv)$\mathbb{U}_t(x) = \mathbb{U}$ is compact.

Assumption 5.2.2 (Condition S) *(i) For every measurable and bounded v on $\mathbb{X}$ (that is, $v \in L_\infty(\mathbb{X};\mathbb{R})$), $\int_{\mathbb{X}} \mathcal{T}(dy|x,u)v(y)$ is a continuous function on $\mathbb{U}$, for every fixed x (in this case, we call $\mathcal{T}$ a strongly continuous transition kernel in u for every fixed x).*

(ii) For every $x \in \mathbb{X}$ the bounded measurable cost function $c(x,u)$ is continuous on $\mathbb{U}$,

(iii)If applicable, c_N is bounded measurable.

(iv)$\mathbb{U}_t(x) = \mathbb{U}$ is compact.

Theorem 5.2.1 *Under Assumption 5.2.1 or Assumption 5.2.2, there exists an optimal solution and the measurable selection hypothesis applies, and there exists a minimizing control policy $f_t : \mathbb{X} \to \mathbb{U}$. Furthermore, under Assumption 5.2.1, J_t is continuous for any $t \geq 0$.*

The result follows from the following three lemmas below:

Lemma 5.2.1 *A continuous function $f : \mathbb{X} \to \mathbb{R}$ over a compact set $A \subset \mathbb{X}$ admits a minimum.*

Proof. Let $\delta = \inf_{x \in A} f(x)$. Let $\{x_i\}$ be a sequence such that $f(x_i)$ converges to δ. Since A is compact $\{x_i\}$ must have a converging subsequence $\{x_{i(n)}\}$. Let the limit of this subsequence be $\bar{x}$. Then, it follows that, $\{x_{i(n)}\} \to \bar{x}$ and thus, by continuity $\{f(x_{i(n)})\} \to f(\bar{x})$. As such $f(\bar{x}) = \delta$. $\diamond$

To see why compactness is important, consider $\inf_{x \in A} \frac{1}{x}$ for $A = [1,2)$ or $A = \mathbb{R}$. In both cases there does not exist an x value in the specified set which attains the infimum.

Lemma 5.2.2 *Let $\mathbb{U}$ be compact, and $c(x,u)$ be continuous on $\mathbb{X} \times \mathbb{U}$. Then, $\min_{u \in \mathbb{U}} c(x,u)$ is continuous on $\mathbb{X}$.*

Proof. Let $x_n \to x$, u_n optimal for x_n and u optimal for x. Such optimal action values exist as a result of compactness of $\mathbb{U}$ and continuity. Now,

$$\begin{aligned} &|\min_u c(x_n,u) - \min_u c(x,u)| \\ &\leq \max\Big(c(x_n,u) - c(x,u), c(x,u_n) - c(x_n,u_n)\Big) \end{aligned} \tag{5.7}$$

The first term above converges to zero since c is continuous in x, u. The second converges also. Suppose otherwise. Then, for some $\epsilon > 0$, there exists a subsequence such that

$$|c(x, u_{k_n}) - c(x_{k_n}, u_{k_n})| \geq \epsilon$$

Consider the sequence (x_{k_n}, u_{k_n}). There exists a further subsequence (in this sequence (x_{k_n}, u_{k_n})) $(x_{k'_n}, u_{k'_n})$ which converges to x, u' for some u' since $\mathbb{U}$ is compact. Hence, for this subsequence, we have convergence of $c(x_{k'_n}, u_{k'_n})$ as well as $c(x, u_{k'_n})$ to the same term, leading to a contradiction. ◇

Lemma 5.2.3 *Let $c(x, u)$ be a continuous function on $\mathbb{U}$ for every x, where $\mathbb{U}$ is a compact set. Then, there exists a Borel measurable function $f : \mathbb{X} \to \mathbb{U}$ such that*

$$c(x, f(x)) = \min_{u \in \mathbb{U}} c(x, u)$$

Proof. A sketch is as follows: Let $\tilde{c}(x) := \min_{u \in \mathbb{U}} c(x, u)$. The function

$$\tilde{c}(x) := \min_{u \in \mathbb{U}} c(x, u),$$

is Borel measurable. This follows from the observation that it is sufficient to prove that $\{x : \tilde{c}(x) > \alpha\}$ is Borel for every $\alpha \in \mathbb{R}$. By continuity of c and compactness of $\mathbb{U}$, with a successively refining quantization of the space of control actions $\mathbb{U}$ (such a sequence of quantizers map $\mathbb{U}$ to a sequence of finite sets (expanding as n increases), so that $\lim_{n\to\infty} \sup_u |Q_n(u) - u| = 0$ and the cardinality $|Q_n(\mathbb{U})| < \infty$ for every n)

$$\{x : \tilde{c}(x) \geq \alpha\} = \bigcap_n \bigcap_{Q_n(u), u \in \mathbb{U}} \{x : c(x, Q_n(u)) \geq \alpha\}$$

the result follows since each of $\{x : c(x, Q_n(u)) \geq \alpha\}$ is Borel. Define $\mathbb{F} := \{(x, u) : c(x, u) = \tilde{c}(x), \quad x \in \mathbb{X}.\}$ This set is a Borel set and for every x, $\{u : (x, u) \in \mathbb{F}\}$ is a closed set. The question is now whether one can construct a measurable (selection) function γ in $\mathbb{F}$ so that $\{(x, \gamma(x)), x \in \mathbb{X}\} \subset \mathbb{F}$. One can construct a measurable function which lives in this set, using the property that $\mathbb{U}$ is a separable metric space: This builds on measurable selection results, e.g. Schäl [283] and [202]; see Theorem C.0.1 (building on [169, Theorem 2], [283], [282] and [202], among others; see Appendix C).

◇

5.2.1 Some Relaxations on the Measurable Selection Conditions

We first note that one can replace the compactness condition with an *inf-compactness* condition, and modify *Condition 1* in Assumption 5.2.1 as below:

Assumption 5.2.3 (Condition 3) *For every $x \in \mathbb{X}$ the cost function to be minimized $c(x, u)$ is continuous on $\mathbb{X} \times \mathbb{U}$; is non-negative; $\{u : c(x, u) \leq \alpha\}$ is compact for all $\alpha > 0$ and all $x \in \mathbb{X}$; $\int_{\mathbb{X}} \mathcal{T}(dy|x, u)v(y)$ is a continuous function on $\mathbb{X} \times \mathbb{U}$ for every continuous and bounded v.*

Theorem 5.2.2 *Under Assumption 5.2.3, the Measurable Selection Hypothesis applies.*

The measurable selection results also hold when $\mathbb{U}(x)$ depends on x so that it is compact for each x and $\{(x, u) : u \in \mathbb{U}(x), x \in \mathbb{X}\}$ is a Borel subset of $\mathbb{X} \times \mathbb{U}$:

Lemma 5.2.4 *[169, Theorem 2], [283] [202] Let $\mathbb{X}, \mathbb{U}$ be standard Borel spaces and $\Upsilon = (x, \psi(x))$ where $\psi(x) \subset \mathbb{U}$ be such that, $\psi(x)$ is compact for each $x \in \mathbb{X}$ and Υ is a Borel measurable set in $\mathbb{X} \times \mathbb{U}$. Let $c(x, u)$ be a continuous function on $\psi(x)$ for every x.*

(i) Then, there exists a Borel measurable function $f : \mathbb{X} \to \mathbb{U}$ such that

$$c(x, f(x)) = \min_{u \in \psi(x)} c(x, u)$$

(ii) If continuity is also to be attained for the value function $c(x, f(x))$ (a close look at the proof of Lemma 5.2.2 reveals that) it suffices if $\mathbb{U}(x)$ is compact and $\mathbb{U}(x)$ is an upper semi-continuous set-valued function (the implication being that: for any $x^n \to x$ and $u'^n \in \mathbb{U}(x^n)$, there exists a subsequence u'^{n_k} which converges to some u' with the property that $u' \in \mathbb{U}(x)$) and c is continuous.

We could relax the continuity condition and change it with lower semi-continuity. A function is lower semi-continuous at x_0 if $\liminf_{x\to x_0} f(x) \geq f(x_0)$. We state the following, see also [165, Theorem 3.3.5] (we note there is a slight typo in [165, Theorem 3.3.5]; in [165, Condition 3.3.2.(c2)] should be assumed and only [165, Condition 3.3.2(c1)] is not sufficient for [165, Condition 3.3.2] to imply measurable selection).

Theorem 5.2.3 *The following hold:*

(a) Suppose that (i) $\mathbb{U}(x)$ is compact for every x and $\{(x, u) : u \in \mathbb{U}(x)\}$ is a Borel subset of $\mathbb{X} \times \mathbb{U}$, (ii) c is lower semi-continuous on $\mathbb{U}(x)$ for every $x \in \mathbb{X}$, and (iii) $\int v(x_{t+1})P(dx_{t+1}|x_t = x, u_t = u)$ is lower semi-continuous on $\mathbb{U}(x)$ for every $x \in \mathbb{X}$ and every measurable and bounded v on $\mathbb{X}$. Then, the measurable selection hypothesis applies.

(b) If (i) c is lower semi-continuous on $\{(x, u) : u \in \mathbb{U}(x), x \in \mathbb{X}\}$, (ii) for every lower semi-continuous function v on $\mathbb{X}$, $\int v(x_{t+1})P(dx_{t+1}|x_t = x, u_t = u)$ is lower semi-continuous on $\{(x, u) : u \in \mathbb{U}(x), x \in \mathbb{X}\}$, and (iii) $\mathbb{U}(x)$ is compact for every $x \in \mathbb{X}$ and $\mathbb{U}(x)$ is an upper semi-continuous set-valued function; then the value function v is lower semi-continuous.

For further related relaxations, see Appendix C and [165, Appendix D].

Universally Measurable Policies. As we discuss in Appendix C, studying the class of *universally measurable* and *semi-analytic* functions allows one to even further relax conditions required for carrying out dynamic programming recursions (and integrations) with regard to their well-posedness properties and for arriving at ϵ-optimal policies via dynamic programming.

For many problems, one can compute an optimal solution directly, without explicitly studying existence. The linear quadratic setup is one such important case.

5.3 The Linear Quadratic Regulator (LQR) Problem

Consider the following linear system

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{5.8}$$

where $x \in \mathbb{R}^n, u \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$. Suppose $\{w_t\}$ is i.i.d. zero-mean with a given covariance matrix $E[w_t w_t^T] = W$ for all $t \geq 0$ (not necessarily Gaussian).

The goal is to obtain

$$\inf_{\gamma\in\Gamma_A} J(x, \gamma),$$

where

$$J(x, \gamma) = E_x^\gamma[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t + x_N^T Q_N x_N], \tag{5.9}$$

with $R = R^T > 0, Q = Q^T \geq 0, Q_N = Q_N^T \geq 0$ (where, for matrices, the notations $>$ and $\geq$ denote the positive-definite and positive semi-definite properties, respectively).

Theorem 5.3.1 *Consider (5.9). The optimal control is linear and has the form:*

$$u_t = -(B^T P_{t+1} B + R)^{-1} B^T P_{t+1} A x_t$$

where P_t solves the Discrete-Time Riccati Equation:

$$P_t = Q + A^T P_{t+1} A - A^T P_{t+1} B (B^T P_{t+1} B + R)^{-1} B^T P_{t+1} A, \tag{5.10}$$

with final condition $P_N = Q_N$. The optimal cost is given by

$$J(x_0) = x_0^T P_0 x_0 + \sum_{t=0}^{N-1} E[w_t^T P_{t+1} w_t]$$

In the following, we study the Riccati equation (5.10). Consider the linear system

$$x_{t+1} = Ax_t + Bu_t, \qquad y_t = Cx_t \tag{5.11}$$

Here, y_t is a measurement variable and x_t is an $\mathbb{R}^n$-valued state variable. Such a system is said to be *controllable* [83], if for any initial x_i and a final x_f, there exists $T \in \mathbb{N}$ and a sequence of control actions $u_0, u_1, \cdots, u_{T-1}$ such that with $x_0 = x_i$, we have $x_T = x_f$. If x_f is restricted to be $0 \in \mathbb{R}^n$, and the above holds (but possibly with $T \to \infty$), the system is said to be *stabilizable*. Thus, the only *modes* in a stabilizable system that are not controllable are the stable ones.

Now let $B = 0$ in (5.11). Such a system is said to be *observable* if by measuring $y_0, y_1, \cdots, y_T$, for some $T \in \mathbb{N}$, x_0 can be uniquely recovered. Such a system is called *detectable* if all unstable modes of A are observable, in the sense that if $y_t \to 0$, it must be that $x_t \to 0$.

There are well-known algebraic tests to verify controllability and observability. A very useful result building on the Cayley-Hamilton theorem is that if a system cannot be moved from any initial state to any final state in n (that is, the dimension of $\mathbb{R}^n$) time stages, the system is not controllable; and if a system's initial state cannot be recovered by having the n measurements $\{y_0, y_1, \cdots, y_{n-1}\}$, the system is not observable. In particular, the linear system above with matrices (A, B) is controllable if and only if

$$\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$$

is full-rank. The pair (A, C) is observable if and only if (A^T, C^T) is controllable.

For a review of linear systems theory, the reader is referred to, e.g. [83].

Theorem 5.3.2 *(i) If (A, B) is controllable there exists a solution to the Riccati equation*

$$P = Q + A^T P A - A^T P B (B^T P B + R)^{-1} B^T P A.$$

(ii) if (A, B) is controllable and, with $Q = C^T C$, (A, C) is observable; as $t \to -\infty$ (or as $N \to \infty$ with $Q_N = \bar{P}$ fixed for an arbitrary positive semi-definite matrix $\bar{P}$), the sequence of Riccati recursions,

$$P_t = Q + A^T P_{t+1} A - A^T P_{t+1} B (B^T P_{t+1} B + R)^{-1} B^T P_{t+1} A,$$

converges to some limit P that satisfies

$$P = Q + A^T P A - A^T P B (B^T P B + R)^{-1} B^T P A.$$

That is, convergence takes place for any initial condition $\bar{P}$. Furthermore, such a P is unique, and is positive definite. Finally, under the optimal stationary control policy

$$u_t = -(B^T P B + R)^{-1} B^T P A x_t,$$

the solution to $x_{t+1} = Ax_t + Bu_t$ is stable; i.e., $x_t \to 0$.

(iii)Under the conditions of part (ii), the stationary policy above minimizes,

$$\limsup_{N\to\infty} \frac{1}{N} E_x^{\gamma}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t], \tag{5.12}$$

for the system (5.8), for every $x \in \mathbb{R}^n$. Furthermore, the optimal cost is $E[w^T P w] = \mathrm{Trace}(PW)$.

Remark 5.3. Part (i) can be relaxed to (A, B) being stabilizable; and part (ii) to (A, C) being detectable for the existence of a unique P and a stable system under the optimal policy. In this case, however, P may only be positive semi-definite.

Proof.

(i) Assume that $w_t = 0$ for all t; the noise does not affect the recursions in the Riccati equation. Now, since the system is controllable there exists a control sequence such that $x_t = 0$ for $t \geq n$ which also satisfies $u_t = 0$ for $t \geq n$. The cost $\sum_{t=0}^{\infty} x_t^T Q x_t + u_t^T R u_t$ induced by this control sequence is finite (and thus bounded by some $M(x_0)$). Now, define $P_0^{(N)}$ through

$$x_0^T P_0^{(N)} x_0 = \inf_{\gamma\in\Gamma_A} E_x^{\gamma}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t]$$

and observe that

$$x_0^T P_0^{(N)} x_0 \leq x_0^T P_0^{(N+1)} x_0 \leq M(x_0).$$

As a result, for a fixed x_0, we can conclude that the sequence $\{x_0^T P_0^{(N)} x_0, T \geq 0\}$ is monotone (non-decreasing) and bounded from above. Thus, the sequence has a limit. By selecting different values of x_0 (e.g., with $x_0 = \begin{bmatrix}1 & 0 & 0 & \cdots & 0\end{bmatrix}^T$, $x_0 = \begin{bmatrix}0 & 1 & 0 & \cdots & 0\end{bmatrix}^T$, $x_0 = \begin{bmatrix}1 & 1 & 0 & \cdots & 0\end{bmatrix}^T$ and so on), we conclude that there is a fixed point P such that $x_0^T P_0^{(N)} x_0 \to x_0^T P x_0$ for any $x_0 \in \mathbb{R}^n$ (i.e., $P_0^{(N)} \to P$ point-wise in the matrix entries).

(ii) As above, assume again that $w_t = 0$ for all $t \geq 0$. Let P be the fixed point in (i). We will show that this is the unique fixed point.

We use the property that, through a change of limit supremum and infimum argument (as in Lemma 5.5.1 further below),

$$\begin{aligned} \infty > M(x) \;&\geq \inf_{\gamma\in\Gamma_A} \limsup_{N\to\infty} E_x^{\gamma}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t] \\ &\geq \limsup_{N\to\infty} \inf_{\gamma\in\Gamma_A} E_x^{\gamma}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t] = x^T P x \end{aligned} \tag{5.13}$$

Now, we show that, the sequence of policies that are optimal for each N converge to the stationary policy by $\gamma^*(x_t) = -(B^T P B + R)^{-1} B^T P A x_t$ and that this policy attains the cost $x^T P x$: Note that, with $x_0 = x$,

$$x^T P x = E^{\gamma^*}[(\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t) + x_N^T P x_N] \tag{5.14}$$

is finite and as a result, under this policy γ^*, we have that $\sum_{t=0}^{\infty} x_t^T Q x_t + u_t^T R u_t$ is finite.

However, since the induced cost is finite and $R > 0$, under this control policy, $u_t \to 0$. Therefore, the policy γ^*

$$u_t = -(B^T P B + R)^{-1} B^T P A x_t = \gamma^*(x_t),$$

is stabilizing: This follows because since $x^T Q x_t \to 0$ (and $u_t \to 0$), by observability of (C, A) it must be that $x_t \to 0$ as well (note that here one should also use that $u_t \to 0$). As a result, we conclude that, by taking $N \to \infty$ in (5.14), γ^* satisfies

$$x^T P x = E^{\gamma^*}[\sum_{t=0}^{\infty} x_t^T Q x_t + u_t^T R u_t],$$

and is therefore optimal (by 5.13).

Given this optimality (which will be useful also for (iii) below), we now show uniqueness. Let

$$x_0^T P_0^{N,\bar{P}} x_0 := \inf_{\gamma \in \Gamma_A} E_x^{\gamma}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t + x_N^T \bar{P} x_N] \tag{5.15}$$

be the solution of the optimization problem where $P_N = \bar{P}$. We will show that $P_0^{N,\bar{P}} \to P$ regardless of the value of the positive semi-definite matrix $\bar{P}$, leading to the uniqueness of the limit.

By writing $E_{x_0}^{\gamma^*}[x_N^T \bar{P} x_N] = E_{x_0}^{\gamma^*}[x_N^T P x_N] + E_{x_0}^{\gamma^*}[x_N^T(\bar{P} - P)x_N]$, noting that, as P is a solution to the Riccati recursion,

$$E_x^{\gamma^*}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t + x_N^T P x_N] = x_0^T P x_0,$$

we have that

$$x_0^T P_0^{(N)} x_0 \leq x_0^T P_0^{N,\bar{P}} x_0 \leq E_x^{\gamma^*}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t + x_N^T P x_N] + E_{x_0}^{\gamma^*}[x_N^T(\bar{P} - P)x_N].$$

or

$$x_0^T P_0^{(N)} x_0 \leq x_0^T P_0^{N,\bar{P}} x_0 \leq x_0^T P x_0 + E_{x_0}^{\gamma^*}[x_N^T(\bar{P} - P)x_N].$$

The above holds as γ^* is not necessarily optimal, and provides an upper bound, for (5.15). However, through the property that $x_N \to 0$ as $N \to \infty$ under $u_t = \gamma^*(x_t)$, we conclude that $P_0^{N,\bar{P}} \to P$ and hence uniqueness follows.

(iii)As in (ii), we use the property that, through a change of limit supremum and infimum argument (as in Lemma 5.5.1 further below),

$$\begin{aligned} &\inf_{\gamma \in \Gamma_A} \limsup_{N \to \infty} \frac{1}{N} E_x^{\gamma}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t] \\ &\geq \limsup_{N \to \infty} \frac{1}{N} \inf_{\gamma \in \Gamma_A} E_x^{\gamma}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t] \end{aligned} \tag{5.16}$$

Since, $\inf_{\gamma \in \Gamma_A} E_x^{\gamma}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t]$ is determined by $P^{(N)}$ that converges to P, leading to the optimality of γ^*, and the policy $u_t = -(BPB + R)^{-1} B^T P A x_t$ is stabilizing, this implies that the policy γ^* is optimal for (5.12) as well; see (5.13) and the following discussion. The optimal cost then is $E[w^T P w] = \text{Trace}(PW)$, via observing that $P_{t+1}^{(N)} \to P$ for all t as $N \to \infty$ and writing

$$\lim_{N \to \infty} \frac{1}{N} \left(x_0^T P_0^{(N)} x_0 + \sum_{t=0}^{N-1} E[w_t^T P_{t+1}^{(N)} w_t] \right) = E[w^T P w].$$

◇

We will discuss average cost optimization problems in further detail in *Chapter* 7.

5.4 Optional: A Strategic Measures Approach

For stochastic control problems, *strategic measures* are defined (see [283], [117] and [124]) as the set of probability measures induced on the product spaces of the state and action pairs by measurable control policies: Given an initial distribution on the state, and a policy, one can uniquely define a probability measure on the product space. Topological

properties, such as measurability and compactness, of sets of strategic measures are studied in [283], [117], [124] and [49].

We assume, as before, that the spaces considered are standard Borel. In the following, we consider a finite horizon problem, with time horizon $N-1$.

Theorem 5.4.1 *Let $L_R(\mu)$ be the set of strategic measures induced by (possibly randomized) Γ_A with $x_0 \sim \mu$. Then, for any $P \in L_R(\mu)$, there exists an augmented space Ω and a probability measure η such that*

$$P(B) = \int_{\Omega} \eta(d\omega) P_{\mu}^{\gamma(\omega)}(B), \quad B \in \mathcal{B}((\mathbb{X} \times \mathbb{U})^N),$$

where each $\gamma(\omega) \in \Gamma_A$ is deterministic admissible.

Proof. Here, we build on Lemma 1.2 in Gikhman and Shorodhod [144] and Theorem 1 in [123]. Any stochastic kernel $P(dx|y)$ can be realized by some measurable function $x = f(y, v)$ where v is a uniformly distributed random variable on $[0, 1]$ and f is measurable (see also [56] for a related argument). One can define a new random variable $(\omega = (v_0, v_1, \cdots, v_{T-1}))$. In particular, η can be taken to be the probability measure constructed on the product space $[0, 1]^N$ by the independent variables $v_k, k \in \{0, 1, \cdots, N-1\}$. ◇

One implication of this theorem is that if one relaxes the measure η to be arbitrary, a convex representation would be possible. That is, the set

$$P(B) = \int_{\Omega} \eta(d\omega) P_{\mu}^{\gamma(\omega)}(B), \quad B \in \mathcal{B}((\mathbb{X} \times \mathbb{U})^N), \eta \in \mathcal{P}(\Omega)$$

is convex, when one does not restrict η to be a fixed measure. Furthermore, the extreme points of these convex sets consist of policies which are deterministic. A further implication then is that, since the expected cost function is linear in the strategic measures, one can without any loss consider the extreme points while searching for optimal policies. In particular,

$$\inf_{\gamma \in \Gamma_{MR}} J(x, \gamma) = \inf_{\gamma \in \Gamma_M} J(x, \gamma)$$

and

$$\inf_{\gamma \in \Gamma_{AR}} J(x, \gamma) = \inf_{\gamma \in \Gamma_A} J(x, \gamma).$$

Thus, deterministic policies are as good as any other. This is not surprising in view of Theorem 5.1.1.

We present the following characterization for strategic measures. Let for all $n \in \mathbb{N}$, $h_n = \{x_0, u_0, \cdots, x_{n-1}, u_{n-1}, x_n, u_n\}$, and $P(dx_n|h_{n-1}) = \mathcal{T}(dx_n|x_{n-1}, u_{n-1})$ be the transition kernel.

Let $L_A(\mu)$ be the set of strategic measures induced by deterministic policies and let $L_R(\mu)$ be the set of strategic measures induced by independently provided randomized policies. Such an individual randomized policy can be represented in a functional form, as noted earlier: for any stochastic kernel Π^k from $\mathbb{Y}^k$ to $\mathbb{U}^k$, there exists a measurable function $\gamma^k : [0, 1] \times \mathbb{Y}^k \to \mathbb{U}^k$ such that

$$m\{r : \gamma^k(r, y^k) \in A\} = \gamma^k(u^k \in A|y^k), \tag{5.17}$$

and m is the uniform distribution (Lebesgue measure) on $[0, 1]$.

Theorem 5.4.2 *A probability measure $P \in \mathcal{P}\left(\prod_{k=1}^{N}(\mathbb{X} \times \mathbb{U})\right)$ is a strategic measure induced by a randomized policy (that is in $L_R(\mu)$) if and only if for every $n \in \mathbb{N}$ and for all continuous and bounded g:*

$$\int P(dh_{n-1}, dx_n) g(h_{n-1}, x_n) = \int P(dh_{n-1}) \left(\int_{\mathbb{X}} g(h_{n-1}, z) \mathcal{T}(dz|h_{n-1}) \right), \tag{5.18}$$

(where we recall that $\mathcal{T}(B|h_{n-1}) = \int_B \mathcal{T}(dx_n|x_{n-1}, u_{n-1})$), and

$$\int P(dh_n)g(h_{n-1}, x_n, u_n) = \int P(dh_{n-1}, dx_n)\Big(\int_{\mathbb{U}^n} g(h_{n-1}, x_n, a)\gamma^n(da|h_{n-1}, x_n)\Big), \tag{5.19}$$

for some stochastic kernel γ^n on $\mathbb{U}^n$ given h_n, x_n, with $P(d\omega_0) = \mu(dw_0)$.

Proof. The proof follows from the fact that testing the equalities such as (5.18-5.19) on continuous and bounded functions implies this property for any measurable and bounded function (that is, continuous and bounded functions form a *separating class*, see e.g. [42, p. 13] or [120, Theorem 3.4.5]) ◇

An implication is the following.

Theorem 5.4.3 *[283] The set of strategic measures induced by admissible randomized policies is compact under the weak convergence topology, if Assumption 5.2.1holdssothatT($dx_{t+1}|x_t = x, u_t = u$) is weakly continuous in x, u and also $\mathbb{X}, \mathbb{U}$ are compact.*

An implication of this result is that optimal policies exist, and are deterministic when the cost function is continuous in x, u.

We note also that Schäl [283] introduces a more general topology, $w - s$ topology, which requires strong continuity in control actions. In this case, one can generalize Theorem 5.4.3 to the setups where Condition 2 applies and existence of optimal policies follows.

We refer the reader to the Appendix, Section D.4, for a definition of the w-s topology.

Theorem 5.4.4 *[283] The set of strategic measures induced by admissible randomized policies is sequentially compact under the w-s topology, if $\mathcal{T}(dx_{t+1}|x_t = x, u_t = u)$ is strongly continuous in u for every x and also $\mathbb{X}, \mathbb{U}$ are compact.*

Thus, the above is a counterpart for when Assumption 5.2.2 holds.

The proofs of Theorems 5.4.3 and 5.4.4 follow from the property that to check whether a conditional independence property, as in (5.18-5.19)), holds testing these on continuous and bounded functions implies this property for any measurable and bounded function. Note that (5.19) holds since there is no conditional independence property condition, and the main issue is to establish that (5.18) holds for any converging sequence of strategic measures. Applying the hypotheses for each of the theorems leads to the desired results.

An implication of Theorem 5.4.4 is that an optimal strategic measure exists under the conditions of the theorem, provided that the $\mathbb{R}_+$-valued cost function c is lower semi-continuous in u for every x. In particular, for any w-s converging sequence of strategic measures which satisfies (5.18)-(5.19), so does the limit. By [283, Theorem 3.7], and the generalization of Portmanteau theorem for the w-s topology, the lower semi-continuity of the integral cost over the set of strategic measures leads to the existence of an optimal strategic measure.

Now, we know that an optimal policy will be deterministic as a consequence of Theorem 5.4.1. Thus, an optimal policy (which is deterministic) exists.

5.5 Infinite Horizon Optimal Discounted Cost Control Problems

When the time horizon becomes unbounded, we cannot directly invoke dynamic programming in the form considered earlier. Infinite horizon problems that we will consider will belong to two classes: *Discounted cost* and *average cost* problems. In the following, we first discuss the discounted cost problem. The average cost problem is discussed in *Chapter 7.*

Under the discounted cost criterion, future cost realizations are discounted: the future is perceived to be less important than the current time with different justifications depending on the applications, e.g. due to the uncertainty in the future leading one become more cautious about optimizing for the distant time stages, or perhaps due to an economic understanding that the current value of a good is more important than its value in the future.

For a given $T \in \mathbb{Z}_+$, the expected discounted cost criterion is given as:

$$J_\beta^T(x_0, \gamma) = E_{x_0}^\gamma [\sum_{t=0}^{T-1} \beta^t c(x_t, u_t)], \tag{5.20}$$

for some $\beta \in (0, 1)$. If there exists a policy γ^* which minimizes this cost, the policy is said to be optimal. We often consider an infinite horizon problem by taking the limit (when c is non-negative)

$$J_\beta(x_0, \gamma) = \lim_{T\to\infty} E_{x_0}^\gamma [\sum_{t=0}^{T-1} \beta^t c(x_t, u_t)],$$

and invoking the monotone convergence theorem:

$$J_\beta(x_0, \gamma) = E_{x_0}^\gamma [\sum_{t=0}^{\infty} \beta^t c(x_t, u_t)]. \tag{5.21}$$

We seek to find

$$J_\beta(x_0) = \inf_{\gamma\in\Gamma_A} J_\beta(x_0, \gamma).$$

Define

$$\inf_{\gamma\in\Gamma_A} J_\beta^T(x_0, \gamma) = J_\beta^T(x_0)$$

Lemma 5.5.1 *Let $\mathbb{A}$ be a set and $\{f_n\}$ be a sequence of maps from $f_n : \mathbb{A} \to \mathbb{R}$ for all $n \in \mathbb{N}$. Then,*

$$\limsup_{n\to\infty} \inf_{x\in\mathbb{A}} f_n(x) \le \inf_{x\in\mathbb{A}} \limsup_{n\to\infty} f_n(x).$$

Proof. For any $n \in \mathbb{N}$ and $y \in \mathbb{A}$ we have

$$\inf_{x\in\mathbb{A}} f_n(x) \le f_n(y).$$

This holds for all n we can take the limit superior of both sides, which yields

$$\limsup_{n\to\infty} \inf_{x\in\mathbb{A}} f_n(x) \le \limsup_{n\to\infty} f_n(y).$$

This inequality holds for all $y \in \mathbb{A}$ and thus

$$\limsup_{n\to\infty} \inf_{x\in\mathbb{A}} f_n(x) \le \inf_{x\in\mathbb{A}} \limsup_{n\to\infty} f_n(x).$$

◇

By Lemma 5.5.1, we change the order of limit and infimum so that

$$J_\beta(x_0) \ge \limsup_{T\to\infty} J_\beta^T(x_0) \tag{5.22}$$

but since $\lim$ exists for the right-hand side as the expression is monotonically increasing the limit superior becomes an actual limit and thus

$$J_\beta(x_0) \ge \lim_{T\to\infty} J_\beta^T(x_0).$$

We will make use of this relation explicitly in Lemma 5.5.4 below. Now, observe that (from (5.20))

$$J_\beta^T(x_0,\gamma) = E_{x_0}^\gamma\Bigg[c(x_0,u_0) + E^\gamma[\sum_{t=1}^{T-1}\beta^t c(x_t,u_t)|x_1,x_0,u_0]\Bigg|x_0,u_0\Bigg],$$

writes as

$$J_\beta^T(x_0,\gamma) = E_{x_0}^\gamma\Bigg[c(x_0,u_0) + \beta E^\gamma[\sum_{t=1}^{T-1}\beta^{t-1} c(x_t,u_t)|x_1,x_0,u_0]\Bigg|x_0,u_0\Bigg].$$

Through the controlled Markov property and the fact that without any loss Markov policies are as good as any other for finite horizon problems, it follows that (if dynamic programming recursions are well-defined)

$$J_\beta^T(x_0) = \inf_{u_0} E_{x_0}^\gamma\Bigg[c(x_0,u_0) + \beta E^\gamma[J_\beta^{T-1}(x_1)|x_0,u_0]\Bigg] \tag{5.23}$$

We also saw in fact, under measurable selection conditions, via Bellman's Theorem 5.1.3, the above is in fact an equality. The goal is now to take $T\to\infty$ and obtain desirable structural properties. The limit

$$\lim_{T\to\infty} J_\beta^T(x_0)$$

will be a lower bound to $J_\beta(x_0)$ by (5.22). But the inequality will turn out to be an equality under mild conditions to be studied in the following. The next result is on the exchange of the order of the minimum and limits.

Lemma 5.5.2 *[165] Let $V_n(x,u)\uparrow V(x,u)$ pointwise. Suppose that V_n and V are continuous in u for every x, and $u\in\mathbb{U}(x)=\mathbb{U}$ is compact. Then,*

$$\lim_{n\to\infty}\min_{u\in\mathbb{U}(x)} V_n(x,u) = \min_{u\in\mathbb{U}(x)} V(x,u)$$

Proof. The proof follows from essentially the same arguments as in the proof of Lemma 5.2.2. Let u_n^* solve $\min_{u\in\mathbb{U}(x)} V_n(x,u)$. Note that

$$|\min_{u\in\mathbb{U}(x)} V_n(x,u) - \min_{u\in\mathbb{U}(x)} V(x,u)| \le V(x,u_n^*) - V_n(x,u_n^*), \tag{5.24}$$

since $V_n(x,u)\uparrow V(x,u)$. Now, suppose that for some $\epsilon>0$

$$V(x,u_n^*) - V_n(x,u_n^*) \ge \epsilon, \tag{5.25}$$

along a subsequence n_k. There exists a further subsequence n_k' such that $u_{n_k'}^*\to\bar{u}$ for some $\bar{u}$. By assumption, for this x and $\bar{u}$, and every $\epsilon>0$, we can find a sufficiently large N such that $V(x,\bar{u}) - V_N(x,\bar{u})\le\epsilon/2$. Fix such an N. Now, for every $n_k'\ge N$, since V_n is monotonically increasing:

$$V(x,u_{n_k'}^*) - V_{n_k'}(x,u_{n_k'}^*) \le V(x,u_{n_k'}^*) - V_N(x,u_{n_k'}^*)$$

However, $V(x,u_{n_k'}^*)$ and for the fixed N, $V_N(x,u_{n_k'}^*)$, are continuous hence these two terms converge to: $V(x,\bar{u}) - V_N(x,\bar{u})$. Hence (5.25) cannot hold. ◇

Recall from dynamic programming equations that with

$$\inf_{\gamma\in\Gamma_A} J_\beta^T(x_0,\gamma) = J_\beta^T(x_0),$$

we have (5.23):

$$J_\beta^T(x_0) = \min_{u_0}\Bigg(c(x_0,u_0) + \beta E^\gamma[J_\beta^{T-1}(x_1)|x_0,u_0]\Bigg).$$

It follows then that

$$J_\beta^\infty(x_0) := \lim_{T\to\infty} J_\beta^T(x_0) = \lim_{T\to\infty}\min_{u_0}\Bigg(c(x_0,u_0) + \beta E[J_\beta^{T-1}(x_1)|x_0,u_0]\Bigg),$$

where the limit exists due to the monotone convergence theorem since the cost is increasing with T: $J_\beta^T(x_1) \uparrow J_\beta^\infty$ as $T \to \infty$. If Lemma 5.5.2 applies (i.e., the continuity condition in actions holds), we obtain that

$$J_\beta^\infty(x_0) = \min_{u_0} \lim_{T\to\infty} \Big(c(x_0, u_0) + \beta E[J_\beta^T(x_1)|x_0, u_0] \Big), \tag{5.26}$$

and thus

$$J_\beta^\infty(x_0) = \min_{u_0} \Big(c(x_0, u_0) + \beta E[J_\beta^\infty(x_1)|x_0, u_0] \Big). \tag{5.27}$$

The following result shows that the fixed point equation (5.27) is closely related to optimality. Define $\mathbb{T}$ as follows:

$$(\mathbb{T}(v))(x) := \min_u \Big(c(x, u) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x, u) \Big).$$

and define the *Discounted Cost Optimality Equation (DCOE)* as follows

$$v(x) = (\mathbb{T}(v))(x), \qquad x \in \mathbb{X} \tag{5.28}$$

Lemma 5.5.3 *[Verification Theorem] [165]*

(i) If v is a measurable $\mathbb{R}_+$-valued function under Assumption 5.2.2 (or continuous and bounded function under Assumption 5.2.1) with $v \geq \mathbb{T}v$, then $v(x) \geq J_\beta(x)$.

(ii) If $\mathbb{T}v \geq v$ and

$$\lim_{n\to\infty} \beta^n E_x^\gamma[v(x_n)] = 0, \tag{5.29}$$

for every policy and initial condition, then $v(x) \leq J_\beta(x)$. As a result, a fixed point to (5.27) leads to an optimal policy under (5.29).

Proof.

(i) For some stationary policy f that achieves (whose existence is justified by the measurable selection conditions)

$$\min(c(x, u) + \beta E[v(x_1)|x_0 = x, u_0 = u]) = c(x, f(x)) + \beta E[v(x_1)|x_0 = x, u_0 = f(x)],$$

apply repeatedly

$$v(x) \geq c(x, f(x)) + \beta \int v(y)\mathcal{T}(dy|x, f(x)) \geq \cdots \geq E_x^f[\sum_{k=0}^{n-1} \beta^t c(x_k, f(x_k))] + \beta^n E_x^f[v(x_n)]$$

Thus, taking the limit and given that v is non-negative valued,

$$v(x) \geq \limsup_{n\to\infty} E_x^f[\sum_{k=0}^{n-1} \beta^t c(x_k, f(x_k))] + \beta^n E_x^f[v(x_n)] \geq \lim_{n\to\infty} E_x^f[\sum_{k=0}^{n-1} \beta^t c(x_k, f(x_k))] \geq J_\beta(x),$$

since $\{f, f, f, \cdots, f, \cdots\}$ is a particular policy and $J_\beta(x)$ is the optimal expected cost among all admissible policies.

(ii) If $\mathbb{T}v(x) \geq v(x)$, then for any $n \in \mathbb{Z}_+$,

$$\begin{aligned} &E_x^\gamma[\beta^{n+1} v(x_{n+1})|h_n] = E_x^\gamma[\beta^{n+1} v(x_{n+1})|x_n, u_n] \\ &= \beta^n \Big(\{c(x_n, u_n) + \beta \int v(z)\mathcal{T}(dz|x_n, u_n)\} - c(x_n, u_n) \Big) \end{aligned}$$

$$\geq \beta^n(v(x_n) - c(x_n, u_n)), \tag{5.30}$$

where we use the inequality $c(x_n, u_n) + \beta \int v(z)\mathcal{T}(dz|x_n, u_n) \geq v(x_n)$. Thus, using the iterated expectations and arranging the terms

$$E_x^\gamma[\sum_{k=0}^{n-1} \beta^n c(x_n, u_n)] \geq E[\sum_{k=0}^{n-1} E[\beta^n v(x_k) - \beta^{n+1} v(x_{k+1})|h_k]]$$

leading to

$$E_x^\gamma[\sum_{k=0}^{n-1} \beta^n c(x_n, u_n)] \geq v(x) - \beta^n E_x^\gamma[v(x_n)]$$

If the last term on the right hand size converges to zero, then the result is obtained so that for any fixed policy, v provides a lower bound on the value function. Taking the infimum over all admissible policies, the desired result $v(x) \leq J_\beta(x)$ is obtained.

◇

We have the following refinement, where we do not need to check (5.29) for every policy.

Lemma 5.5.4 *If*

$$v(x) = \lim_{T\to\infty} J_\beta^T(x)$$

is so that $v = \mathbb{T}(v)$ *where*

$$\mathbb{T}(v)(x) = c(x, f(x)) + \beta E[v(x_1)|x_0 = x, u_0 = f(x)]$$

is such that with $\gamma = \{f, f, \cdots\} \in \Gamma_S$*,*

$$\lim_{n\to\infty} \beta^n E_x^\gamma[v(x_n)] = 0, \tag{5.31}$$

then γ *is optimal.*

Proof. Equation (5.22) implies that $J_\beta(x) \geq v(x)$ since v is the pointwise limit of the discounted cost functions as the horizon increases. Now, since the stationary policy f achieves

$$v(x) = \min_{u\in\mathbb{X}} \Big(c(x, u) + \beta E[v(x_1)|x_0 = x, u_0 = u] \Big) = c(x, f(x)) + \beta E[v(x_1)|x_0 = x, u_0 = f(x)],$$

applying this repeatedly to $v(x_0), v(x_1)$ and then up to $v(x_{n-1})$ leads to

$$v(x) = c(x, f(x)) + \beta \int v(x_1)\mathcal{T}(dx_1|x, f(x)) = \cdots = E_x^\gamma[\sum_{k=0}^{n-1} \beta^k c(x_k, f(x_k))] + \beta^n E_x^\gamma[v(x_n)].$$

Taking the limit, we have

$$v(x) = E_x^\gamma[\sum_{k=0}^{\infty} \beta^k c(x_k, f(x_k))]$$

implying that $\gamma = \{f, f, \cdots\}$ is optimal. ◇

An implication of the proof of the result above is that for any stationary policy $\gamma = \{f, f, \cdots, f, \cdots\}$, we have the following equation:

$$J_\beta(x, \gamma) = c(x, f(x)) + \beta E[J_\beta(x_1, \gamma)|x_0 = x, u_0 = f(x)] \tag{5.32}$$

provided that

$$\lim_{n\to\infty} \beta^n E^\gamma[J_\beta(x_n, \gamma)] = 0.$$

This will be useful later on when we study numerical methods.

A sufficient condition for (5.31) is that the cost function c is bounded, though this is certainly not necessary.

5.5.1 Value Iteration Algorithm and Regularity of Value Functions

By dynamic programming, the Bellman optimality recursion for every finite horizon $T \in \mathbb{N}$ be written as

$$J_t^T(x) = \mathbb{T}(J_{t+1}^T)(x) = \min_u \left(c(x,u) + \beta \int_{\mathbb{X}} J_{t+1}^T(y) \mathcal{T}(dy|x,u) \right), \qquad t = T-1, T-2, \cdots, 0, \tag{5.33}$$

with

$$J_T^T(x) = 0.$$

This sequence will lead to a solution for a T-stage discounted optimal cost problem. In particular, if we define $v_0 := J_T^T$, and $v_n := J_{T-n}^T$, we obtain the recursions for $n = 1, 2, \cdots$,

$$v_{n+1} = \mathbb{T}(v_n)(x),$$

which will form the basis of a very important algorithm, known as the value iteration algorithm, to be presented below. The following then is a consequence of Lemma 5.5.4.

Theorem 5.5.1 *[Value Iteration Algorithm: General Cost Setup] Suppose the cost function is non-negative. Consider the successive iteration*

$$v_n(x) = \min_u \{ c(x,u) + \beta \int_{\mathbb{X}} v_{n-1}(y) \mathcal{T}(dy|x,u) \}, \forall x, n \geq 1 \tag{5.34}$$

with $v_0(x) = 0$ for all $x \in \mathbb{X}$. Then, v_n is a monotonically non-decreasing sequence. If this sequence converges pointwise to a function v where

$$v(x) = c(x, f(x)) + \beta \int v(y) \mathcal{T}(dy|x, f(x))$$

is such that with $\gamma = \{f, f, \cdots\}$, (5.31) holds, then γ is optimal and v is the value function.

A sufficient condition for the iterations in (5.33) to converge is the following. Suppose that measurable selection conditions apply so that the iterations are well defined for every $n \in \mathbb{Z}_+$. Let there exist a policy which leads to a finite cost for every initial state and that by dynamic programming the recursions for every T given in (5.33) hold. This sequence will lead to a solution for a T-stage discounted cost problem. Since $J_t^T(x) \leq J_t^{T+1}(x)$, if there exists some J_t^∞ such that $J_t^T(x) \uparrow J_t^\infty(x)$, we could invoke Lemma 5.5.2 to argue that

$$J_t^\infty(x) = \mathbb{T}(J_{t+1}^\infty)(x) = \min_u \{ c(x,u) + \beta \int_{\mathbb{X}} J_{t+1}^\infty(y) \mathcal{T}(dy|x,u) \}.$$

Such a limit exists, by the monotone convergence theorem since $J_t^\infty(x) < \infty$ due to the assumption that there exists a policy leading to a finite cost for every initial state. Hence, a limit satisfying (5.28) indeed exists. If

$$\{ c(x,u) + \beta \int_{\mathbb{X}} J_{t+1}^\infty(y) \mathcal{T}(dy|x,u) \}$$

and

$$\{ c(x,u) + \beta \int_{\mathbb{X}} J_{t+1}^T(y) \mathcal{T}(dy|x,u) \}$$

are continuous in u for every x and every T and t, by (5.22), a lower bound to an optimal solution will have to satisfy a fixed point equation (5.28). The result then would follow from Lemma 5.5.4.

In the bounded cost case, we can obtain a very strong result with a direct argument.

Lemma 5.5.5 *(i) The space of measurable functions $\mathbb{X} \to \mathbb{R}$ endowed with the $||.||_\infty$ norm (also called the supremum norm) is a Banach space, that is*

$$L_\infty(\mathbb{X};\mathbb{R}) = \{f : \mathbb{X} \to \mathbb{R} : ||f||_\infty = \sup_x |f(x)| < \infty\}$$

is a Banach space.
(ii) The space of continuous and bounded functions from $\mathbb{X} \to \mathbb{R}$, $C_b(\mathbb{X})$, endowed with the $||.||_\infty$ norm is a Banach space.

Theorem 5.5.2 *[Value Iteration Algorithm - Bounded Cost Setup] Suppose the cost function is bounded, non-negative, and one of the measurable selection conditions (Condition WF in Assumption 5.2.1 or Condition S in Assumption 5.2.2) applies. Then, there exists a unique solution to the discounted cost problem which solves the fixed point equation.*

$$v(x) = \min_u\{c(x,u) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x,u)\}, \qquad x \in \mathbb{X}$$

Furthermore, the optimal cost (value function) is obtained by a successive iteration (known as the Value Iteration Algorithm):

$$v_n(x) = \min_u\{c(x,u) + \beta \int_{\mathbb{X}} v_{n-1}(y)\mathcal{T}(dy|x,u)\}, \qquad \forall x, n \in \mathbb{N} \tag{5.35}$$

For any $v_0 \in L_\infty(\mathbb{X};\mathbb{R})$, the sequence converges to a unique fixed point. If $v_0(x) = 0, x \in \mathbb{X}$, then $v_n(x) \uparrow v(x)$ for all $x \in \mathbb{X}$ (that is, v_n monotonically converges to v). If Condition WF applies, then v is also continuous.

Proof of Theorem 5.5.2 Depending on the measurable selection conditions, we can take the value functions to be either measurable and bounded, or continuous and bounded. (i) Suppose that we consider the measurable and bounded case (Assumption 5.2.2). We observe that the vector J^∞ lives in $L_\infty(\mathbb{X};\mathbb{R})$ (since the cost is bounded, there is a uniform bound for every x). We will show that the iteration given by

$$\mathbb{T}(v)(x) = \min_u\{c(x,u) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x,u)\}$$

is a contraction in $L_\infty(\mathbb{X};\mathbb{R})$. Let

$$\begin{aligned}
&||\mathbb{T}(v) - \mathbb{T}(v')||_\infty = \sup_{x\in\mathbb{X}} |\mathbb{T}(v)(x) - \mathbb{T}(v')(x)| \\
&= \sup_{x\in\mathbb{X}} \left| \min_u\{c(x,u) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x,u)\} - \min_u\{c(x,u) + \beta \int_{\mathbb{X}} v'(y)\mathcal{T}(dy|x,u)\} \right| \\
&\leq \sup_{x\in\mathbb{X}} \left(1_{\{x\in A1\}} \left\{ c(x,u_x^*) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x,u_x^*) - c(x,u_x^*) - \beta \int_{\mathbb{X}} v'(y)\mathcal{T}(dy|x,u_x^*) \right\} \right. \\
&\left. + 1_{\{x\in A2\}} \left\{ - c(x,u_x^{**}) - \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x,u_x^{**}) + c(x,u_x^{**}) + \beta \int_{\mathbb{X}} v'(y)\mathcal{T}(dy|x,u_x^{**}) \right\} \right) \\
&= \sup_{x\in\mathbb{X}} \left(1_{\{x\in A1\}} \{\beta \int_{\mathbb{X}} (v(y) - v(y'))\mathcal{T}(dy|x,u_x^*)\} \right) + \sup_{x\in\mathbb{X}} \left(1_{\{x\in A2\}} \{\beta \int_{\mathbb{X}} (v'(y) - v(y))\mathcal{T}(dy|x,u_x^{**})\} \right) \\
&\leq \beta||v - v'||_\infty \{1_{\{x\in A1\}} \int_{\mathbb{X}} \mathcal{T}(dy|x,u_x^{**}) + 1_{\{x\in A2\}} \int_{\mathbb{X}} \mathcal{T}(dy|x,u_x^*)\} \\
&= \beta||v - v'||_\infty
\end{aligned} \tag{5.36}$$

Here

$$A_1 = \left\{ x : \min_u\{c(x,u) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x,u)\} \geq \min_u\{c(x,u) + \beta \int_{\mathbb{X}} v'(y)\mathcal{T}(dy|x,u)\} \right\},$$

and A_2 denotes the complementary event, u_x^{**} is a minimizing control for $\{c(x,u) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x,u)\}$ and u_x^* is a minimizer for $\{c(x,u) + \beta \int_{\mathbb{X}} v'(y)\mathcal{T}(dy|x,u)\}$. As a result $\mathbb{T}$ defines a contraction on the Banach space $L_\infty(\mathbb{X};\mathbb{R})$,

and there exists a unique fixed point. Thus, the sequence of iterations in (5.33),

$$J_t^T(x) = \mathbb{T}(J_{t+1}^T)(x) = \{\min_u\{c(x,u) + \beta \int_{\mathbb{X}} J_{t+1}^T(y)\mathcal{T}(dy|x,u)\}\},$$

converges to $J_\infty^T(x) = J_0^\infty(x)$.

In particular, if one lets $v_0(x) = 0$ for all $x \in \mathbb{X}$, the iterations increase monotonically and converges to the value function. If one is only interested in convergence (and not the monotone behaviour), any initial function $v_0 \in L_\infty(\mathbb{X};\mathbb{R})$ is sufficient.

(ii) The above discussion also applies by considering a contraction on the space $C_b(\mathbb{X})$, if Condition WF (Assumption 5.2.1) holds; in this case, the value function sequence v_n is continuous for every $n \in \mathbb{Z}_+$, and by the completeness of $C_b(\mathbb{X})$ under the supremum norm, so is the limit. ◇

Example 5.4. Consider a controlled Markov chain with state space $\mathbb{X} = \{0,1\}$, action space $\mathbb{U} = \{0,1\}$, and transition kernel for $t \in \mathbb{Z}_+$:

$$P(x_{t+1} = 1|x_t = 0, u_t = 1) = P(x_{t+1} = 1|x_t = 1, u_t = 1) = \alpha$$
$$P(x_{t+1} = 1|x_t = 0, u_t = 0) = P(x_{t+1} = 1|x_t = 1, u_t = 0) = 1 - \alpha.$$

where $\alpha \in (0,1)$. Let a cost function $c(x,u)$, with $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ be given by

$$c(0,1) = c(0,0) = 1 \qquad\qquad c(1,0) = c(1,1) = 2.$$

Suppose that the goal is to minimize the quantity

$$E_0^\gamma[\sum_{t=0}^{\infty} \beta^t c(x_t,u_t)],$$

for a fixed $\beta \in (0,1)$, over all admissible policies $\gamma \in \mathit{\Gamma}_A$. Find an optimal policy and the optimal expected cost explicitly, as a function of α, β (note that the initial condition is $x_0 = 0$).

Solution. Apply value iteration. Take $v_0 \equiv 0$. Then,

$$v_1(0) = \min_u(c(0,u) + \beta E[v_0(x_1)|x_0 = 0, u_0 = u]) = 1$$

$$v_1(1) = \min_u(c(1,u) + \beta E[v_0(x_1)|x_0 = 1, u_0 = u]) = 2$$

and

$$\begin{aligned} v_2(0) &= \min_u(c(0,u) + \beta E[v_1(x_1)|x_0 = 0, u_0 = u]) \\ &= \min_u \Big(c(0,0) + \beta(\alpha v_1(0) + (1-\alpha)v_1(1)), c(0,1) + \beta((1-\alpha)v_1(0) + \alpha v_1(1))\Big) \\ v_2(1) &= \min_u(c(1,u) + \beta E[v_1(x_1)|x_0 = 1, u_0 = u]) \\ &= \min_u \Big(c(1,0) + \beta(\alpha v_1(0) + (1-\alpha)v_1(1)), c(1,0) + \beta((1-\alpha)v_1(0) + \alpha v_1(1))\Big) \end{aligned}$$

We see that if $\alpha < \frac{1}{2}$, then the optimal selection is $u = 1$, leading to

$$v_2(0) = c(0,1) + \beta((1-\alpha)v_1(0) + \alpha v_1(1))$$
$$v_2(1) = c(1,1) + \beta((1-\alpha)v_1(0) + \alpha v_1(1))$$

We see that state 0 is always more desirable than state 1, in that $v_2(0) < v_2(1)$ and this holds for all time stages. Then, $v_n(0) \uparrow v(0)$ and $v_n(1) \uparrow v(1)$ with $v(0) < v(1)$. As a result, for the discounted cost optimality equation

$$v(0) = \min_u (c(0,u) + \beta E[v(x_1)|x_0 = 0, u_0 = u])$$
$$v(1) = \min_u (c(1,u) + \beta E[v(x_1)|x_1 = 0, u_0 = u])$$

using $v(0) < v(1)$, we have that:

$$v(0) = c(0,1) + \beta((1-\alpha)v(0) + \alpha v(1))$$

$$v(1) = c(1,1) + \beta((1-\alpha)v(0) + \alpha v(1))$$

Noting that $v(1) = v(0) + 1$ and solving for v, we obtain

$$v(0) = \frac{1+\beta\alpha}{1-\beta}.$$

For $\alpha \geq \frac{1}{2}$ a parallel argument can be made. ◇

5.5.2 Lipschitz Regularity of Value Functions and the Case with Unbounded Costs

Lipschitz regularity of value functions

A further regularity property is the following. In the following W_1 is the Wasserstein metric on probability measures; see Appendix D. The following property will be useful later, when we study approximation and learning theoretic applications.

Assumption 5.5.1 *[Condition WA] Let $d(\cdot,\cdot)$ denote the metric on $\mathbb{X}$. We assume that for some K_1, K_2:*

(a) $|c(x,u) - c(y,u)| \leq K_1 d(x,y)$; that is, $c(\cdot,u)$ is K_1-Lipschitz (denoted with the notation $c(\cdot,u) \in \mathrm{Lip}(\mathbb{X}, K_1)$).

(b) $W_1(\mathcal{T}(dx_1|x_0 = x, u_0 = u), \mathcal{T}(dx_1|x_0 = y, u_0 = u)) \leq K_2 d(x,y)$.

Theorem 5.5.3 *[170] [270, Theorem 4.37] Suppose that Assumptions 5.2.1 and 5.5.1 hold, and $\beta K_2 < 1$. Then, the solution to*

$$v = \mathbb{T}(v)$$

is Lipschitz with coefficient $K = \frac{K_1}{1-\beta K_2}$.

Proof Let $f \in \mathrm{Lip}(\mathbb{X}, k)$, that is f be k-Lipschitz for some $k \in \mathbb{R}$. Then,

$$\begin{aligned} |\,\mathbb{T}f(z) - \mathbb{T}f(y)\,| &\leq \max_{u \in \mathrm{U}} \Bigg\{ |c(z,u) - c(y,u)| \\ &\quad + \beta \left| \int_{\mathbb{X}} f(x_1)\eta(dx_1|z,u) - \int_{\mathbb{X}} f(x_1)\eta(dx_1|y,u) \right| \Bigg\} \\ &\leq K_1 d(z,y) + \beta k K_2 d(z,y) = (K_1 + \beta k K_2)\, d(z,y) =: M_1 d(z,y). \end{aligned} \tag{5.37}$$

By induction we have for all $n \geq 2$

$$\mathbb{T}^n f \in \mathrm{Lip}\,(\mathbb{X}, M_n)\,,$$

where $M_n = K_1 + \beta K_2 M_{n-1}$ and thus $M_n = K_1 \sum_{i=0}^{n-1} (\beta K_2)^i + k\,(\beta K_2)^n$. Taking $k \leq \frac{K_1}{1-\beta K_2}$, we certify that the fixed point satisfies the desired Lipschitz continuity, as the sequence M_n monotonically converges to $\frac{K_1}{1-\beta K_2}$. Hence, $\mathbb{T}^n f \in \mathrm{Lip}\left(\mathbb{X}, \frac{K_1}{1-\beta K_2}\right)$ for all n, and therefore, the unique solution to the fixed point equation satisfies

$$v \in \mathrm{Lip}\left(\mathbb{X}, \frac{K_1}{1-\beta K_2}\right) \tag{5.38}$$

$\diamond$

When $\mathbb{X}$ is not compact, note that we may still need to verify (5.31) to claim optimality.

Remark 5.5. We finally note that similar contraction arguments can also be applied to functions that are not necessarily continuous, but only lower semi-continuous bounded functions, which also constitute a Banach space under the supremum norm.

A further contraction argument for unbounded costs

As discussed earlier, one could follow the iteration method for the unbounded case (as in the proof of Theorem 5.5.1), whereas the contraction method in the proof of Theorem 5.5.2 holds for the bounded cost case. The contraction method can also be adjusted for the unbounded case under further conditions: If the cost is not bounded, one can define a weighted sup-norm (called an f-norm): $\|c\|_f = \sup_x |\frac{c(x)}{f(x)}|$, where f is a positive function uniformly bounded from below by a positive number. The contraction discussion above will apply to this context with such a consideration, provided that the value function v used in the contraction analysis can be shown to satisfy $\|v\|_f < \infty$. For a suitable function w, let $B_w(\mathbb{X})$ denote the Banach space of measurable functions with a bounded w-norm. We state the corresponding results formally in the following. We state two sets of conditions, one corresponds an unbounded function generalization of strong continuity and the other of weak continuity conditions.

Assumption 5.5.2 *(i) The one stage cost function $c(x,u)$ is nonnegative and continuous in u for every x.*

(ii) The stochastic kernel $\mathcal{T}(\cdot\,|x,u)$ is strongly continuous in u for every x, i.e., if $u_k \to u$, then $\int u(y)\mathcal{T}(dy|x,u_k) \to \int u(y)\mathcal{T}(dy|x,u)$ for every measurable and bounded function u.

(iii) $\mathbb{U}$ is compact.

(iv) There exist nonnegative real numbers M and $\alpha \in [1, \frac{1}{\beta})$, and a weight function $w : \mathbb{X} \to [1,\infty)$ such that for each $z \in \mathbb{X}$, we have

$$\sup_{u\in\mathbb{U}} |c(x,u)| \le Mw(x), \tag{5.39}$$

$$\sup_{u\in\mathbb{U}} \int_{\mathbb{X}} w(y)\mathcal{T}(dy|x,u) \le \alpha w(x), \tag{5.40}$$

and $\int_{\mathbb{X}} w(y)\mathcal{T}(dy|x,u)$ is continuous in u for every x.

Assumption 5.5.3 *(i) The one stage cost function $c(x,u)$ is nonnegative and continuous in (x,u).*

(ii) The stochastic kernel $\mathcal{T}(\cdot\,|x,u)$ is weakly continuous in $(x,u) \in \mathbb{X}\times\mathbb{U}$, i.e., if $(x_k,u_k) \to (x,u)$, then $\mathcal{T}(\cdot\,|x_k,u_k) \to \mathcal{T}(\cdot\,|x,u)$ weakly.

(iii) $\mathbb{U}$ is compact.

(iv) There exist nonnegative real numbers M and $\alpha \in [1, \frac{1}{\beta})$, and a continuous weight function $w : \mathbb{X} \to [1,\infty)$ such that for each $z \in \mathbb{X}$, we have

$$\sup_{u\in\mathbb{U}} |c(x,u)| \le Mw(x), \tag{5.41}$$

$$\sup_{u\in\mathbb{U}} \int_{\mathbb{X}} w(y)\mathcal{T}(dy|x,u) \le \alpha w(x), \tag{5.42}$$

and $\int_{\mathbb{X}} w(y)\mathcal{T}(dy|x,u)$ is continuous in (x,u).

Define the operator $\mathbb{T}$ on the set of real-valued measurable functions on $\mathbb{X}$ as

$$\mathbb{T}v(z) = \min_{a\in\mathbb{U}} \left\{ c(z,a) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|z,a) \right\}. \tag{5.43}$$

It can be proved that $\mathbb{T}$ is a contraction operator mapping $B_w(\mathbb{X})$ into itself with modulus $\sigma = \beta\alpha$ (see [166, Lemma 8.5.5]); that is,

$$\|\mathbb{T}u - \mathbb{T}v\|_w \leq \beta\|u - v\|_w \text{ for all } u, v \in B_w(\mathbb{X}).$$

Theorem 5.5.4 *[166, Theorem 8.3.6] [166, Lemma 8.5.5] Suppose Assumption 5.5.2 (or 5.5.3) holds. Then, the value function J^* is the unique fixed point in $B_w(\mathbb{X})$ (or $B_w(\mathbb{X}) \cap C(\mathbb{X})$ under Assumption 5.5.3) of the contraction operator $\mathbb{T}$, i.e.,*

$$J^* = \mathbb{T}J^*. \tag{5.44}$$

Furthermore, a deterministic stationary policy f^* is optimal if and only if

$$J^*(z) = c(z, f^*(z)) + \beta \int_{\mathbb{X}} J^*(y)\mathcal{T}(dy|z, f^*(z)). \tag{5.45}$$

Finally, there exists a deterministic stationary policy f^* which is optimal (and thus satisfies (5.45)).

The proof follows from [166, Theorem 8.3.6]. See also [166, Lemma 8.5.5].

5.6 Regularity of Transition Kernels and Optimal Value Functions

We have seen in the chapter that continuity and regularity of transition kernels play a significant role for carrying out optimality analysis. Later on we will see that these are also important for approximations, robustness, and learning theoretic results and applications.

We review the following regularity properties for the transition kernels:

(i) $\mathcal{T}(\cdot|x,u)$ is said to be weakly continuous (weak Feller) in (x,u), if $\mathcal{T}(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ weakly for any $(x_n,u_n) \to (x,u)$.

(ii) $\mathcal{T}(\cdot|x,u)$ is said to be strongly continuous (strong Feller) in u for every x, if $\mathcal{T}(\cdot|x,u_n) \to \mathcal{T}(\cdot|x,u)$ setwise for any $u_n \to u$ for every fixed $x \in \mathbb{X}$.

(iii) $\mathcal{T}(\cdot|x,u)$ is said to be continuous under total variation in (x,u), if $\|\mathcal{T}(\cdot|x_n,u_n) - \mathcal{T}(\cdot|x,u)\|_{TV} \to 0$ for any $(x_n,u_n) \to (x,u)$.

(iv) $\mathcal{T}(\cdot|x,u)$ is said to be continuous under the first order Wasserstein distance in (x,u), if

$$W_1(\mathcal{T}(\cdot|x_n,u_n), \mathcal{T}(\cdot|x,u)) \to 0$$

for any $(x_n,u_n) \to (x,u)$. To ensure continuity of $\mathcal{T}$ with respect to the first order Wasserstein distance, in addition to weak continuity, we may assume that there exists a function $g : [0,\infty) \to [0,\infty)$ such that as $t \to \infty$, $\frac{g(t)}{t} \uparrow \infty$, and

$$\sup_{(x,u)\in K\times\mathbb{U}} \int g(\|y\|)\, \mathcal{T}(dy|x,u) < \infty$$

for any compact $K \subset \mathbb{X}$. Note that the latter condition implies uniform integrability of the collection of random variables with probability measures $\mathcal{T}(dx_1|X_0 = x_n, U_0 = u_n)$ as $(x_n,u_n) \to (x,u)$, which coupled with weak convergence can be shown to imply convergence under the Wasserstein distance. To see this, let $(x_n,u_n) \to (x_\infty,u_\infty)$ and X_n random variables with law $\mathcal{T}(\cdot|x_n,u_n)$ satisfying $X_n \to X_\infty$ a.s., which is possible by Skorohod's theorem

(Theorem B.3.5). Then the above condition implies uniform integrability of $\{X_n\}$, and thus $E[\|X_n - X_\infty\|] \to 0$. Then $W_1(\mathcal{T}(\cdot|x_n, u_n), \mathcal{T}(\cdot|x_\infty, u_\infty)) \to 0$.

Example 5.6. Some example models satisfying these regularity properties are as follows:

(i) For a model with the dynamics $x_{t+1} = f(x_t, u_t, w_t)$, the induced transition kernel $\mathcal{T}(\cdot|x, u)$ is weakly continuous in (x, u) if $f(x, u, w)$ is a continuous function of (x, u), since for any continuous and bounded function g

$$\int g(x_1)\mathcal{T}(dx_1|x_n, u_n) = \int g(f(x_n, u_n, w))\mu(dw)$$
$$\to \int g(f(x, u, w))\mu(dw) = \int g(x_1)\mathcal{T}(dx_1|x, u)$$

where μ denotes the probability measure of the noise process. If we also have that $\mathbb{X}$ is compact, the transition kernel $\mathcal{T}(\cdot|x, u)$ is also continuous under the first order Wasserstein distance.

(ii) For a model with the dynamics $x_{t+1} = f(x_t, u_t) + w_t$, the induced transition kernel $\mathcal{T}(\cdot|x, u)$ is continuous under total variation in (x, u) if $f(x, u)$ is a continuous function of (x, u), and w_t admits a continuous density function.

(iii) In general, if the transition kernel admits a continuous density function f so that $\mathcal{T}(dx_1|x, u) = f(x_1, x, u)dx_1$, then $\mathcal{T}(dx_1|x, u)$ is continuous in total variation. This follows from an application of Scheffé's Lemma [44, Theorem 16.12]. In particular, we can write that

$$\|\mathcal{T}(\cdot|x_n, u_n) - \mathcal{T}(\cdot|x, u)\|_{TV} = \int_{\mathbb{X}} |f(x_1, x_n, u_n) - f(x_1, x, u)|dx_1 \to 0.$$

(iv) For a model with the dynamics $x_{t+1} = f(x_t, u_t, w_t)$, if f is Lipschitz continuous in (x, u) pair such that, there exists some $\alpha < \infty$ with

$$|f(x_n, u_n, w) - f(x, u, w)| \leq \alpha \left(|x_n - x| + |u_n - u|\right),$$

we can then bound the first order Wasserstein distance between the corresponding kernels with α:

$$W_1\left(\mathcal{T}(\cdot|x_n, u_n), \mathcal{T}(\cdot|x, u)\right) = \sup_{Lip(g)\leq 1} \left| \int g(x_1)\mathcal{T}(dx_1|x_n, u_n) - \int g(x_1)\mathcal{T}(dx_1|x, u) \right|$$
$$= \sup_{Lip(g)\leq 1} \left| \int g(f(x_n, u_n, w))\mu(dw) - \int g(f(x, u, w))\mu(dw) \right|$$
$$\leq \int |f(x_n, u_n, w) - f(x, u, w)|\, \mu(dw) \leq \alpha \left(|x_n - x| + |u_n - u|\right).$$

We next review the following regularity properties, which serves as a summary of Theorem 5.2.1, and Theorems 5.5.2 and 5.5.3 lead to the following.

Theorem 5.6.1 (Regularity for Finite Horizon Cost Criterion) *Consider (5.1). Under Assumptions 5.2.1 and 5.2.2 there exists a minimizing control policy $\{f_t, t \geq 0\}$ which is Markov (and thus in Γ_M). Furthermore, under Assumption 5.2.1, the function J_t, for all t, is continuous, under 5.2.2 J_t is Borel measurable, and under Assumptions 5.5.1 and 5.2.1 it is Lipschitz (in the latter case if c_N exists, it is assumed to be Lipschitz).*

Theorem 5.6.2 (Regularity for Discounted Cost Criterion) *Consider (5.21). Under Assumptions 5.2.1 and 5.2.2 there exists a minimizing control policy which is stationary (and thus in Γ_S) without loss. Furthermore, under Assumption 5.2.1, the function J_β is continuous, under 5.2.2, it is Borel measurable, and under Assumption 5.5.1 (with $\beta K_2 < 1$) and Assumption 5.2.1, it is Lipschitz*

5.7 Exercises

Exercise 5.7.1 *An investor's wealth dynamics is given by the following:*

$$x_{t+1} = u_t w_t,$$

where $\{w_t\}$ is an i.i.d. $\mathbb{R}_+$-valued stochastic process with $E[\sqrt{w_t}] = 1$ and u_t is the investment of the investor at time t. The investor has access to the past and current wealth information and his previous actions. The goal is to maximize:

$$J(x_0, \gamma) = E^{\gamma}_{x_0}[\sum_{t=0}^{T-1} \sqrt{x_t - u_t}].$$

The investor's action set for any given x is: $\mathbb{U}(x) = [0, x]$. His initial wealth is given by x_0.

Formulate the problem as an optimal stochastic control problem by clearly identifying the state space, the control action space, the information available at the controller at any time, the transition kernel and a cost functional mapping the actions and states to $\mathbb{R}$.

Find an optimal policy.

Hint: For $\alpha > 0$, $\sqrt{x-u} + \alpha\sqrt{u}$ is a concave function of u for $0 \leq u \leq x$ and its maximum is computed when the derivative of $\sqrt{x-u} + \alpha\sqrt{u}$ is set to zero.

Exercise 5.7.2 *Consider the following linear system:*

$$x_{t+1} = Ax_t + Bu_t + w_t,$$

where $x \in \mathbb{R}^n, u \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$. Suppose $\{w_t\}$ is i.i.d. zero-mean Gaussian with a given covariance matrix $E[w_t w_t^T] = W$ for all $t \geq 0$.

The goal is to obtain

$$\inf_{\gamma} J(x, \gamma),$$

where

$$J(x, \gamma) = E^{\gamma}_{x}[\sum_{t=0}^{T-1} x_t^T Q x_t + u_t^T R u_t + x_T^T Q_T x_T],$$

with $R, Q, Q_T > 0$ (that is, these matrices are positive definite).

a) Show that there exists an optimal policy.

b) Obtain the Dynamic Programming recursion for the optimal control problem. Is the optimal control policy Markov? Is it stationary?

c) For $T \to \infty$, if (A, B) is controllable and with $Q = C^T C$ and (A, C) is observable, prove that the optimal policy is stationary.

Exercise 5.7.3 (Optimality of Threshold Policies) *[[36]] Consider an inventory-production system given by*

$$x_{t+1} = x_t + u_t - w_t,$$

where x_t is $\mathbb{R}$-valued, with the one-stage cost

$$c(x_t, u_t, w_t) = bu_t + h\max(0, x_t + u_t - w_t) + p\max(0, w_t - x_t - u_t)$$

Here, b is the unit production cost, h is the unit holding (storage) cost and p is the unit shortage cost; here we take $p > b$. At any given time, the decision maker can take $u_t \in \mathbb{R}_+$. The demand variable $w_t \sim \mu$ is a $\mathbb{R}_+$-valued i.i.d.

process, independent of x_0, with a finite mean where μ is assumed to admit a probability density function. The goal is to minimize

$$J(x,\gamma) = E_x^{\gamma}[\sum_{t=0}^{T-1} c(x_t, u_t, w_t)]$$

The controller at time t has access to $I_t = \{x_s, u_s, s \leq t-1\} \cup \{x_t\}$.

Obtain a recursive form for the optimal solution. In particular, show that the solution is of threshold type: There exists a sequence of real-numbers s_t so that the optimal solution is of the form: $u_t = 0 \times 1_{\{x_t \geq s_t\}} + (s_t - x_t) \times 1_{\{x_t < s_t\}}$. See [35] for a detailed analysis of this problem.

We note that the above is a relatively simplified model (e.g., the inventory can be negative and the cost/reward function is simple) and such inventory problems can be made more general. Nonetheless, the solution method, via a convex value function analysis and associated optimality arguments leading to threshold optimality, applies nearly identically to many problems in stochastic control.

Exercise 5.7.4 (Optimal Stopping) *[36]*

Consider a burglar who is considering retirement. His goal is to maximize his earnings up to time T. At any time, he can either continue his profession to steal an amount of w_t which is an i.i.d. $\mathbb{R}_+$-valued random process (he adds this amount to his wealth), or retire.

However, each time he attempts burglary, there is a chance that he gets caught and he loses all of his savings (and cannot work any further); this happens according to an i.i.d. Bernoulli process so that he gets caught with probability p at each time stage when he is attempting to steal.

Assume that his initial wealth is $x_0 = 0$. His goal is to maximize $E[x_T]$. Find his optimal policy for $0 \leq t \leq T-1$.

Note: Such problems where a decision maker can quit or stop a process are known as optimal stopping problems.

Exercise 5.7.5 (The Secretary Problem) *Consider a manager who interviews N candidates for a position. The manager wishes to maximize the probability of finding the best candidate. The candidates are interviewed in succession according to a random order (uniformly distributed given all possible permutations). If a candidate is rejected, that candidate is no longer available and if a candidate is selected, the process is over. The decisions must be made causally given the available information up to that time, that is if the order is $X_1, X_2, \cdots, X_t$, the policy can only use the information generated by $\sigma(X_1, \cdots, X_t)$. What is the optimal strategy?*

Hint: Apply dynamic programming. At time N, $J_N = \frac{1}{N}$ since at time N the past is given and the best one can hope for is that the best candidate is the final one, which happens with probability $\frac{1}{N}$. Now, consider $m = N-1$: $J_{N-1} = \frac{1}{N-1}(\max(\frac{N-1}{N}), J_N) + \frac{N-2}{N-1} J_N$. Here, the first event is the probability that the $N-1$st candidate is the best among the first $N-1$ candidates, and in this case the manager needs to decide to stop or wait for the future (which he does by comparing whether the best among the first $N-1$ is the best among all, or whether he should skip and move to the next time stage, N). The second event is with probability $\frac{N-2}{N-1}$ in which case the best among the first $N-1$ candidates is not the $N-1$st one, in which case the manager has to wait. Continuing on with this logic:

$$J_m = \frac{1}{m}\max(\frac{m}{N}, J_{m+1}) + \frac{m-1}{m} J_{m+1}$$

where $\frac{m}{N}$ is the probability that the best in the first m is the best among all. Now, computing explicitly, we obtain

$$J_N = 1/N; J_{N-1} = \frac{N-2}{N}(\frac{1}{N-2} + \frac{1}{N-1}); \cdots ..$$

and with $J_m = \frac{m-1}{N}(\frac{1}{m-1} + \frac{1}{m} + \cdot + \frac{1}{N-1})$, and we stop when $\frac{m}{N} \geq J_{m+1} = \frac{m}{N}(\frac{1}{m} + \frac{1}{m+1} + \cdot + \frac{1}{N-1})$. When N is large enough, the above suggests that we stop as soon as the best candidate thus far has been spotted at time m when $1 \geq (\frac{1}{m} + \frac{1}{m+1} + \cdot + \frac{1}{N-1}) \approx \log(N/m)$ which means that an optimal policy is around when $\sum_m^{N-1} \frac{1}{k} < 1$.

Approximately, this means that $m^* = \frac{N}{e}$ is a nearly optimal rule for large N if the m^*th candidate is the best candidate seen until then.

Exercise 5.7.6 *A fishery manager annually has x_t units of fish and sells $u_t x_t$ of these where $u_t \in [0,1]$. With the remaining ones, the next year's production is given by the following model*

$$x_{t+1} = w_t x_t (1 - u_t) + v_t,$$

with x_0 is given and $\{w_t, v_t\}$ is a sequence of mutually independent, identically distributed sequence of random variables with $w_t \geq 0$, $v_t \geq 0$ for all t and therefore $E[w_t] = \bar{w} \geq 0$ and $E[v_t] = \bar{v} > 0$.

At time T, he sells all of the fish. The goal is to maximize the profit over the time horizon $0 \leq t \leq T-1$.

a) Formulate the problem as an optimal stochastic control problem by clearly identifying the state, the control actions, the information available at the controller, the transition kernel and a cost functional mapping the actions and states to $\mathbb{R}$.

b) Does there exist an optimal policy? If it does, compute the optimal control policy as a dynamic programming recursion.

Exercise 5.7.7 *A common example in mathematical finance applications is the portfolio selection problem where a controller (investor) would like to optimally allocate his wealth between a stochastic stock market and a market with a guaranteed income : Consider a stock with an i.i.d. random return σ_t and a bank account with fixed interest rate $r > 0$. These are modeled by:*

$$X_{t+1} = X_t u_t (1 + \sigma_t) + X_t (1 - u_t)(1 + r), \quad X_0 = 1$$

and

$$X_{t+1} = X_t (1 + r + u_t (\sigma_t - r))$$

Here, $u_t \in [0,1]$ denotes the proportion of the money that the investor invests in the stock market. Suppose that the goal is to maximize $E[\log(X_T)]$. Then, we can write:

$$\log(X_T) = \log(\prod_{k=0}^{T-1} \frac{X_{k+1}}{X_k}) = \sum_{k=0}^{T-1} \log((1 + r + u_t(\sigma_t - r))) \tag{5.46}$$

Formulate the problem as an optimal stochastic control problem by clearly identifying the state and the control action spaces, the information available at the controller, the transition kernel, and a cost functional mapping the actions and states to $\mathbb{R}$. Find the optimal policy.

Exercise 5.7.8 *We will illustrate dynamic programming by considering a simplified version of a setup in [156]. Consider a two server-station network; where a router routes the incoming traffic, as is depicted in Figure 5.1.*

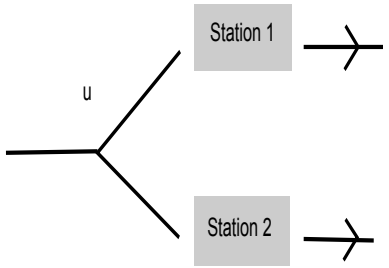

Fig. 5.1

Customers arrive according to a (continuous-time) Poisson process of rate λ. The router routes to station 1 with probability u and second station with probability $1-u$. The router has access to the number of customers at both of the queues, while implementing her policy.

Station 1 has a service time distribution which is exponential with rate μ_1, and Station 2 with $\mu_2 = \mu_1$, as well. After some computation, we find out that the controlled transition kernel is given by the following:

$$P(q_{t+1}^1 = q_t^1 + 1, q_{t+1}^2 = q_t^2 | q_t^1, q_t^2) = \lambda \frac{u}{\lambda + 2\mu_1}$$

$$P(q_{t+1}^1 = q_t^1, q_{t+1}^2 = q_t^2 + 1 | q_t^1, q_t^2) = \lambda \frac{(1-u)}{\lambda + 2\mu_1}$$

$$P(q_{t+1}^1 = \max(q_t^1 - 1, 0), q_{t+1}^2 = q_t^2 | q_t^1, q_t^2) = \frac{\mu_1}{\lambda + 2\mu_1}$$

$$P(q_{t+1}^1 = q_t^1, q_{t+1}^2 = \max(0, q_t^2 - 1) | q_t^1, q_t^2) = \frac{\mu_1}{\lambda + 2\mu_1}$$

There is also a holding cost per unit time. The holding cost at Station 1 is $c_1 > 0$ and the cost at Station 2 is $c_2 > 0$. That is if there are q_t^1 customers, the cost is $c_1 q_t^1$ at station 1 at time t and likewise for Station 2.

The goal of the router is to minimize the expected total holding cost from time 0 to some time $T \in \mathbb{N}$, where the total cost is

$$\sum_{t=0}^{T} c_1 q_t^1 + c_2 q_t^2.$$

a) Express the problem as a dynamic programming problem, up until time T. That is; where does the control action live? What is the state space? What is the transition kernel for the controlled Markov Chain?

Write down the dynamic programming recursion, starting from time T and going backwards.

b) Suppose that $c_1 = c_2$. Let $J_t(q_t^1, q_t^2)$ be the value function at time t (that is the current cost and the cost to go).

Via dynamic programming, prove the following:

For a given t, if, whenever $0 \leq q_t^1 \leq q_t^2$ we have that

$$J_t(q_t^1, q_t^2) \leq J_t(q_t^1 - 1, q_t^2 + 1),$$

then the same applies for $J_{t-1}(.,.)$, for $t \geq 1$. With the above, prove that an optimal control policy is given by:

$$u_t = 1_{\{q_t^1 \leq q_t^2\}},$$

for all t values.

Exercise 5.7.9 *Consider a scalar linear system with the following dynamics:*

$$x_{t+1} = ax_t + bu_t + w_t,$$

where $\{w_t\}$ is i.i.d Gaussian with zero-mean and unit variance. Suppose that the controller has access to $I_t = \{x_{[0,t]}, u_{[0,t-1]}\}$ at time t. Suppose that the initial state is $x_0 = x$ for some $x \in \mathbb{R}$. We wish to find for some $\beta \in (0,1)$:

$$\inf_{\gamma} J(x_0, \gamma) = E_x^{\gamma}[\sum_{t=0}^{\infty} \beta^t (qx_t^2 + ru_t^2)],$$

for $q \geq 0$ and $r > 0$.

Compute the optimal control policy and the optimal cost.

Hint: Use Lemma 5.5.3. Start with a finite horizon version, and apply dynamic programming, obtain the solution and take the finite horizon to infinity. This is also equivalent to applying value iteration with $v_0(x) = 0$ for all $x \in \mathbb{R}$. You will see that a recursion with $v_t(x) = C_t x^2 + D_t$ will be obtained and C_t and D_t will have limits as $t \to \infty$, C and D, respectively. The optimal control will be stationary and deterministic:

$$u_t = \gamma(x_t) = -(r + \beta C b^2)^{-1} \beta a b C x_t, \qquad t \geq 0.$$

Thus, you need to find C and D.

Exercise 5.7.10 *Consider a controlled Markov chain with state space $\mathbb{X} = \{0, 1\}$, action space $\mathbb{U} = \{0, 1\}$, and transition kernel for $t \in \mathbb{Z}_+$:*

$$P(x_{t+1} = 1 | x_t = 0, u_t = 1) = P(x_{t+1} = 1 | x_t = 0, u_t = 0) = \alpha$$

where $\alpha \in (0, 1)$. Furthermore,

$$P(x_{t+1} = 1 | x_t = 1, u_t = 0) = P(x_{t+1} = 1 | x_t = 1, u_t = 1) = \frac{1}{2}.$$

Let a cost function $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ be given by

$$c(0, 1) = \kappa \in \mathbb{R}_+, \quad c(0, 0) = 1$$

$$c(1, 0) = \frac{1}{2}, \qquad c(1, 1) = 1.$$

Suppose that the goal is to minimize the quantity

$$E_0^{\gamma}[\sum_{t=0}^{\infty} \beta^t c(x_t, u_t)],$$

for a fixed $\beta \in (0, 1)$, over all admissible policies $\gamma \in \Gamma_A$.

Find an **optimal policy** *and the* **optimal expected cost** *explicitly, as a function of α, β, κ.*

6

Partially Observed Markov Decision Processes, Non-Linear Filtering, and the Kalman Filter

As discussed earlier in Section 2.4, for a large class of problems the controller does not have access to the state process, but may have access to some partial information obtained via noisy measurements. In particular, in this chapter, we consider systems of the form:

$$x_{t+1} = f(x_t, u_t, w_t), \qquad y_t = g(x_t, v_t). \tag{6.1}$$

Here, x_t is the state (with $x_0 \sim \mu_0$), $u_t \in \mathbb{U}$ is the control, (w_t, v_t) are $(\mathbb{W} \times \mathbb{V})$-valued i.i.d noise processes where w_t is independent of v_t.

The controller only has causal access to the second component $\{y_t\}$ of the process, together with the past applied control actions. An admissible policy $\gamma = \{\gamma_t, t \in \mathbb{Z}_+\}$ is a collection of measurable functions so that γ_t is measurable with respect to $\sigma(I_t)$ with $I_t = \{y_{[0,t]}, u_{[0,t-1]}\}$ at time t. We emphasize the implicit assumption here that the control policy can also (and typically does) depend on the prior probability measure μ_0. We denote the observed history space as: $H_0 := \mathbb{Y}, \quad H_t = H_{t-1} \times \mathbb{Y} \times \mathbb{U}$.

In the following $\mathcal{P}(\mathbb{X})$ denotes the space of probability measures on $\mathbb{X}$, which we assume to be a Polish space. Under the topology of weak convergence, $\mathcal{P}(\mathbb{X})$ is also a Polish space (see Appendix D).

6.1 Enlargement of the State-Space and the Construction of a Controlled Markov Chain

We will see in this section that one could always transform a Partially Observable Markov Decision Problem (POMDP) to a Fully Observed Markov Decision Problem (called a belief-MDP) via an enlargement of the state space and a reformulation of the model. In particular, when $\mathbb{X}, \mathbb{Y}, \mathbb{U}$ are countable (the more general case will be studied later in the chapter in Section 6.3), we obtain via the properties of total probability the following recursion for conditional probability measures, given an admissible policy,

$$\begin{aligned}
\pi_t(x) := P(x_t = x | y_{[0,t]}, u_{[0,t-1]}) &= \frac{P(x_t = x, y_t, u_{t-1} | y_{[0,t-1]}, u_{[0,t-2]})}{\sum_{x \in \mathbb{X}} P(x_t = x, y_t, u_{t-1} | y_{[0,t-1]}, u_{[0,t-1]})} \\
&= \frac{\sum_{x_{t-1} \in \mathbb{X}} P(y_t|x_t) P(x_t|x_{t-1}, u_{t-1}) P^{\gamma}(u_{t-1}|y_{[0,t-1]}, u_{[0,t-2]}) \pi_{t-1}(x_{t-1})}{\sum_{x_{t-1} \in \mathbb{X}} \sum_{x \in \mathbb{X}} P(y_t|x_t = x) P(x_t = x|x_{t-1}, u_{t-1}) P^{\gamma}(u_{t-1}|y_{[0,t-1]}, u_{[0,t-2]}) \pi_{t-1}(x_{t-1})}. \\
&= \frac{\sum_{x_{t-1} \in \mathbb{X}} P(y_t|x_t) P(x_t|x_{t-1}, u_{t-1}) \pi_{t-1}(x_{t-1})}{\sum_{x_{t-1} \in \mathbb{X}} \sum_{x \in \mathbb{X}} P(y_t|x_t = x) P(x_t = x|x_{t-1}, u_{t-1}) \pi_{t-1}(x_{t-1})}. \\
&=: F(\pi_{t-1}, y_t, u_{t-1})(x)
\end{aligned} \tag{6.2}$$

Notice that the right hand side does not depend on the policy γ, therefore the conditional expectation is policy-independent. We will see shortly that the conditional measure process forms a controlled Markov chain in $\mathcal{P}(\mathbb{X})$.

Note that in the above analysis $P(u_{t-1}|y_{[0,t-1]}, u_{[0,t-2]})$ is determined by the control policy, and $P(x_t|x_{t-1}, u_{t-1})$ is determined by the transition kernel $\mathcal{T}$ of the controlled Markov chain.

The result above leads to the following.

Theorem 6.1.1 *The process $\{\pi_t, u_t\}$ is a controlled Markov chain. That is, under any admissible control policy, given the action at time $t \geq 0$ and π_t, π_{t+1} is conditionally independent from $\{\pi_s, u_s, s \leq t-1\}$.*

We will prove the result for the case where $\mathbb{Y}$ is countable. For the more general case, see Section 6.3.

Proof. Let $D \in \mathcal{B}(\mathcal{P}(\mathbb{X}))$. From (6.2), with Y_{y+1} denoted with the capital letter to emphasize its randomness, we have under any admissible policy,

$$
\begin{aligned}
&P(\pi_{t+1} \in D|\pi_s, u_s, s \leq t) = P(F(\pi_t, Y_{t+1}, u_t) \in D|\pi_s, u_s, s \leq t) \\
&= P(F(\pi_t, Y_{t+1}, u_t) \in D, Y_t \in \mathbb{Y}|\pi_s, u_s, s \leq t) \\
&= \sum_{y \in \mathbb{Y}} P(F(\pi_t, y_{t+1}, u_t) \in D, y_{t+1} = y|\pi_s, u_s, s \leq t) \\
&= \sum_{y \in \mathbb{Y}} P(F(\pi_t, y_{t+1}, u_t) \in D|y_{t+1} = y, \pi_s, u_s, s \leq t) P(y_{t+1} = y|\pi_s, u_s, s \leq t) \\
&= \sum_{y \in \mathbb{Y}} 1_{\left\{F(\pi_t, y, u_t) \in D\right\}} P(y_{t+1} = y|\pi_t, u_t) \\
&= \sum_{y \in \mathbb{Y}} 1_{\left\{F(\pi_t, y, u_t) \in D\right\}} \left(\sum_{x' \in \mathbb{X}} \sum_{x \in \mathbb{X}} P(y_{t+1} = y|x_{t+1} = x') P(x_{t+1} = x'|x_t = x, u_t) \pi_t(x) \right) \\
&= P(\pi_{t+1} \in D|\pi_t, u_t) =: \eta(D|\pi_t, u_t)
\end{aligned} \tag{6.3}
$$

Observe that the kernel η does not depend on the policy (and thus it is *policy-independent*). We still need to show that the expression $P(\pi_{t+1} \in \cdot|\pi_t, u_t) : \mathcal{P}(\mathbb{X}) \times \mathbb{U} \to \mathcal{P}(\mathcal{P}(\mathbb{X}))$ is a regular conditional probability measure; that is, for every fixed $D \in \mathcal{B}(\mathcal{P}(\mathbb{X}))$, $P(\pi_{t+1} \in D|\pi_t, u_t)$ is a measurable function on $\mathcal{P}(\mathbb{X}) \times \mathbb{U}$ and for every π_t, u_t, the map is a probability measure on $\mathcal{P}(\mathbb{X})$. The rest of the proof follows in Section 6.3. ◇

Let the cost function to be minimized be

$$E^{\gamma}_{\mu_0}[\sum_{t=0}^{T-1} c(x_t, u_t)],$$

where $E^{\gamma}_{\mu_0}[\cdot]$ denotes the expectation over all sample paths with initial state measure given by μ_0 under policy $\gamma = \{\gamma_0, \gamma_1, \cdots\}$. We can transform the system into a fully observed Markov model as follows. Using the law of the iterated expectations (Theorem 4.1.3), write the total cost as

$$E^{\gamma}_{\mu_0}[\sum_{t=0}^{T-1} c(x_t, u_t)] = E^{\gamma}_{\mu_0}[\sum_{t=0}^{T-1} E^{\gamma}[c(x_t, u_t)|I_t]].$$

Given a policy γ with $u_t = \gamma_t(I_t)$, we have that

$$
\begin{aligned}
&E^{\gamma}_{\mu_0}[\sum_{t=0}^{T-1} c(x_t, u_t)] = E^{\gamma}_{\mu_0}\left[\sum_{t=0}^{T-1} E^{\gamma}\left[c(x_t, \gamma_t(I_t))|I_t\right]\right] \\
&= E^{\gamma}_{\mu_0}\left[\sum_{t=0}^{T-1} \left(\sum_{x \in \mathbb{X}} \sum_{u \in \mathbb{U}} P^{\gamma}(x_t = x|I_t, u_t) P^{\gamma}(u_t = u|I_t) c(x, u) \right)\right] \\
&= E^{\gamma}_{\mu_0}\left[\sum_{t=0}^{T-1} \left(\sum_{x \in \mathbb{X}} P^{\gamma}(x_t = x|I_t) c(x, \gamma_t(I_t)) \right)\right]
\end{aligned}
$$

$$= E^{\gamma}_{\mu_0}\Big[\sum_{t=0}^{T-1}\Big(\sum_{x\in\mathbb{X}}\pi_t(x)c(x,\gamma_t(I_t))\Big)\Big] \tag{6.4}$$

Notice that $P^{\gamma}(x_t = x|I_t, u_t) = P^{\gamma}(x_t = x|I_t) = P(x_t = x|I_t)$ is policy-independent as noted earlier. At this point, we should pause and reflect on Theorem 6.1.1 and (Blackwell's) Theorem 5.1.1, to conclude that without any loss a policy, for finite horizons, could use π_t and t, by the following reasoning: Define a stage-wise cost function $\tilde{c} : \mathcal{P}(\mathbb{X}) \times \mathbb{U} \to \mathbb{R}_+$ as

$$\tilde{c}(\pi, u) = \sum_{\mathbb{X}} c(x, u)\pi(x), \quad \pi \in \mathcal{P}(\mathbb{X}), \tag{6.5}$$

Observe that an admissible control policy will select u_t as a function of I_t. However, we know that for a finite horizon problem, by Blackwell's Theorem 5.1.1, any admissible policy can be replaced with one which only uses π_t without any loss (since π_t, u_t forms a controlled Markov chain), and therefore without any loss, we can restrict our search space to policies which are Markov (that is which only use π_t and t).

In view of the preceding discussion, it follows then that an optimal solution to the following minimization for the problem

$$E^{\gamma}_{\mu_0}[\sum_{t=0}^{T-1}\tilde{c}(\pi_t, u_t)],$$

for the controlled belief-MDP model with kernel η (6.3), is also optimal for the original problem (6.4), where the initial state distribution μ_0 for the belief-MDP is the probability measure on $\pi_0(\cdot) = P(x_0 \in \cdot|y_0)$ induced by the initial probability measure μ_0 on x_0 and the measurement variable y_0.

Let η be the transition kernel defined with (6.3). It follows then that $(\mathcal{P}, \mathbb{U}, \eta, \tilde{c})$ defines a completely observable controlled Markov process (also called a belief-MDP).

Thus, one can obtain the optimal solution by using the solution of the filtering equation (6.2) as a sufficient statistic, as Markov policies (policies that use the Markov state as their sufficient statistics) are optimal for control of Markov chains, under the previously studied measurable selection conditions (see Section 5.2) which require some regularity conditions. We will discuss these later in the chapter.

We call the control policies which use π as their information to generate control as **separated** control policies; as one first generates the belief π_t via the filtering equation, and then generates the control via π_t.

We note here that some of the first separation results for partially observed Markov Decision Processes were reported in [352], [296], and [262], among others.

Separation will be particularly consequential in the context of linear Gaussian systems: A Gaussian probability measure can be uniquely identified by knowing the mean and the covariance of the Gaussian random variable. This makes the analysis for estimating a Gaussian random variable particularly simple to perform, since the conditional estimate of a partially observed (through an additive Gaussian noise) Gaussian random variable is a linear/affine function of the observed variable and the non-linear filtering equation (6.2) becomes significantly simpler. Recall that a Gaussian measure with mean μ and covariance matrix K_{XX} has the following density:

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|K_{XX}|^{1/2}}e^{-1/2((x-\mu)^T K_{XX}^{-1}(x-\mu))},$$

and thus it suffices to compute the mean and the covariance matrix to define the Gaussian probability measure.

6.2 The Linear Quadratic Gaussian (LQG) Problem and Kalman Filtering

6.2.1 A Supporting Result on Estimation

Lemma 6.2.1 *Let X be a random variable (defined on a probability space $(\Omega, \mathcal{F}, P)$) with a finite second moment and $R > 0$ (that is, a positive definite matrix). The following holds*

$$\inf_{g \in \mathbb{M}(\mathbb{Y})} E[(X - g(Y))^T R(X - g(Y))] = E[(X - G(Y))^T R(X - G(Y))],$$

where $\mathbb{M}(\mathbb{Y})$ denotes the set of measurable functions from $\mathbb{Y}$ to $\mathbb{R}$ and where $G(y) = E[X|Y = y]$ almost surely.

Before we state the proof, it is useful to emphasize that there are setups where the measurability assumption is not superfluous. See Exercise 4.5.13.

Proof. Let $G(y) = E[X|Y = y] + h(y)$, for some measurable h; we then have the following through the law of the iterated expectations:

$$\begin{aligned}
&E[(X - E[X|Y] - h(Y))^T R(X - E[X|Y] - h(Y))] \\
&= E[(X - E[X|Y])^T R(X - E[X|Y])] + 2E[(X - E[X|Y])^T Rh(Y)] + E[h^T(Y)Rh(Y)] \\
&= E[(X - E[X|Y])^T R(X - E[X|Y])] + 2E[E[(X - E[X|Y])^T Rh(Y)|Y]] + E[h^T(Y)Rh(Y)] \qquad (6.6) \\
&= E[(X - E[X|Y])^T R(X - E[X|Y])] + E[h^T(Y)Rh(Y)] + 2E\Big[E[(X - E[X|Y])^T|Y]Rh(Y)\Big] \qquad (6.7) \\
&= E[(X - E[X|Y])^T R(X - E[X|Y])] + E[h^T(Y)Rh(Y)] \\
&\geq E[(X - E[X|Y])^T R(X - E[X|Y])],
\end{aligned}$$

where in (6.6) we use Theorem 4.1.3 and in (6.7) we use Theorem 4.1.4. Note that without any loss we can assume that $E[h^T(Y)h(Y)] < \infty$ (by the above analysis for otherwise the expectation above would be unbounded) and therefore

$$E[|(X - E[X|Y])^T Rh(Y)|] \leq \Big(E\Big[(X - E[X|Y])^T R(X - E[X|Y])\Big]\Big)^{1/2} \Big(E[h^T(Y)Rh(Y)]\Big)^{1/2}$$

by the Cauchy-Schwarz inequality, so that $(X - E[X|Y])^T Rh(Y)$ is integrable, validating the use Theorem 4.1.3. Thus, for an optimal policy, we must have that $E[h^T(Y)Rh(Y)] = 0$. ⋄

Remark 6.1. We note that the above admits a Hilbert space interpretation or formulation: Let $\mathbb{H}$ denote the space of random variables (defined on a probability space) on which the inner product $\langle X, Z\rangle := E[X^T RZ]$ is defined; this defines a Hilbert space. Let M_H be a subspace of $\mathbb{H}$, defined as the closed subspace of random variables that are measurable on $\sigma(Y)$. Then, the *projection theorem* [220] leads to the observation that an optimal $g(Y) \in M_H$ minimizing $\|X - g(Y)\|_2^2$, denoted here by $G(Y)$, is one which satisfies:

$$\langle X - G(Y), h(Y)\rangle = E[X - G(Y)^T Rh(Y)] = 0, \qquad \forall h \in M_H$$

The conditional expectation satisfies this condition as

$$E[(X - E[X|Y])^T Rh(Y)] = E[E[(X - E[X|Y])^T Rh(Y)|Y]] = E[E[(X - E[X|Y])^T|Y]Rh(Y)] = 0,$$

since P a.s., $E[(X - E[X|Y])^T|Y] = 0$.

6.2.2 The Linear Quadratic Gaussian Problem

Consider the following linear system:

$$x_{t+1} = Ax_t + Bu_t + w_t,$$
$$y_t = Cx_t + v_t, \tag{6.8}$$

where $x \in \mathbb{R}^n, u \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$, $y \in \mathbb{R}^p$, $v \in \mathbb{R}^p$. Suppose $\{w_t, v_t\}$ are zero-mean i.i.d. random Gaussian vectors with given covariance matrices $E[w_t w_t^T] = W$ and $E[v_t v_t^T] = V$ for all $t \geq 0$.

The goal is to obtain

$$\inf_{\gamma} J(\gamma, \mu_0),$$

where

$$J(\mu_0, \gamma) = E^{\gamma}_{\mu_0}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t + x_N^T Q_N x_N], \tag{6.9}$$

with $R > 0$ and $Q, Q_N \geq 0$ (that is, these matrices are positive definite and positive semi-definite) and μ_0 is an initial prior probability measure (on x_0) assumed to be zero-mean Gaussian.

Building on Lemma 6.2.1, we will show in the following that the optimal control is linear in its expectation and has the form

$$u_t = -(B^T K_{t+1} B + R)^{-1} B^T K_{t+1} A E[x_t | I_t]$$

where K_t solves the Discrete-Time Riccati Equation:

$$K_t = Q + A^T K_{t+1} A - A^T K_{t+1} B (B^T K_{t+1} B + R)^{-1} B^T K_{t+1} A,$$

with final condition $K_N = Q_N$.

In the following, we start with the estimation problem.

6.2.3 Estimation and Kalman Filtering

In this section, we discuss the control-free setup and derive the celebrated *Kalman Filter*. In the following to make certain computations more explicit and easier to follow, we will use capital letters to denote the random variables and small letters for the realizations of these variables.

For a linear Gaussian system, the state process has a Gaussian probability measure. A Gaussian probability measure can be uniquely identified by knowing the mean and the covariance of the Gaussian random variable. This makes the analysis for estimating a Gaussian random variable particularly simple to perform, since the conditional estimate of a partially observed (through an additive Gaussian noise) Gaussian random variable is a linear/affine function of the observed variable.

Recall that a Gaussian measure with mean μ and covariance matrix Σ_{XX} has the following density:

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_{XX}|^{1/2}} e^{-1/2((x-\mu)^T \Sigma_{XX}^{-1}(x-\mu))}$$

Lemma 6.2.2 *Let X, Y be zero-mean Gaussian vectors. Then $E[X|Y = y]$ is linear in y: With $\Sigma_{XY} = E[XY^T]$ and $\Sigma_{YY} = E[YY^T]$,*

$$E[X|Y = y] = \Sigma_{XY} \Sigma_{YY}^{-1} y, \tag{6.10}$$

and

$$E[(X - E[X|Y])(X - E[X|Y])^T] = \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^T =: D.$$

In particular,

$$E[(X - E[X|Y])(X - E[X|Y])^T | Y = y]$$

does not depend on the realization y of Y and is equal to D.

We note that if the random variables are not-zero mean, one needs to add a constant correction term making the estimate an affine function (of the measurement).

Proof. By Bayes' rule and the fact that the processes admit densities: $p(x|y) = \frac{p(x,y)}{p(y)}$.With $K_{XY} := E\left[\begin{bmatrix} X \\ Y \end{bmatrix}[X^T Y^T]\right]$, we have that

$$K_{XY} = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

It follows that K_{XY}^{-1} is also symmetric (since the eigenvectors are the same as those of K_{XY} and the eigenvalues are inverted) and given by:

$$K_{XY}^{-1} = \begin{bmatrix} \Psi_{XX} & \Psi_{XY} \\ \Psi_{YX} & \Psi_{YY} \end{bmatrix},$$

Thus, for some normalization constant C,

$$\frac{p(x,y)}{p(y)} = C\frac{e^{-1/2(x^T\Psi_{XX}x + 2x^T\Psi_{XY}y + y^T\Psi_{YY}y)}}{e^{-1/2(y^T K_{YY}^{-1}y)}}$$

By the *completion of the squares method* for the expression in the exponent, for some matrix D we obtain

$$(x^T\Psi_{XX}x + 2x^T\Psi_{XY}y + y^T\Psi_{YY}y - y^T K_{YY}^{-1}y) = (x - Hy)^T D^{-1}(x - Hy) + Q(y),$$

it follows that $H = -\Psi_{XX}^{-1}\Psi_{XY}$ and $D = \Psi_{XX}^{-1}$. Since $K_{XY}^{-1}K_{XY} = I$ (and thus $\Psi_{XX}\Sigma_{XY} + \Psi_{XY}\Sigma_{YY} = 0$), H is also equal to $\Sigma_{XY}\Sigma_{YY}^{-1}$. Here $Q(y)$ is a quadratic expression in y. As a result, one obtains

$$p(x|y) = Ce^{-\frac{1}{2}Q(y)}e^{-1/2(x-Hy)^T D^{-1}(x-Hy)}.$$

Since $\int p(x|y)dx = 1$ (as it is a conditional probability density function), it follows that $Ce^{-\frac{1}{2}Q(y)} = \frac{1}{(2\pi)^{\frac{n}{2}}|D|^{1/2}}$ and is in fact independent of y. Then, we finally have that D, which does not depend on y, equals (see Lemma 6.2.4 below)

$$E[(X - E[X|Y])(X - E[X|Y])^T] = E[XX^T] - E[(E[X|Y])(E[X|Y])^T] = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^T \quad (6.11)$$

◇

Remark 6.2. The fact that $Q(y)$ above does not depend on y reveals an interesting result that the conditional covariance of $X - E[X|Y]$ viewed as a Gaussian random variable is identical for all y values. This is a crucial fact that will be utilized in the derivation of the Kalman Filter.

Remark 6.3. Even if the random variables X, Y are not Gaussian (but zero-mean), through another Hilbert space formulation and an application of the *Projection Theorem* (see Remark 6.1), it can be shown that the expression $\Sigma_{XY}\Sigma_{YY}^{-1}y$ is the best linear estimate, that is the solution to $\inf_K E[(X - KY)^T(X - KY)]$. One can naturally generalize this for random variables with non-zero mean.

We will derive the Kalman filter in the following. The following two lemmas are instrumental.

Lemma 6.2.3 *If $E[X] = 0$ and Z_1, Z_2 are orthogonal zero-mean Gaussian processes (with $E[Z_1 Z_2^T] = 0$), then $E[X|Z_1 = z_1, Z_2 = z_2] = E[X|Z = z_1] + E[X|Z_2 = z_2]$.*

Proof. The proof follows by writing $z = [z_1, z_2]^T$, noting that Σ_{ZZ} is diagonal and $E[X|z] = \Sigma_{XZ}\Sigma_{ZZ}^{-1}z$. ◇

Lemma 6.2.4 *$E[(X - E[X|Y])(X - E[X|Y])^T]$ is given by $D = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^T$ above.*

Proof. Note that

$$E[X(E[X|Y])^T] = E[(X - E[X|Y] + E[X|Y])(E[X|Y])^T] = E[E[X|Y](E[X|Y])^T]$$

since $X - E[X|Y]$ is orthogonal to $E[X|Y]$[1]. As a result,

$$\begin{aligned}&E[(X - E[X|Y])(X - E[X|Y])^T] = E[XX^T] - 2E[X(E[X|Y])^T] + E[E[X|Y](E[X|Y])^T] \\ &= E[XX^T] - E[E[X|Y](E[X|Y])^T],\end{aligned}$$

and the result follows from Lemma 6.2.2 in view of (6.10). ◇

Now, we can move on to the derivation of the Kalman Filter.

Consider

$$x_{t+1} = Ax_t + w_t, \qquad y_t = Cx_t + v_t,$$

with $E[w_t w_t^T] = W$ and $E[v_t v_t^T] = V$ where $\{w_t\}$ and $\{v_t\}$ are mutually independent i.i.d. zero-mean Gaussian processes.

Define

$$m_t = E[x_t|y_{[0,t-1]}]$$

$$\Sigma_{t|t-1} = E[(x_t - E[x_t|y_{[0,t-1]}])(x_t - E[x_t|y_{[0,t-1]}])^T|y_{[0,t-1]}]$$

and note that since the estimation error covariance does not depend on the realization $y_{[0,t-1]}$ (see Remark 6.2), we write also

$$\Sigma_{t|t-1} = E[(x_t - E[x_t|y_{[0,t-1]}])(x_t - E[x_t|y_{[0,t-1]}])^T]$$

Theorem 6.2.1 *The following holds:*

$$m_{t+1} = Am_t + A\Sigma_{t|t-1}C^T(C\Sigma_{t|t-1}C^T + V)^{-1}(y_t - Cm_t) \tag{6.12}$$

$$\Sigma_{t+1|t} = A\Sigma_{t|t-1}A^T + W - (A\Sigma_{t|t-1}C^T)(C\Sigma_{t|t-1}C^T + V)^{-1}(C\Sigma_{t|t-1}A^T) \tag{6.13}$$

with

$$m_0 = E[x_0]$$

and

$$\Sigma_{0|-1} = E[x_0 x_0^T]$$

Proof. With $x_{t+1} = Ax_t + w_t$, the following hold:

$$\begin{aligned}m_{t+1} &= E[Ax_t + w_t|y_{[0,t]}] = E[Ax_t|y_{[0,t]}] = E[Am_t + A(x_t - m_t)|y_{[0,t]}] \\ &= Am_t + E[A(x_t - m_t)|y_{[0,t-1]}, y_t - E[y_t|y_{[0,t-1]}]] \\ &= Am_t + E[A(x_t - m_t)|y_{[0,t-1]}] + E[A(x_t - m_t)|y_t - E[y_t|y_{[0,t-1]}]] \qquad (6.14) \\ &= Am_t + E[A(x_t - m_t)|y_t - E[y_t|y_{[0,t-1]}]] \\ &= Am_t + E[A(x_t - m_t)|Cx_t + v_t - E[Cx_t + v_t|y_{[0,t-1]}]] \\ &= Am_t + E[A(x_t - m_t)|C(x_t - m_t) + v_t] \qquad (6.15)\end{aligned}$$

In the above, (6.14) follows from Lemma 6.2.3. We also use the fact that w_t is independent of (and hence orthogonal to) $y_{[0,t]}$. Let $X = A(x_t - m_t)$ and $Y = y_t - E[y_t|y_{[0,t-1]}] = y_t - Cm_t = C(x_t - m_t) + v_t$. Then, by Lemma 6.2.2, $E[X|Y] = \Sigma_{XY}\Sigma_{YY}^{-1}Y$ and thus,

[1]Note that by iterated expectations, we have $E[(X-E[X|Y])(E[X|Y])^T] = E\Big[E[(X-E[X|Y])(E[X|Y])^T|Y]\Big] = E\Big[E[(X-E[X|Y])|Y](E[X|Y])^T\Big] = 0$.

$$m_{t+1} = Am_t + A\Sigma_{t|t-1}C^T(C\Sigma_{t|t-1}C^T + V)^{-1}(y_t - Cm_t)$$

Likewise,

$$x_{t+1} - m_{t+1} = A(x_t - m_t) + w_t - A\Sigma_{t|t-1}C^T(C\Sigma_{t|t-1}C^T + V)^{-1}(y_t - Cm_t),$$

leads to, after a few lines of calculations:

$$\Sigma_{t+1|t} = A\Sigma_{t|t-1}A^T + W - (A\Sigma_{t|t-1}C^T)(C\Sigma_{t|t-1}C^T + V)^{-1}(C\Sigma_{t|t-1}A^T)$$

$\diamond$

The above is the celebrated Kalman filter.

Define now

$$\tilde{m}_t := E[x_t|y_{[0,t]}] = m_t + E[x_t - m_t|y_{[0,t]}]$$

Following the analysis above, we obtain

$$\tilde{m}_t = m_t + E[x_t - m_t|y_{[0,t-1]}] + E\Big[x_t - m_t|y_t - E[y_t|y_{[0,t-1]}]\Big].$$

Note that we also have $m_t = A\tilde{m}_{t-1}$. The following result then follows:

Theorem 6.2.2 *The recursions for $\tilde{m}_t$ satisfy*

$$\tilde{m}_t = A\tilde{m}_{t-1} + \Sigma_{t|t-1}C^T(C\Sigma_{t|t-1}C^T + V)^{-1}(y_t - CA\tilde{m}_{t-1}), \tag{6.16}$$

with $\tilde{m}_0 = E[x_0|y_0]$.

We observe that the zero-mean variable $x_t - \tilde{m}_t$ is orthogonal to $y_{[0,t]}$, in the sense that the error is independent of the information available at the controller, and since the information available is Gaussian, independence and orthogonality are equivalent.

We observe that the recursion (6.13) in Theorem 6.2.1 is essentially identical to the recursions in Theorem 5.3.2 with writing $A = A^T, W = Q, V = R, C^T = B$. This leads to the following result (as a corollary of Theorem 6.2.1).

Theorem 6.2.3 *Suppose (A^T, C^T) is controllable (this is equivalent to saying that (A, C) is observable) and $V > 0$. Then, the recursions for the covariance matrices Σ_t in Theorem 6.2.1 admit a fixed point. If, in addition, with $W = BB^T$, (A^T, B^T) is observable (that is (A, B) is controllable), the fixed point solution is unique, and is positive definite. As noted earlier, these can be relaxed to stabilizability (of (A, B)) and detectability of (A, C) but in this case the fixed point solution may only be positive semi-definite.*

Remark 6.4. The above suggest that if the observations are sufficiently informative, then the Kalman filter converges to a solution (with an appropriate initialization), even in the absence of an irreducibility condition (i.e., the controllability condition for (A, B) above) on the original state process x_t; under irreducibility, however, the solution is unique. This intuition has been shown to find a precise generalization in the non-linear filtering context [85, 227, 314], see Definition 6.4.10.

6.2.4 Optimal Control of Partially Observed LQG Systems

Let us revisit (6.9). With the analysis of optimal linear estimation above, we will now reformulate the quadratic optimization problem (6.9) in terms of $\tilde{m}_t$, u_t and $x_t - \tilde{m}_t$ as follows. First, let us note the following:

Theorem 6.2.4 *Consider the controlled linear system (6.8). Then, with*

$$m_t = E[x_t|y_{[0,t-1]}, u_{[0,t-1]}]$$

and

$$\Sigma_{t|t-1} = E[(x_t - E[x_t|y_{[0,t-1]}, u_{[0,t-1]})(x_t - E[x_t|y_{[0,t-1]}, u_{[0,t-1]}])^T],$$

the following hold:

$$m_{t+1} = Am_t + Bu_t + A\Sigma_{t|t-1}C^T(C\Sigma_{t|t-1}C^T + V)^{-1}(y_t - Cm_t)$$
$$\Sigma_{t+1|t} = A\Sigma_{t|t-1}A^T + W - (A\Sigma_{t|t-1}C^T)(C\Sigma_{t|t-1}C^T + V)^{-1}(C\Sigma_{t|t-1}A^T)$$

with

$$m_0 = E[x_0]$$

and

$$\Sigma_{0|-1} = E[x_0 x_0^T]$$

The proof follows that of Theorem 6.2.1: the only difference is the presence of control. Observe that, the estimation can be viewed to be that of estimating:

$$x_n = \left(A^n x_0 + \sum_{k=0}^{n-1} A^{n-k-1} w_k\right) + \sum_{k=0}^{n-1} A^{n-k-1} B u_k =: \bar{x}_n + \sum_{k=0}^{n-1} A^{n-k-1} B u_k$$

where

$$\bar{x}_{n+1} = A\bar{x}_n + w_n$$

is the control-free system. But since $\sum_{k=0}^{n-1} A^{n-k-1} B u_k$ is known at time n (by the controller), the estimation problem is essentially that of estimating the control-free system $\bar{x}_n$. Furthermore, the control adds no additional information with regard to estimating $\bar{x}_n$, that is, the information generated by

$$\bar{y}_n = C\bar{x}_n + v_n$$

up to time n contains the same information with regard to $\bar{x}_n$ as that contained by $\{y_k, u_k\}$ up to time n, because (i) $\bar{x}_n$ is not affected by the control, and (ii) the information that control actions contain are already available in the information content of the current and past $\bar{y}_n$ variables under any measurable policy (to see this, note that u_0 is a function of $\bar{y}_0$, and u_1 is a function of u_0 and $\bar{y}_{[0,1]}$, and thus really only that of $\bar{y}_{[0,1]}$, and so on for $n > 1$ by an inductive reasoning). That is, under any policy γ, for any Borel B and any n:

$$P^\gamma(\bar{x}_n \in B|\bar{y}_{[0,n]}) = P^\gamma(\bar{x}_n \in B|\bar{y}_{[0,n]}, u_{[0,n-1]})$$

What the above implies is that, under any policy γ

$$E^\gamma[x_n|y_{[0,n]}, u_{[0,n-1]}] = \sum_{k=0}^{n-1} A^{n-k-1} B u_k + E[\bar{x}_n|y_{[0,n]}, u_{[0,n-1]}] = \sum_{k=0}^{n-1} A^{n-k-1} B u_k + E[\bar{x}_n|\bar{y}_{[0,n]}].$$

Furthermore, $x_n - E[x_n|y_{[0,n-1]}, u_{[0,n-1]}]$ is sample path equivalent to $\bar{x}_n - E[\bar{x}_n|\bar{y}_{[0,n-1]}]$, and these are determined solely by $x_0, w_{[0,n-1]}, v_{[0,n-1]}$.

Now, for the controlled case, let us define

$$\tilde{m}_t = E[x_t|y_{[0,t]}, u_{[0,t-1]}].$$

and observe that the Kalman filtering recursions apply almost verbatim with the control actions added in an additive fashion:

$$\tilde{m}_t = A\tilde{m}_{t-1} + Bu_{t-1} + \Sigma_{t|t-1}C^T(C\Sigma_{t|t-1}C^T + V)^{-1}\Big(y_t - C(A\tilde{m}_{t-1} + Bu_{t-1})\Big) \tag{6.17}$$

Let $I_t = \{y_{[0,t]}, u_{[0,t-1]}\}$. Observe now that

$$
\begin{aligned}
&E[x_t^T Q x_t] = E[(x_t - \tilde{m}_t + \tilde{m}_t)^T Q (x_t - \tilde{m}_t + \tilde{m}_t)] \\
&= E[(x_t - \tilde{m}_t)^T Q (x_t - \tilde{m}_t)] + E[\tilde{m}_t^T Q \tilde{m}_t] + 2E[(x_t - \tilde{m}_t)^T Q \tilde{m}_t] \\
&= E[(x_t - \tilde{m}_t)^T Q (x_t - \tilde{m}_t)] + E[\tilde{m}_t^T Q \tilde{m}_t] + 2E[E[(x_t - \tilde{m}_t)^T Q \tilde{m}_t | I_t]] \\
&= E[(x_t - \tilde{m}_t)^T Q (x_t - \tilde{m}_t)] + E[\tilde{m}_t^T Q \tilde{m}_t] \qquad (6.18)
\end{aligned}
$$

by the orthogonality property of the conditional estimation error to $\tilde{m}_t$ (recall that $\tilde{m}_t$ is a function of I_t and $E[(x_t - \tilde{m}_t)^T Q \tilde{m}_t | I_t] = 0$). In particular, the cost:

$$
J(\gamma, \mu_0) = E^{\gamma}_{\mu_0}[\sum_{t=0}^{N-1} x_t^T Q x_t + u_t^T R u_t + x_N^T Q_N x_N], \qquad (6.19)
$$

writes as:

$$
\begin{aligned}
E^{\gamma}_{\mu_0}[\sum_{t=0}^{N-1} \tilde{m}_t^T Q \tilde{m}_t + u_t^T R u_t + \tilde{m}_N^T Q_N \tilde{m}_N] + E^{\gamma}_{\mu_0}[\sum_{t=0}^{N-1} (x_t - \tilde{m}_t)^T Q (x_t - \tilde{m}_t)] \\
+ E^{\gamma}_{\mu_0}[(x_N - \tilde{m}_N)^T Q_N (x_N - \tilde{m}_N)] \qquad (6.20)
\end{aligned}
$$

for the fully observed system (see (6.17)):

$$
\tilde{m}_t = A \tilde{m}_{t-1} + B u_{t-1} + \tilde{w}_{t-1}, \qquad (6.21)
$$

with

$$
\tilde{w}_{t-1} = \Sigma_{t|t-1} C^T (C \Sigma_{t|t-1} C^T + V)^{-1} \Big(y_t - C(A \tilde{m}_{t-1} + B u_{t-1}) \Big)
$$

Furthermore, the estimation errors in (6.20) (the second and the third terms) **do not depend on the control policy** γ so that the expected cost writes as

$$
\begin{aligned}
E^{\gamma}_{\mu_0}[\sum_{t=0}^{N-1} \tilde{m}_t^T Q \tilde{m}_t + u_t^T R u_t + \tilde{m}_N^T Q_N \tilde{m}_N] + E_{\mu_0}[\sum_{t=0}^{N-1} (x_t - \tilde{m}_t)^T Q (x_t - \tilde{m}_t)] \\
+ E_{\mu_0}[(x_N - \tilde{m}_N)^T Q_N (x_N - \tilde{m}_N)] \qquad (6.22)
\end{aligned}
$$

Thus, the optimal control problem is equivalent to the control of the fully observed state $\tilde{m}_t$, with additive time-varying independent Gaussian noise process $\{\tilde{w}_t\}$ given in (6.21).

Here, that the error term $(x_t - \tilde{m}_t)$ does not depend on the control policy is a consequence of what is known as the *lack of dual effect of control*: the control actions up to any time t do not affect the state estimation error process for the future time stages. Using our earlier analysis, it follows then that the optimal control has the form stated in the following:

Theorem 6.2.5 *Consider (6.8) with cost criterion given in (6.9). The optimal control is given with*

$$
u_t = -(B^T P_{t+1} B + R)^{-1} B^T P_{t+1} A E[x_t | I_t] = -(B^T P_{t+1} B + R)^{-1} B^T P_{t+1} A \tilde{m}_t,
$$

with $\tilde{m}_t$ computed as in (6.21), and P_t generated as in Theorem 5.3.1 with $P_N = Q_N$. The optimal cost writes as

$$
E[\tilde{m}_0^T P_0 \tilde{m}_0] + E\Big[\sum_{k=0}^{N-1} \bar{w}_k^T P_{k+1} \bar{w}_k + (x_k - \tilde{m}_k)^T Q (x_k - \tilde{m}_k)\Big] + E[(x_N - \tilde{m}_N)^T Q_N (x_N - \tilde{m}_N)]
$$

In the above problem, we observed that the optimal control has a separation structure: The controller first estimates the state, and then applies its control action, by regarding the estimate as the state itself.

Separation of Estimation and Control. In the above we observe that the optimal control policy is the same as that in the fully observed setup in Theorem 5.3.1, except that the state is replaced with its estimate. The sufficiency of

conditional expectation in optimal control is generally known as the *separation of estimation and control* [141] [206] [334] [219]–that is, the separation principle is said to hold when an optimal control exists in a subset of admissible policies where the control depends on the information only through the conditional expectation of the state given the information available–, and for this particular case, a more special version of it, known as the *certainty equivalence* principle, applies: As expressed in [29, eqs. (2.20)–(2.22)], a control problem possesses the **certainty equivalence** (CE) property if the closed-loop optimal control policy has the same form as the deterministic optimal control policy under perfect state observation and in the absence of process noise. More precisely, if in the absence of noise the optimal control policy for the deterministic system is

$$u_k = \phi_k(x_k), \tag{6.23}$$

and CE holds, then the optimal closed-loop control policy for the noisy and not necessarily fully observed system is

$$u_k^{\text{CE}} = \phi_k(E[x_k|y_{[0,k]}, u_{[0,k-1]}]), \forall k. \tag{6.24}$$

As observed above, the absence of *dual effect* plays a key part in this analysis leading the *separation of estimation and control principle*, in taking $E[(x_t - \tilde{m})^T Q(x_t - \tilde{m})]$ out of the optimization over control policies, since it does not depend on the policy.

Remark 6.5 (Dual Effect vs. Certainty Equivalence). In [30], dual effect is introduced as the property that the moments of $(x_t - \tilde{m}_t)$ do not depend on the past applied control actions (leading to a form of probabilistic independence). A more general condition would be that $(x_t - \tilde{m}_t)$ does not depend on the past control *policies* (and not necessarily the actions) in that the control policies do not alter the realization of the random variable $(x_t - \tilde{m}_t)$ (see [103] for an explicit analysis and relaxations). This distinction is important in certain applications in networked control systems where separation results are particularly important [345] (as probabilistic independence is often too restrictive when one goes beyond the Gaussian setup). We also note that separation also applies in the linear quadratic setup when the noise processes are not Gaussian, though of course the conditional estimations will no longer be linear [35, Lemma 5.2.1]. For results involving non-linear measurement models and for a detailed literature review, the reader is referred to [103].

In many problems, the dual effect of the control is present and, depending on the control policy, the estimation quality at the controller regarding future states may be affected. As an example, consider a linear system controlled over an erasure channel, where the controller applies a control, but does not know whether the control reaches the destination or not. In this case, the control signal which was intended to be sent, does affect the estimation error [176, 286].

6.3 On the Controlled Markov Construction in the Space of Probability Measures and Extension to General Spaces

In Section 6.1, we observed that we can replace the state with a probability measure valued state. It is important to provide notions of convergence and continuity on the spaces of probability measures to be able to apply the machinery of *Chapter 5*. In view of Theorem 6.1.1, if we can invoke the measurable selection conditions studied earlier (such as Assumption 5.2.1), we can use the machinery of optimal stochastic control (such as Bellman's principle) for partially observed models.

The reader is referred to Appendix D for review of some concepts involving convergences of probability measures.

6.3.1 Non-linear Filter in the Standard Borel setup

The analysis in Section 6.1 applies essentially identically to the standard Borel setup.

We consider (6.1). Let $\mathbb{X}$ be a standard Borel set from which the controlled Markov process $\{x_t,\, t \in \mathbb{Z}_+\}$ takes its values with transition kernel $\mathcal{T}$. Let $\mathbb{Y}$ be a standard Borel space, and let the observation channel Q be defined as the stochastic kernel (regular conditional probability) from $\mathbb{X} \times \mathbb{U}$ to $\mathbb{Y}$ such that $Q(\,\cdot\,|x,u)$ is a probability measure on the Borel σ-algebra $\mathcal{B}(\mathbb{Y})$ of $\mathbb{Y}$ for every $(x,u) \in \mathbb{X} \times \mathbb{U}$ and $Q(A|\,\cdot\,) : \mathbb{X} \times \mathbb{U} \to [0,1]$ is a Borel measurable function for every $A \in \mathcal{B}(\mathbb{Y})$.

Let a decision maker (DM) be located at the output of an observation channel Q, with inputs x_t and outputs y_t. An *admissible policy* γ is a sequence of control functions $\{\gamma_t,\, t \in\}$ such that γ_t is measurable with respect to the σ-algebra generated by the information variables

$$I_t = \{y_{[0,t]}, u_{[0,t-1]}\}, \quad t \in \mathbb{N}, \qquad I_0 = \{y_0\},$$

where

$$u_t = \gamma_t(I_t), \quad t \in \tag{6.25}$$

are the $\mathbb{U}$-valued control action variables. We define $\mathit{\Gamma}_A$ to be the set of all such admissible policies. The joint distribution of the state, control, and observation processes is determined by (12.1) and the following system dynamics:

$$P\big((x_0,y_0) \in B\big) = \int_B Q(dy_0|x_0)P_0(dx_0), B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}),$$

where P_0 is the prior distribution of the initial state x_0 and Q_0 is the observation channel, and for $t \in \mathbb{N}$

$$\begin{aligned} &P\bigg((x_t,y_t) \in B \,\bigg|\, (x,y,u)_{[0,t-1]} = (x,y,u)_{[0,t-1]}\bigg) \\ &\quad = \int_B Q(dy_t|x_t)\mathcal{T}(dx_t|x_{t-1},u_{t-1}), B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}), \end{aligned}$$

where $\mathcal{T}(\cdot|x,u)$ is a stochastic kernel from $\mathbb{X} \times \mathbb{U}$ to $\mathbb{X}$. This completes the probabilistic description of the partially observed model. Let a one-stage cost function $c : \mathbb{X} \times \mathbb{U} \to [0,\infty)$, which is a Borel measurable function from $\mathbb{X} \times \mathbb{U}$ to $[0,\infty)$, be given. Then, we denote by $J(\gamma)$ the cost function of the policy $\gamma \in \mathit{\Gamma}_A$, which can be, for instance, finite horizon, discounted cost or average cost criteria. The goal of the control problem is to find an optimal policy γ^* that minimizes J.

As studied earlier, any such problem can be reduced to a completely observable Markov process [352], [262], whose states are the posterior state distributions or 'beliefs' of the observer; that is, the state at time t is

$$\pi_t(\,\cdot\,) := P\{X_t \in \,\cdot\,|y_0,\ldots,y_t,u_0,\ldots,u_{t-1}\} \in \mathcal{P}(\mathbb{X}).$$

We call this equivalent process the filter process . The filter process has state space $\mathcal{P}(\mathbb{X})$ and action space $\mathbb{U}$. Recall again that $\mathcal{P}(\mathbb{X})$ is equipped with the Borel σ-algebra generated by the topology of weak convergence, where, under this topology $\mathcal{P}(\mathbb{X})$ is also a standard Borel space.

The transition probability of the filter process can be constructed as follows. As in the countable setup case, we have the following explicit Bayesian recursion to define F under a mild regularity condition: Let Q be *dominated* in the sense that there exists a dominating reference measure λ such that $\forall x \in \mathbb{X}, Q(dy|x_n = x) \ll \lambda$. Then, define the Radon-Nikodym derivative

$$g(x,y) = \frac{dG(y_n \in \cdot|x_n = x)}{d\lambda}(y)$$

as the likelihood function (serving as a conditional probability density function) and we can write the filter π_{n+1} recursively in terms of π_n and y_{n+1}, u_n explicitly as a Bayesian update:

$$\pi_{n+1}(dx_{n+1}) =: F(\pi_n, y_{n+1}, u_n)(dx_{n+1}) = \frac{\int_{\mathbb{X}} g(x_{n+1},y_{n+1})\mathcal{T}(dx_{n+1}|x_n,u_n)\pi_n(dx_n)}{\int_{\mathbb{X}}\int_{\mathbb{X}} g(x_{n+1},y_{n+1})\mathcal{T}(dx_{n+1}|x_n,u_n)\pi_n(dx_n)} \tag{6.26}$$

As earlier in the countable space setup, the transition probability η of the filter process is constructed as follows. If we define the measurable function $F(\pi, y, u) := P\{x_{t+1} \in \cdot\,|\pi_t = \pi, u_t = u, y_{t+1} = y\}$ from $\mathcal{P}(\mathbb{X}) \times \mathbb{U} \times \mathbb{Y}$ to $\mathcal{P}(\mathbb{X})$ and use the stochastic kernel $P(\cdot\,|\pi, u) = P\{y_{t+1} \in \cdot\,|\pi_t = \pi, u_t = u\}$ from $\mathcal{P}(\mathbb{X}) \times \mathbb{U}$ to $\mathbb{Y}$, we can write η as

$$\eta(\cdot\,|\pi, u) = \int_{\mathbb{Y}} 1_{\{F(\pi,y,u)\in\cdot\}} P(dy|\pi, u). \tag{6.27}$$

In (6.2), we need to show that the expression $P(\pi_{t+1} \in D|\pi_t, u_t)$ is a regular conditional probability measure; that is, for every fixed $D \in \mathcal{B}(\mathcal{P}(\mathbb{X}))$, this is a measurable function on $\mathcal{P}(\mathbb{X}) \times \mathbb{U}$ and for every π_t, u_t, it is a conditional probability measure on $\mathcal{P}(\mathbb{X})$. Furthermore, we need to ensure that $\tilde{c}$, the equivalent cost function, is also a measurable function.

A proof of the first result below can be found in [4] (see Theorem 15.13 in [4] or p. 215 in [53])

Theorem 6.3.1 *Let $\mathbb{S}$ be a Polish space and M be the set of all measurable and bounded functions $f : \mathbb{S} \to \mathbb{R}$. Then, for any $f \in M$, the integral*

$$\int \pi(dx) f(x)$$

defines a measurable function on $\mathcal{P}(\mathbb{S})$ under the topology of weak convergence.

This is a useful result since it allows us to view integral forms as measurable functions on the space of probability measures when we work with the topology of weak convergence. The second useful result follows from Theorem 6.3.1 and Theorem 2.1 of Dubins and Freedman [110] and Proposition 7.25 in Bertsekas and Shreve [37].

Theorem 6.3.2 *Let $\mathbb{S}$ be a Polish space. A function $F : \mathcal{P}(\mathbb{S}) \to \mathcal{P}(\mathbb{S})$ is measurable on $\mathcal{B}(\mathcal{P}(\mathbb{S}))$ (under weak convergence), if for all $B \in \mathcal{B}(\mathbb{S})$ $(F(\cdot))(B) : \mathcal{P}(\mathbb{S}) \to \mathbb{R}$ is measurable under weak convergence on $\mathcal{P}(\mathbb{S})$, that is for every $B \in \mathcal{B}(\mathbb{S})$, $(F(\pi))(B)$ is a measurable function when viewed as a function from $\mathcal{P}(\mathbb{S})$ to $\mathbb{R}$.*

By Theorem 6.3.2, we have that F is a Borel measurable function, and η is a stochastic kernel.

The above thus establish that under weak convergence topology, (π_t, u_t) forms a standard Borel controlled Markov chain.

As in the countable setup, in the belief-MDP formulation, the one-stage cost function $\tilde{c} : \mathcal{P}(\mathbb{X}) \times \mathbb{U} \to [0, \infty)$ for the filter process is given by

$$\tilde{c}(\pi, u) := \int_{\mathbb{X}} c(x, u)\pi(dx),$$

With cost function $c(x, u)$ continuous and bounded on $\mathbb{X}) \times \mathbb{U}$, by an application of the generalized dominated convergence theorem (see Theorem D.3.1 [212, Theorem 3.5] [287, Theorem 3.5]), we have that that $\tilde{c}(\pi, u) = E^{\pi}[c(x, u)] := \int \pi(dx)c(x, u) : \mathcal{P}(\mathbb{X}) \times \mathbb{U} \to \mathbb{R}$ is also continuous and bounded, and thus Borel measurable as a map from $\mathcal{P}(\mathbb{X}) \times \mathbb{U}$ to $\mathbb{R}$.

Hence, the filter process defines a completely observable Markov process with the components $(\mathcal{P}(\mathbb{X}), \mathbb{U}, \tilde{c}, \eta)$.

For the filter process, let us define another information variable sequence as

$$\tilde{I}_t = \{\pi_{[0,t]}, u_{[0,t-1]}\}, \quad t \in \mathbb{N}, \qquad \tilde{I}_0 = \{\pi_0\}.$$

Now, building all these together, as in the countable setup, in view of the results in *Chapter 5* (notably Theorem **??**5.1.1Theorem 5.1.1hen follows that an optimal control policy of the original POMDP will use the belief π_t as a sufficient statistic for optimal policies (see [352], [262]). More precisely, the filter process is equivalent to the original POMDP in the sense that for any optimal policy using the filter process, one can construct a policy for the original POMDP which is optimal, or more generally, for any policy which uses $\tilde{I}_t$ there exists another one which only uses the filter process π_t and which is at least as good as the original policy.

6.3.2 Continuity Properties of Belief-MDP: Weak Continuity and Wasserstein Continuity of Filter Kernels

Weak Feller Continuity of the Filter Kernel η. Building on [93], [184] and [128], we first study the weak Feller property of the filter process; that is, the weak Feller property of the kernel defined in (6.3) under two different sets of assumptions.

Assumption 6.3.1 *(i) The transition probability $\mathcal{T}(\cdot|x,u)$ is weakly continuous in (x,u), i.e., for any $(x_n,u_n) \to (x,u)$, $\mathcal{T}(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ weakly.*

(ii) The observation channel $Q(\cdot|x,u)$ is continuous in total variation, i.e., for any $(x_n,u_n) \to (x,u)$, $Q(\cdot|x_n,u_n) \to Q(\cdot|x,u)$ in total variation.

Assumption 6.3.2 *(i) The transition probability $\mathcal{T}(\cdot|x,u)$ is continuous in total variation in (x,u), i.e., for any $(x_n,u_n) \to (x,u)$, $\mathcal{T}(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ in total variation.*

(ii) The observation channel $Q(\cdot|x)$ is independent of the control variable.

Theorem 6.3.3 *[128] Under Assumption 6.3.1, the transition probability $\eta(\cdot|z,u)$, given in (6.27), of the filter process is weakly continuous in (z,u).*

See also [93] for an earlier though slightly more restrictive result along the above.

Theorem 6.3.4 *[184] Under Assumption 6.3.2, the transition probability $\eta(\cdot|z,u)$, given in (6.27), of the filter process is weakly continuous in (z,u).*

A proof for these is given in Section 6.6.

If the cost function c is continuous and bounded, an application of the dominated convergence theorem implies that $\tilde{c}(\pi,u)$ is also continuous and bounded. If the action set is compact, then under the weak continuity condition noted above on the non-linear filter, we have that the measurable selection conditions apply, and solutions to the Bellman or discounted cost optimality equations exist, and accordingly an optimal control policy exists.

See [188, Theorem 7], which builds on [184], for further refinements with explicit moduli of continuity for the weak Feller property.

We refer the reader also to [128, Theorem 7.1] which establishes weak Feller property under further sets of assumptions. See [1, 127, 129, 188] for further results on the above weak Feller property.

As examples, taken from [184], suppose that the system dynamics and the observation channel are represented as follows:

$$\begin{aligned} x_{t+1} &= H(x_t, u_t, w_t), \\ y_t &= G(x_t, u_{t-1}, v_t), \end{aligned}$$

where w_t and v_t are i.i.d. noise processes.

(i) Suppose that $H(x,u,w)$ is a continuous function in x and u. Then, the corresponding transition kernel is weakly continuous. To see this, observe that, for any $c \in C_b(\mathbb{X})$, we have

$$\begin{aligned} &\int c(x_1)\mathcal{T}(dx_1|x_0^n,u_0^n) = \int c(H(x_0^n,u_0^n,w_0))\mu(dw_0) \\ &\to \int c(H(x_0,u_0,w_0))\mu(dw_0) = \int c(x_1)\mathcal{T}(dx_1|x_0,u_0), \end{aligned}$$

where we use μ to denote the probability model of the noise.

(ii) Suppose that $G(x,u,v) = g(x,u)+v$, where g is a continuous function and V_t admits a continuous density function φ with respect to some reference measure ν. Then, the channel is continuous in total variation. Notice that under this setup, we can write $Q(dy|x,u) = \varphi(y - h(x,u))\nu(dy)$. Hence, the density of $Q(dy|x_n,u_n)$ converges to the density of $Q(dy|x,u)$ pointwise, and so, $Q(dy|x_n,u_n)$ converges to $Q(dy|x,u)$ in total variation by Scheffé's Lemma [43]. Hence, $Q(dy|x,u)$ is continuous in total variation under these conditions.

(iii)Suppose that we have $H(x,u,w) = h(x,u)+w$, where f is continuous and w_t admits a continuous density function φ with respect to some reference measure ν. Then, the transition probability is continuous in total variation: with this setup we have $\mathcal{T}(dx_1|x_0,u_0) = \varphi(x_1-h(x_0,u_0))\nu(dx_1)$. Thus, continuity of φ and h guarantees the pointwise convergence of the densities, so we can conclude that the transition probability is continuous in total variation by again Scheffé's Lemma.

Under the above, it follows that for the belief-MDP $(\mathcal{P}(\mathbb{X}),\mathbb{U},\tilde{c},\eta)$, Assumption 5.2.1 holds and therefore Theorem 5.2.1 applies: For a finite horizon cost minimization problem, there exists an optimal control policy which is of Markov type (Markov in the belief state, π_t). The discounted and average cost criteria will be presented in the following section.

Remark 6.6 (Existence results without separation / belief-MDP reduction). Consider a partially observable stochastic control problem (POMDP) with the following dynamics.

$$x_{t+1} = f(x_t,u_t,w_t), \qquad y_t = g(x_t,v_t).$$

If $f(\cdot,\cdot,w)$ is continuous and g has the form: $y_t = g(x_t)+v_t$, with g continuous and w_t admitting a continuous density function η, an existence result can be established building on the measurable selection criteria under weak continuity in view of Theorem 6.3.3.

Without adopting the belief-MDP reduction method, such an existence result can also be established by a measure transformation argument and using a strategic measures approach: With η denoting the density of v_n, we have $P(y_n \in B|x_n) = \int_B \eta(y - g(x_n))dy$. With η and g continuous and bounded, taking $y^n := y_n$, by writing $x_{n+1} = f(x_n,u_n,w_n) = f(f(x_{n-1},u_{n-1},w_{n-1}),u_n,w_n)$, and iterating inductively to obtain

$$x_{n+1} = h_n(x_0,\mathbf{u}_{[0,n-1]},\mathbf{w}_{[0,n-1]}),$$

for some h_n which is continuous in $\mathbf{u}_{[0,n-1]}$ for every fixed $x_0,\mathbf{w}_{[0,n-1]}$, one obtains an effective *reduced* cost (10.31) that is a continuous function in the control actions. [346, Section 5.4.2] then implies the existence of an optimal control policy. This reasoning is also applicable when the measurements are not additive in the noise but with $P(y_n \in B|x_n = x) = \int_B m(y,x)\eta(dy)$ for some m continuous in x and η a reference measure.

It may be important to note that Bismut [46] arrived at related results for partially observed models in continuous-time, through an approach which also avoids separation / the construction of a belief-MDP. Please see Section 10.8.1 for further discussion.

Wasserstein Continuity of the Filter Kernel η. Regularity under the Wasserstein metric has been studied in [1, 100] and [101]:

Assumption 6.3.3

1. *$(\mathbb{X},d)$ is a bounded compact metric space with diameter D (where $D = \sup_{x,y\in\mathbb{X}} d(x,y)$).*
2. *The transition probability $\mathcal{T}(\cdot \mid x,u)$ is continuous in total variation in (x,u), i.e., for any $(x_n,u_n) \to (x,u), \mathcal{T}(\cdot \mid x_n,u_n) \to \mathcal{T}(\cdot \mid x,u)$ in total variation.*
3. *There exists $\alpha \in R^+$ such that*
$$\|\mathcal{T}(\cdot \mid x,u) - \mathcal{T}(\cdot \mid x',u)\|_{TV} \le \alpha d(x,x')$$
for every $x,x' \in \mathbb{X}$, $u \in \mathbb{U}$.
4. *There exists $K_1 \in \mathbb{R}_+$ such that*
$$|c(x,u) - c(x',u)| \le K_1 d(x,x').$$

for every $x, x' \in \mathbb{X}$, $u \in \mathbb{U}$.

5. *The cost function* c *is bounded and continuous.*

Theorem 6.3.5 *[100] Assume that* $\mathbb{X}$ *and* $\mathbb{Y}$ *are Polish spaces. If Assumptions 6.3.3-1,3 are fulfilled, then we have*

$$W_1\left(\eta(\cdot \mid z_0, u), \eta\left(\cdot \mid z_0', u\right)\right) \leq K_2 W_1\left(z_0, z_0'\right),$$

with $K_2 := \frac{\alpha D(3-2\delta(Q))}{2}$ *for all* $z_0, z_0' \in \mathcal{P}(\mathbb{X})$, $u \in \mathbb{U}$.

Assumption 6.3.4 *(i)* $(\mathbb{X}, d)$ *is a compact metric space.*

(ii) There exists a constant $\theta \in (0,1)$ *such that*

$$W_1\left(\mathcal{T}(\cdot \mid x, u) - \mathcal{T}\left(\cdot \mid x', u\right)\right) \leq \theta \cdot d\left(x, x'\right)$$

for every $x, x' \in \mathbb{X}, u \in \mathbb{U}$.

(iii)There exists a constant $\gamma \in \mathbb{R}^+$ *such that*

$$\|Q(\cdot \mid x) - Q\left(\cdot \mid x'\right)\|_{TV} \leq \gamma \cdot d\left(x, x'\right)$$

for every $x, x' \in \mathbb{X}$.

Theorem 6.3.6 *[99, Theorem 2.4] Assume that* $\mathbb{X}$ *and* $\mathbb{Y}$ *are Polish spaces. Under Assumption 6.3.4, we have*

$$W_1\left(\eta\left(\cdot \mid z_0, u\right), \eta\left(\cdot \mid z_0', u\right)\right) \leq \left(\theta + \frac{3\theta\gamma D}{2}\right) W_1\left(z_0, z_0'\right)$$

for all $z_0, z_0' \in \mathcal{Z}, u \in \mathbb{U}$, *where* $D = \sup_{x,y \in \mathbb{X}} d(x,y)$.

Remark 6.7. [348] has presented an alternative approach, without belief-separation, and has arrived further conditions for the existence of optimal policies for discounted and average cost problems as well as the unique ergodicity property for both controlled and control-free setups. Such an approach leads to complementary conditions on the weak Feller property on the state, which considers the entire past as the state endowed with the product topology.

6.3.3 Existence of Optimal Policies: Discounted Cost and Average Cost

Consider the minimization of either the discounted cost criterion (for some $\beta \in (0,1)$

$$J(\mu, \gamma) := E_\mu^\gamma[\sum_{k=0}^{\infty} \beta^k c(x_k, u_k)] \tag{6.28}$$

or the average cost criterion

$$J(\mu, \gamma) := \limsup_{N \to \infty} \frac{1}{N} E_\mu^\gamma[\sum_{k=0}^{N-1} c(x_k, u_k)], \tag{6.29}$$

over all admissible control policies $\gamma = \{\gamma_0, \gamma_1, \cdots\} \in \mathit{\Gamma}$ with $x_0 \sim \mu$.

Discounted Cost Cost.

Theorem 6.3.7 *If the cost function* $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ *is continuous and bounded, and* $\mathbb{U}$ *is compact, under under Theorems 6.3.3 or 6.3.4, for any* $\beta \in (0,1)$, *there exists an optimal solution to the discounted cost optimality problem*

with a continuous and bounded value function. Furthermore, under Assumption 6.3.3, with $K_2 = \frac{\alpha D(3-2\delta(Q))}{2}$*, if* $\beta K_2 < 1$ *the value function is Lipschitz continuous.*

Proof. An application of the dominated convergence theorem implies that $\tilde{c}(\pi, u)$ is also continuous and bounded. If the action set is compact, then under Theorems 6.3.3 or 6.3.4, which imply that η is weakly continuous, we have that the measurable selection conditions (see e.g. [165]) apply, and solutions to the Bellman or discounted cost optimality equations exist, and accordingly an optimal control policy exists. For the second result, Theorem 5.5.3 (see [270, Theorem 4.37]) leads to Lipschitz regularity under the Wasserstein continuity condition on the kernel. ◇

Average Cost. The average cost is a significantly more challenging problem as the typical contraction conditions via minorization (7.2.2) for the kernel η is too demanding. As is studied in detail in *Chapter* 7, the average cost optimality equation (ACOE) plays a crucial role for the analysis and the existence results of MDPs under the infinite horizon average cost optimality criteria. The triplet (h, ρ^*, γ^*), where $h, \gamma : \mathcal{P}(\mathbb{X}) \to \mathcal{R}$ are measurable functions and $\rho* \in R$ is a constant forms the ACOE if

$$\begin{aligned} h(z) + \rho^* &= \inf_{u \in U} \left\{ \tilde{c}(z, u) + \int h(z_1) \eta(dz_1 | z, u) \right\} \\ &= \tilde{c}(z, \gamma^*(z)) + \int h(z_1) \eta(dz_1 | z, \gamma^*(z)) \end{aligned} \tag{6.30}$$

for all $z \in \mathcal{P}(\mathbb{X})$. It is well known that (see e.g. [165, Theorem 5.2.4]) if (6.30) is satisfied with the triplet (h, ρ^*, γ^*), and furthermore if h satisfies

$$\sup_{\gamma \in \Gamma} \lim_{t \to \infty} \frac{E_z^{\gamma}[h(Z_t)]}{t} = 0, \quad \forall z \in \mathcal{P}(\mathbb{X})$$

then γ^* is an optimal policy for the POMDP under the infinite horizon average cost optimality criteria, and

$$J^*(z) = \inf_{\gamma \in \Gamma} J(z, \gamma) = \rho^* \quad \forall z \in \mathcal{P}(\mathbb{X}).$$

Theorem 6.3.8 *[100] Under Assumption 6.3.3, with* $K_2 = \frac{\alpha D(3-2\delta(Q))}{2} < 1$*, a solution to the average cost optimality equation (ACOE) exists. This leads to the existence of an optimal control policy, and optimal cost is constant for every initial state.*

The proof follows from Corollary 7.3.1. For belief-MDPs, we should emphasize that minorization conditions (as in Assumption 7.2.2) are typically not applicable.

6.3.4 A useful structural result: Concavity of the value function in the priors

The following theorem establishes concavity of the optimal cost in a single-stage stochastic control problem over the space of initial distributions and this also applies for multi-stage setups.

Theorem 6.3.9 *Let* $\int c(x, \gamma(y)) PQ(dx, dy)$ *exist for all* $\gamma \in \Gamma$ *and* $P \in \mathcal{P}(\mathbb{X})$*. Then,*

$$J^*(P, Q) = \inf_{\gamma \in \Gamma} E_P^{Q, \gamma}[c(x, \gamma(y))]$$

is concave in P*.*

Proof. For $a \in [0, 1]$ and $P', P'' \in \mathcal{P}(\mathbb{X})$ we let $P = aP' + (1 - a)P''$. Note that $PQ = aP'Q + (1 - a)P''Q$. We have

$$\begin{aligned}
J(aP' + (1-a)P'', Q) &= J(P, Q) \\
&= \inf_{\gamma \in \Gamma} E_P^{Q,\gamma}[c(x, \gamma(y))] \\
&= \inf_{\gamma \in \Gamma} \int c(x, \gamma(y)) PQ(dx, dy) \\
&= \inf_{\gamma \in \Gamma} \Bigg(a \int c(x, \gamma(y)) P'Q(dx, dy) \\
&\quad + (1-a) \int c(x, \gamma(y)) P''Q(dx, dy) \Bigg) \\
&\geq \inf_{\gamma \in \Gamma} \Bigg(a \int c(x, \gamma(y)) P'Q(dx, dy) \Bigg) \\
&\quad + \inf_{\gamma \in \Gamma} \Bigg((1-a) \int c(x, \gamma(y)) P''Q(dx, dy) \Bigg) \\
&= aJ(P', Q) + (1-a)J(P'', Q)
\end{aligned}$$

$\diamond$

6.4 Filter Stability

The filter stability problem refers to the correction of an incorrectly initialized non-linear filter for a partially observed stochastic dynamical system (controlled or control-free) with increasing measurements. Let us describe this property more explicitly: Given a prior $\mu \in \mathcal{P}(\mathbb{X})$ and a policy $\gamma \in \Gamma$ we can then define the filter and predictor for a POMDP using the (strategic) measure $P^{\mu,\gamma}$.

Definition 6.4.1 *(i) We define the one step predictor process as the sequence of conditional probability measures*

$$\pi_{n-}^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n-1]}, U_{[0,n-1]}) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n-1]}) \quad n \in \mathbb{N}$$

(ii) We define the filter process as the sequence of conditional probability measures

$$\pi_n^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n]}, U_{[0,n-1]}) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n]}), \quad n \in \mathbb{Z}_+ \tag{6.31}$$

Remark 6.8. Recall that the $U_{[0,n-1]}$ are all functions of the $Y_{[0,n-1]}$, so conditioning on the control actions is not necessary in the above definitions. Yet this conditional probability would be *policy dependent*; if we condition on the past actions, this conditioning would be *policy-independent*.

Say a prior $\mu \in \mathcal{P}(\mathbb{X})$ and a policy $\gamma \in \Gamma$ are chosen, an observer sees measurements $Y_{[0,\infty)}$ generated via the strategic measure $P^{\mu,\gamma}$. The observer is aware that the policy applied is γ, but incorrectly thinks the prior is $\nu \neq \mu$. The observer will then compute the incorrectly initialized filter $\pi_n^{\nu,\gamma}$ while the true filter is $\pi_n^{\mu,\gamma}$. The filter stability problem is concerned with the merging of $\pi_n^{\nu,\gamma}$ and $\pi_n^{\mu,\gamma}$ as n goes to infinity.

It will be useful to note that the filter is the strategic measure conditioned on the sigma field $\mathcal{F}_{0,n}^{\mathbb{Y}}$ and restricted to the sigma field $\mathcal{F}_n^{\mathbb{X}}$.

$$\pi_n^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n]}) = P^{\mu,\gamma}|_{\mathcal{F}_n^{\mathbb{X}}} | \mathcal{F}_{0,n}^{\mathbb{Y}}$$

In the literature, there are a number of merging notions when one considers stability which we enumerate here. Let $C_b(\mathbb{X})$ represent the set of continuous and bounded functions from $\mathbb{X} \to \mathbb{R}$.

Definition 6.4.2 *Two sequences of probability measures P_n, Q_n merge weakly if $\forall\, f \in C_b(\mathbb{X})$ we have* $\lim_{n\to\infty} \left| \int f dP_n - \int f dQ_n \right| = 0$.

Definition 6.4.3 *For two probability measures P and Q we define the total variation norm as $\|P-Q\|_{TV} = \sup_{\|f\|_\infty \leq 1} \left|\int f dP - \int f dQ\right|$ where f is assumed measurable. We say two sequences of probability measures P_n, Q_n merge in total variation if $\|P_n - Q_n\|_{TV} \to 0$ as $n \to \infty$.*

Definition 6.4.4

(i) For two probability measures P and Q we define the relative entropy as $D(P\|Q) = \int \log \frac{dP}{dQ} dP = \int \frac{dP}{dQ} \log \frac{dP}{dQ} dQ$ where we assume $P \ll Q$ and $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative of P with respect to Q.

(ii) Let X and Y be two random variables, let P and Q be two different joint measures for (X,Y) with $P \ll Q$. Then we define the (conditional) relative entropy between $P(X|Y)$ and $Q(X|Y)$ as

$$\begin{aligned} D(P(X|Y)\|Q(X|Y)) &= \int \log \left(\frac{dP_{X|Y}}{dQ_{X|Y}}(x,y) \right) dP(x,y) \\ &= \int \left(\int \log \left(\frac{dP_{X|Y}}{dQ_{X|Y}}(x,y) \right) dP(x|Y=y) \right) dP(y) \end{aligned} \tag{6.32}$$

We define here the different notions of stability for the filter:

Definition 6.4.5 *(i) A filter process is said to be stable in the sense of weak merging with respect to a policy γ $P^{\mu,\gamma}$ almost surely (a.s.) if there exists a set of measurement sequences $A \subset \mathcal{Y}^{\mathbb{Z}_+}$ with $P^{\mu,\gamma}$ probability 1 such that for any sequence in A; for any $f \in C_b(\mathcal{X})$ and any prior ν with $\mu \ll \nu$ (i.e., for all Borel B $\nu(B) = 0 \implies \mu(B) = 0$) we have $\lim_{n\to\infty} \left|\int f d\pi_n^{\mu,\gamma} - \int f d\pi_n^{\nu,\gamma}\right| = 0$.*

(ii) A filter process is said to be stable in the sense of total variation in expectation with respect to a policy γ if for any measure ν with $\mu \ll \nu$ we have $\lim_{n\to\infty} E^{\mu,\gamma}[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}] = 0$.

(iii)A filter process is said to be stable in the sense of total variation with respect to a policy γ $P^{\mu,\gamma}$ a.s. if there exists a set of measurement sequences $A \subset \mathcal{Y}^{\mathbb{Z}_+}$ with $P^{\mu,\gamma}$ probability 1 such that for any sequence in A; for any measure ν with $\mu \ll \nu$ we have $\lim_{n\to\infty} \|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV} = 0$ $P^{\mu,\gamma}$ $a.s.$.

(iv)A filter process is said to be stable in the sense of relative entropy with respect to a policy γ if for any measure ν with $\mu \ll \nu$ we have $\lim_{n\to\infty} E^{\mu,\gamma}[D(\pi_n^{\mu,\gamma}\|\pi_n^{\nu,\gamma})] = 0$.

(v) The filter is said to be universally stable in one of the above notions if the notion holds with respect to every admissible policy $\gamma \in \Gamma$.

Predictor stability is defined in an analogous fashion for each of the criteria above.

Total variation merging implies weak merging, and relative entropy merging (i.e. $D(P_n\|Q_n) \to 0$) implies total variation merging via Pinsker's inequality [94].

One of the main differences between control-free and controlled partially observed Markov chains is that the filter is always Markovian under the former, whereas under a controlled model the filter process may not be Markovian since the control policy may depend on past measurements in an arbitrary (measurable) fashion. This complicates the dependency structure and therefore results from the control-free case do not directly apply to the controlled setup.

We made the observation earlier that under observability and a controllability assumption, any incorrectly initialized filter will converge to the correct Kalman filter (we note that partial convergence and robustness results on the asymptotic equivalence of conditional expectations and linear estimates for non-Gaussian priors for linear systems are reported in [292]). In the following, we will present a concise discussion on how such results carry over to the stochastic nonlinear setup.

Much of the results on filter stability involves control-free systems. Thus, results have considered partially observed Markov processes (POMP) as opposed to partially observed Markov *decision* processes (POMDP). Since there is no control in such systems, there is no past dependency in the system and the pair $(X_n, Y_n)_{n=0}^{\infty}$ is always a Markov chain. For such control-free models, filter stability has been studied extensively and we refer the reader to [86] for

a comprehensive review and a collection of different approaches. As discussed in [86], filter stability arises via two separate mechanisms:

1. The transition kernel is in some sense *sufficiently* ergodic, forgetting the initial measure and therefore passing this insensitivity (to incorrect initializations) on to the filter process.
2. The measurement channel provides sufficient information about the underlying state, allowing the filter to track the true state process.

To be able to present a concise discussion, building on some prior material in the notes, for both controlled and control-free setups we review conditions in [225] based on Dobrushin's coefficients of the measurement channel and the controlled transition kernel. Recall (3.29). We consider a slight generalization in the following.

Definition 6.4.6 *[106, Equation 1.16] For a kernel operator $K : S_1 \to \mathcal{P}(S_2)$ (that is a regular conditional probability from S_1 to S_2) for standard Borel spaces S_1, S_2, we define the Dobrushin coefficient as:*

$$\delta(K) = \inf \sum_{i=1}^{n} \min(K(x, A_i), K(y, A_i)) \tag{6.33}$$

where the infimum is over all $x, y \in S_1$ and all partitions $\{A_i\}_{i=1}^n$ of S_2.

Let us define

$$\tilde{\delta}(T) := \inf_{u \in \mathbb{U}} \delta(\mathcal{T}(\cdot|\cdot, u)).$$

The following can be viewed as a generalization of Theorem 3.1.8.

Theorem 6.4.1 *[225, Theorem 3.3] Assume that for $\mu, \nu \in \mathcal{P}(\mathbb{X})$, we have $\mu \ll \nu$. Then we have*

$$E^{\mu,\gamma}\left[\|\pi_{n+1}^{\mu,\gamma} - \pi_{n+1}^{\nu,\gamma}\|_{TV}\right] \leq (1 - \tilde{\delta}(T))(2 - \delta(Q))E^{\mu,\gamma}\left[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}\right].$$

In particular, defining $\alpha := (1 - \tilde{\delta}(T))(2 - \delta(Q))$, we have

$$E^{\mu,\gamma}\left[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}\right] \leq 2\alpha^n.$$

By applying the Borel-Cantelli lemma and Markov's inequality, we have that exponential stability in expectation implies the same result in an almost sure sense as well: assume that the filter is exponentially stable with coefficient $\alpha = (1 - \delta(T))(2 - \delta(Q) < 1$ and let ρ be a value $\rho < \frac{1}{\alpha}$. Then we have for every $\epsilon > 0$,

$$\begin{aligned}
\sum_{k=0}^{\infty} P^{\mu}(\rho^k \|\pi_n^{\mu} - \pi_n^{\nu}\|_{TV} \geq \epsilon) &\leq \sum_{k=0}^{\infty} \rho^k \frac{E^{\mu}[\|\pi_n^{\mu} - \pi_n^{\nu}\|_{TV}]}{\epsilon} \\
&\leq \frac{\|\mu - \nu\|_{TV}}{\epsilon} \sum_{k=0}^{\infty} (\rho\alpha)^k \\
&= \frac{\|\mu - \nu\|_{TV}}{\epsilon} \frac{1}{1 - \rho\alpha} \\
&< \infty
\end{aligned}$$

thus by Borel Cantelli Lemma $\rho^k \|\pi_n^{\mu} - \pi_n^{\nu}\|_{TV} \to 0$ P^{μ} a.s. for any $\rho < \frac{1}{\alpha}$. See [225, Remark 3.10].

This also establishes that the rate of convergence is uniform over all priors ν as long as $\mu \ll \nu$.

A further approach to ensuring filter stability, via sample paths, is by an analysis which builds on the Hilbert projective metric: [146]

Definition 6.4.7 *Two non-negative measures μ, ν on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ are comparable, if there exist positive constants $0 < a \leq b$, such that*

$$a\nu(A) \leq \mu(A) \leq b\nu(A)$$

for any Borel subset $A \subset \mathbb{X}$.

Definition 6.4.8 (Mixing kernel) *The non-negative kernel K defined on $\mathbb{X}$ is mixing, if there exists a constant $0 < \varepsilon \leq 1$, and a non-negative measure λ on $\mathbb{X}$, such that*

$$\varepsilon\lambda(A) \leq K(x, A) \leq \frac{1}{\varepsilon}\lambda(A)$$

for any $x \in \mathbb{X}$, and any Borel subset $A \subset \mathbb{X}$.

Definition 6.4.9 *(Hilbert metric). Let μ, ν be two non-negative finite measures. We define the Hilbert metric on such measures as*

$$h(\mu, \nu) = \begin{cases} \log\left(\frac{\sup_{A|\nu(A)>0} \frac{\mu(A)}{\nu(A)}}{\inf_{A|\nu(A)>0} \frac{\mu(A)}{\nu(A)}}\right) & \text{if } \mu, \nu \text{ are comparable} \\ 0 & \text{if } \mu = \nu = 0 \\ \infty & \text{else} \end{cases} \tag{6.34}$$

Note that $h(a\mu, b\nu) = h(\mu, \nu)$ for any positive scalars a, b. Therefore, the Hilbert metric is a useful metric for nonlinear filters since it is invariant under normalization, and the following lemma demonstrates that it bounds the total-variation distance.

Lemma 6.4.1 *[146, Lemma 3.4] Let μ, ν be two non-negative finite measures,*

i. $\|\mu - \nu\|_{TV} \leq \frac{2}{\log 3} h(\mu, \nu)$.

ii. If the nonnegative kernel K is a mixing kernel (see Definition 6.4.8) with constant ϵ, then $h(K\mu, K\nu) \leq \frac{1}{\varepsilon^2}\|\mu - \nu\|_{TV}$.

Lemma 6.4.2 ([146], Lemma 3.8) *The nonnegative linear operator τ on $\mathcal{M}^+(\mathbb{X})$ (positive measures on $\mathbb{X}$) associated with a nonnegative kernel K defined on $\mathbb{X}$*

$$\tau(K) := \sup_{0<h(\mu,\nu)<\infty} \frac{h(K\mu, K\nu)}{h(\mu, \nu)} = \tanh\left[\frac{1}{4}H(K)\right]$$

where

$$H(K) := \sup_{\mu,\nu} h(K\mu, K\nu)$$

is over nonnegative measures, is a contraction (called the Birkhoff contraction coefficient) , is a contraction under the Hilbert metric if $H(K) < \infty$ (which implies $\tau(K) < 1$).

A controlled version of a contraction via the Hilbert metric [146] can be obtained [101]:

Recall that

$$F(z, y, u)(\cdot) = P\{x_{k+1} \in \cdot \mid \pi_k = z, Y_{k+1} = y, U_k = u\}$$

Assumption 6.4.1 *1. $Q(y|x) \geq \epsilon > 0$ for every $x \in \mathbb{X}$ and $y \in \mathbb{Y}$.*

2. The transition kernel $\mathcal{T}(.|., u)$ is a mixing kernel (see Definition 6.4.8) for every $u \in \mathbb{U}$.

Lemma 6.4.3 *[1] Under Assumption 8.3.1, there exists a constant $r < 1$ such that*

$$h(F(\mu,y,u),F(\nu,y,u)) \leq rh(\mu,\nu) \tag{6.35}$$

for every comparable $\mu,\nu \in \mathcal{P}(\mathbb{X})$ and for every $u \in \mathbb{U}$ and $y \in \mathbb{Y}$. Here $r = \frac{1-\epsilon_u^2\epsilon}{1+\epsilon_u^2\epsilon}$, ϵ_u is the mixing constant of the kernel $\mathcal{T}(.|.,u)$.

Another filter stability result which will also be useful in numerical methods for POMDPs to be considered later is via the following *stochastic non-linear observability* definition.

Definition 6.4.10 *[Stochastic Observability for Non-Linear Systems] [226] A POMDP is called one step observable (universal in admissible control policies) if for every $f \in C_b(\mathbb{X})$ and every $\epsilon > 0$ there exists a measurable and bounded function g such that*

$$\|f(\cdot) - \int_{\mathbb{Y}} g(y)Q(dy|\cdot)\|_\infty < \epsilon \tag{6.36}$$

Theorem 6.4.2 *[226] Assume that $\mu \ll \nu$ and that the POMDP is one step observable. Then the predictor is universally stable weakly a.s. .*

We now present an example for observability.

Example 6.9. [227] Consider a finite setup $\mathbb{X} = \{a_1,\cdots,a_n\}$ and let the noise space be $\mathbb{V} = \{b_1,\cdots,b_m\}$. Now, assume $y = h(x,v)$ has K distinct outputs, where $1 \leq K \leq (n)(m)$ and $\mathbb{Y} = \{c_1,\cdots,c_K\}$. We note that for such a setup, there is already a sufficient and necessary condition for filter stability provided in [315, Theorem V.2] (see also [313]. We examine this case to show that Definition 6.37 above leads to filter stability.

For each x, h_x can be viewed as a partition of $\mathbb{V}$, assigning each $b_i \in \mathbb{X}$ to an output level $c_j \in \mathbb{Y}$. We can track this by the matrix $H_x(i,j) = 1$ if $h_x(b_i) = y_j$ and zero else. Let Q be the $1 \times m$ vector representing the probability measure of the noise. We consider the one step observability (though this can be generalized for the control-free case to N-step observability for $N > 1$). Let $g(c_i) = \alpha_i$, with $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{bmatrix}$ and $\int_{\mathbb{V}} g(h(x,v))Q(dv) =: (QH_x)\alpha$. Any function $f(x)$ can be expressed as a $n \times 1$ vector and hence the question reduces to finding a vector α so that $f = QH\alpha$, and the system is one step observable if and only if the matrix $A \equiv \begin{bmatrix} QH_{a_1} \\ \vdots \\ QH_{a_n} \end{bmatrix}$ is rank n.

Since the spaces are finite, Theorem 6.4.4, to be presented below, leads to filter stability, both in total variation and weakly, in expectation.

Further examples for measurement channels satisfying Definition 6.4.10 have been reported in [227, Section 3][2].

The observability notion defined above only results in stability of the predictor in the weak sense $P^{\mu,\gamma}$ a.s. .We next extend this stability to total variation $P^{\mu,\gamma}$ a.s. .Let the measurement channel Q be *dominated* in the sense that there

[2]For control-free systems, [227] defines the following: A control-free filter is *N-step observable* if for every $f \in C_b(\mathcal{X})$ and every $\epsilon > 0$ there exists a measurable and bounded function g such that

$$\|f(\cdot) - \int_{\mathcal{Y}} g(y_{[1,N]})Q(dy_{[1,N]}|X_1 = \cdot)\|_\infty < \epsilon \tag{6.37}$$

A further notion is observability: A POMP is *observable* if for every $f \in C_b(\mathcal{X})$ and every $\epsilon > 0$ there exists $N \in \mathbb{N}$ and a measurable and bounded function g such that (6.37) applies. Due to the presence of dual effect, these N-step definitions require a more refined approach for filter stability.

exists a reference measure λ such that $\forall x \in \mathcal{X}, Q(Y \in \cdot | x_n = x) \ll \lambda(\cdot)$. Then, we define the Radon-Nikodym derivative

$$q(x, y) := \frac{dQ(Y_n \in \cdot | x_n = x)}{d\lambda}(y) \tag{6.38}$$

which serves as a likelihood function. We will consider one of the following assumptions.

Assumption 6.4.2 *(i) $\mathcal{T}(\cdot|x, u)$ is absolutely continuous with respect to a dominating measure ϕ for every $x \in \mathcal{X}, u \in \mathcal{U}$, so that $t(x_1, x, u) = \frac{d\mathcal{T}(\cdot|x,u)}{d\phi}(x_1)$ where t is continuous in x for every $x_1 \in \mathcal{X}$ and $u \in \mathcal{U}$.*

(ii) $q(x, y)$ is bounded and continuous in x for every fixed y. Furthermore, $q(x, y) > 0$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$.

Assumption 6.4.3 *$\mathcal{T}(\cdot|x, u)$ is absolutely continuous with respect to a dominating measure ϕ for every $x \in \mathcal{X}, u \in \mathcal{U}$, so that $s(x_1, x, u) = \frac{d\mathcal{T}(\cdot|x,u)}{d\phi}(x_1)$. The family of (conditional densities) $\{s(\cdot, x, u)\}_{x\in\mathcal{X},u\in\mathcal{U}}$ is uniformly bounded and equicontinuous.*

Theorem 6.4.3 *[226] Let $\mu \ll \nu$. Let Assumption 6.4.2 or Assumption 6.4.3 hold. If the predictor is universally stable in the weak sense a.s. then it is also universally stable in total variation a.s. .*

One of the key steps in the proof of Theorem 6.4.2 is that $P^{\mu,\gamma}(Y_n \in \cdot | Y_{[0,n-1]})$ and $P^{\nu,\gamma}(Y_n \in \cdot | Y_{[0,n-1]})$ merge in total variation $P^{\mu,\gamma}$ a.s. as $n \to \infty$. To achieve this in a POMDP, we apply Blackwell and Dubins [50] to the measurement process $\{Y_n\}_{n=0}^{\infty}$. However, [50] is fundamentally about predictive measures of the future given the past, and hence only directly implies predictor stability results, not the filter. Filter stability is studied next.

Assumption 6.4.4 *The measurement channel Q is continuous in total variation. That is, for any sequence $a_n \to a \in \mathcal{X}$ we have $\|Q(\cdot|a_n) - Q(\cdot|a)\|_{TV} \to 0$ or in other words $\|P(Y_0 \in \cdot | X_0 = a_n) - P(Y_0 \in \cdot | X_0 = a)\|_{TV} \to 0$.*

Assumption 6.4.2(ii), together with the related domination condition (6.38), implies Assumption 6.4.4 (see [184, Section 2.3]); see also [184, Theorem 3] for a partial converse result.

Theorem 6.4.4 *[226]*

(i) Let Assumption 6.4.4 hold. If the predictor is universally stable in weak merging a.s. , then the filter is universally stable in weak merging in expectation.

(ii) The filter is universally stable in total variation in expectation if and only if the predictor is universally stable in total variation in expectation.

(iii)The filter is universally stable in total variation in expectation if and only if it is universally stable in total variation a.s. .

(iv)Let $\mu \ll \nu$, and assume for any policy γ there exists some finite n such that $E^{\mu,\gamma}[D(\pi_n^{\mu,\gamma} \| \pi_n^{\mu,\gamma})] < \infty$ and some m such that $E^{\mu,\gamma}[D(P^{\mu,\gamma}|_{\mathcal{F}_{0,m}^{\mathcal{Y}}} \| (P^{\nu,\gamma}|_{\mathcal{F}_{0,m}^{\mathcal{Y}}})] < \infty$. Then the filter is universally stable in relative entropy if and only if it is universally stable in total variation in expectation.

Applications of these will be discussed in the context of numerical methods for POMDPs later in the notes. Filter stability is also related to robustness of optimal costs to incorrect initializations for controlled models [226].

6.5 Bibliographic Notes

Earlier work on separation results for partially observed Markov Decision Processes include [352], [296], [262]. For linear systems, classical texts include [10, 11, 35, 74, 200, 206, 207]. See [86] for a comprehensive review on filter stability. A very comprehensive recent book on POMDPs is [197].

It has been shown relatively recently that one could approach the Riccati/Kalman Filter updates as a contraction map in positive-definite matrices [69] (see also [215] and [217]), leading to a concise and direct proof of convergence as well as stability (though with strict controllability and observability conditions, instead of detectability and stabilizability). On filter stability, related work in the control-free domain includes [85, 86, 159].

6.6 Appendix

6.6.1 Proof of Theorems 6.3.4 and 6.3.3.

We present the unified proof given in [184]. We first recall (D.3) which is used to metrize weak convergence. The following result plays a key role.

Lemma 6.6.1 *[184] Let $\mathbb{X}$ be a Borel space. Suppose that we have a family of uniformly bounded real Borel measurable functions $\{f_{n,\lambda}\}_{n\geq 1,\lambda\in\Lambda}$ and $\{f_\lambda\}_{\lambda\in\Lambda}$, for some set Λ. If, for any $x_n \to x$ in $\mathbb{X}$, we have*

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}|f_{n,\lambda}(x_n)-f_\lambda(x)|=0 \tag{6.39}$$

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}|f_\lambda(x_n)-f_\lambda(x)|=0, \tag{6.40}$$

then, for any $\mu_n \to \mu$ weakly in $\mathcal{P}(\mathbb{X})$, we have

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}\left|\int_{\mathbb{X}} f_{n,\lambda}(x)\mu_n(dx)-\int_{\mathbb{X}} f_\lambda(x)\mu(dx)\right|=0.$$

In Theorem 6.3.3 and Theorem 6.3.4, we need to show that, for every $(z_0^n,u_n)\to(z_0,u)$ in $\mathbb{Z}\times\mathbb{U}$, we have

$$\sup_{\|f\|_{BL}\leq 1}\left|\int_{\mathbb{Z}} f(z_1)\eta(dz_1|z_0^n,u_n)-\int_{\mathbb{Z}} f(z_1)\eta(dz_1|z_0,u)\right|\to 0,$$

where we equip $\mathbb{Z}$ with the metric ρ to define bounded-Lipschitz norm $\|f\|_{BL}$ of any Borel measurable function $f:\mathbb{Z}\to\mathbb{R}$. We can equivalently write this as

$$\sup_{\|f\|_{BL}\leq 1}\left|\int_{\mathbb{Y}} f(z_1(z_0^n,u_n,y_1))P(dy_1|z_0^n,u_n)-\int_{\mathbb{Y}} f(z_1(z_0,u,y_1))P(dy_1|z_0,u)\right|\to 0. \tag{6.41}$$

The term in equation (6.41) can be upper bounded as follows:

$$\begin{aligned}
&\sup_{\|f\|_{BL}\leq 1}\left|\int_{\mathbb{Y}} f(z_1(z_0^n,u_n,y_1))P(dy_1|z_0^n,u_n)-\int_{\mathbb{Y}} f(z_1(z_0,u,y_1))P(dy_1|z_0,u)\right|\\
&\leq \sup_{\|f\|_{BL}\leq 1}\left|\int_{\mathbb{Y}} f(z_1(z_0^n,u_n,y_1))P(dy_1|z_0^n,u_n)-\int_{\mathbb{Y}} f(z_1(z_0^n,u_n,y_1))P(dy_1|z_0,u)\right|\\
&\quad+\sup_{\|f\|_{BL}\leq 1}\int_{\mathbb{Y}}\left|f(z_1(z_0^n,u_n,y_1))-f(z_1(z_0,u,y_1))\right|P(dy_1|z_0,u)\\
&\leq \|P(\cdot|z_0^n,u_n)-P(\cdot|z_0,u)\|_{TV}\\
&\quad+\sup_{\|f\|_{BL}\leq 1}\int_{\mathbb{Y}}\left|f(z_1(z_0^n,u_n,y_1))-f(z_1(z_0,u,y_1))\right|P(dy_1|z_0,u),
\end{aligned} \tag{6.42}$$

where, in the last inequality, we have used $\|f\|_\infty \leq \|f\|_{BL}\leq 1$. To prove that (6.42) (and so (6.41)) goes to 0, it is sufficient to establish the following results:

(i) $P(dy_1|z_0,u_0)$ is continuous in total variation,

(ii) $\lim_{n\to\infty}\int_{\mathbb{Y}}\rho(z_1(z_0^n,u_n,y_1),z_1(z_0,u,y_1))P(dy_1|z_0,u)=0$ as $(z_0^n,u_n)\to(z_0,u)$.

Indeed, suppose that (i) and (ii) hold. Then, the first term in (6.42) goes to 0 as $P(\cdot|z_0,u)$ is continuous in total variation. For the second term in (6.42), we have

$$\begin{aligned}
&\sup_{\|f\|_{BL}\le 1}\int_{\mathbb{Y}}\big|f(z_1(z_0^n,u_n,y_1))-f(z_1(z_0,u,y_1))\big|P(dy_1|z_0,u)\\
&\le\int_{\mathbb{Y}}\rho(z_1(z_0^n,u_n,y_1),z_1(z_0,u,y_1))P(dy_1|z_0,u)\\
&\to 0 \text{ as } n\to\infty \quad \text{(by (ii))}.
\end{aligned}$$

Therefore, to complete the proof of Theorem 6.3.3 and Theorem 6.3.4, we will prove (i) and (ii).

Proof of Theorem 6.3.3

We first prove (i); that is, $P(dy_1|z_0,u)$ is continuous in total variation. To this end, let $(z_0^n,u_n)\to(z_0,u)$. Then, we write

$$\begin{aligned}
&\sup_{A\in\mathcal{B}(\mathbb{Y})}\big|P(A|z_0^n,u_n)-P(A|z_0,u)\big|\\
&=\sup_{A\in\mathcal{B}(\mathbb{Y})}\left|\int_{\mathbb{X}}Q(A|x_1,u_n)\mathcal{T}(dx_1|z_0^n,u_n)-\int_{\mathbb{X}}Q(A|x_1,u)\mathcal{T}(dx_1|z_0,u)\right|,
\end{aligned}$$

where $\mathcal{T}(dx_1|z_0^n,u_n):=\int_{\mathbb{X}}\mathcal{T}(dx_1|x_0,u_n)z_0^n(dx_0)$. Note that, by Lemma 6.6.1, we can show that $\mathcal{T}(dx_1|z_0^n,u_n)\to\mathcal{T}(dx_1|z_0,u)$ weakly. Indeed, if $g\in C_b(\mathbb{X})$, then we define $r_n(x_0)=\int_{\mathbb{X}}g(x_1)\mathcal{T}(dx_1|x_0,u_n)$ and $r(x_0)=\int_{\mathbb{X}}g(x_1)\mathcal{T}(dx_1|x_0,u)$. Since $\mathcal{T}(dx_1|x_0,u)$ is weakly continuous, we have $r_n(x_0^n)\to r(x_0)$ when $x_0^n\to x_0$. Hence, by Lemma 6.6.1, we have

$$\lim_{n\to\infty}\left|\int_{\mathbb{X}}r_n(x_0)z_0^n(dx_0)-\int_{\mathbb{X}}r(x_0)z_0(dx_0)\right|=0.$$

Hence, $\mathcal{T}(dx_1|z_0^n,u_n)\to\mathcal{T}(dx_1|z_0,u)$ weakly. Moreover, the families of functions $\{Q(A|\cdot,u_n)\}_{n\ge1,A\in\mathcal{B}(\mathbb{Y})}$ and $\{Q(A|\cdot,u)\}_{A\in\mathcal{B}(\mathbb{Y})}$ satisfy the conditions of Lemma 6.6.1 as Q is continuous in total variation distance. Therefore, Lemma 6.6.1 yields that

$$\lim_{n\to\infty}\sup_{A\in\mathcal{B}(\mathbb{Y})}\left|\int_{\mathbb{X}}Q(A|x_1,u_n)\mathcal{T}(dx_1|z_0^n,u_n)-\int_{\mathbb{X}}Q(A|x_1,u)\mathcal{T}(dx_1|z_0,u)\right|=0.$$

Thus, $P(dy_1|z_0,u)$ is continuous in total variation.

To prove (ii), we write

$$\begin{aligned}
&\int_{\mathbb{Y}}\rho(z_1(z_0^n,u_n,y_1),z_1(z_0,u,y_1))P(dy_1|z_0,u)\\
&=\int_{\mathbb{Y}}\sum_{m=1}^{\infty}2^{-m+1}\left|\int_{\mathbb{X}}f_m(x_1)z_1(z_0^n,u_n,y_1)(dx_1)\right.\\
&\qquad\qquad\left.-\int_{\mathbb{X}}f_m(x_1)z_1(z_0,u,y_1)(dx_1)\right|P(dy_1|z_0,u)\\
&=\sum_{m=1}^{\infty}2^{-m+1}\int_{\mathbb{Y}}\left|\int_{\mathbb{X}}f_m(x_1)z_1(z_0^n,u_n,y_1)(dx_1)\right.\\
&\qquad\qquad\left.-\int_{\mathbb{X}}f_m(x_1)z_1(z_0,u,y_1)(dx_1)\right|P(dy_1|z_0,u),
\end{aligned}$$

where we have used Fubini's theorem with the fact that $\sup_m\|f_m\|_\infty\le1$. For each m, let us define

$$\begin{aligned}
I_+^{(n)}&:=\left\{y_1\in\mathbb{Y}:\int_{\mathbb{X}}f_m(x_1)z_1(z_0^n,u_n,y_1)(dx_1)>\int_{\mathbb{X}}f_m(x_1)z_1(z_0,u,y_1)(dx_1)\right\}\\
I_-^{(n)}&:=\left\{y_1\in\mathbb{Y}:\int_{\mathbb{X}}f_m(x_1)z_1(z_0^n,u_n,y_1)(dx_1)\le\int_{\mathbb{X}}f_m(x_1)z_1(z_0,u,y_1)(dx_1)\right\}.
\end{aligned}\tag{6.43}$$

Then, we can write

$$\begin{aligned}
&\int_{\mathbb{Y}} \left| \int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) - \int_{\mathbb{X}} f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right| P(dy_1|z_0, u) \\
&= \int_{I_+^{(n)}} \left(\int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) - \int_{\mathbb{X}} f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right) P(dy_1|z_0, u) \\
&+ \int_{I_-^{(n)}} \left(\int_{\mathbb{X}} f_m(x_1) z_1(z_0, u, y_1)(dx_1) - \int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) \right) P(dy_1|z_0, u).
\end{aligned}$$

In the sequel, we only consider the term with the set $I_+^{(n)}$. The analysis for the other one follows from the same steps. We have

$$\begin{aligned}
&\int_{I_+^{(n)}} \left(\int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) - \int_{\mathbb{X}} f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right) P(dy_1|z_0, u) \\
&\leq \int_{I_+^{(n)}} \int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) P(dy_1|z_0, u) \\
&\qquad - \int_{I_+^{(n)}} \int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) P(dy_1|z_0^n, u_n) \\
&+ \int_{I_+^{(n)}} \int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) P(dy_1|z_0^n, u_n) \\
&\qquad - \int_{I_+^{(n)}} \int_{\mathbb{X}} f_m(x_1) z_1(z_0, u, y_1)(dx_1) P(dy_1|z_0, u) \\
&\leq \|P(dy_1|z_0, u) - P(dy_1|z_0^n, u_n)\|_{TV} \\
&\quad + \int_{\mathbb{X}} \int_{I_+^{(n)}} f_m(x_1) Q(dy_1|x_1, u_n) \mathcal{T}(dx_1|z_0^n, u_n) \\
&\qquad\qquad - \int_{\mathbb{X}} \int_{I_+^{(n)}} f_m(x_1) Q(dy_1|x_1, u) \mathcal{T}(dx_1|z_0, u),
\end{aligned}$$

where we have used $\|f_m\|_\infty \leq 1$ in the last inequality. The first term above goes to 0 since $P(dy_1|z_0, u)$ is continuous in total variation. For the second term, we use Lemma 6.6.1. Indeed, families of functions $\{f_m(\cdot)Q(A|\cdot, u_n) : n \geq 1, A \in \mathcal{B}(\mathbb{Y})\}$ and $\{f_m(\cdot)Q(A|\cdot, u) : A \in \mathcal{B}(\mathbb{Y})\}$ satisfy the conditions in Lemma 6.6.1 as Q is continuous in total variation. Hence, the second term converges to 0 by Lemma 6.6.1 since $\mathcal{T}(dx_1|z_0^n, u_n) \to \mathcal{T}(dx_1|z_0, u)$ weakly. Hence, for each m, we have

$$\begin{aligned}
\lim_{n\to\infty} \int_{\mathbb{Y}} \Bigg| \int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) \\
- \int_{\mathbb{X}} f_m(x_1) z_1(z_0, u, y_1)(dx_1) \Bigg| P(dy_1|z_0, u) = 0.
\end{aligned}$$

By the dominated convergence theorem, we then have

$$\begin{aligned}
&\lim_{n\to\infty} \int_{\mathbb{Y}} \rho(z_1(z_0^n, u_n, y_1), z_1(z_0, u, y_1)) P(dy_1|z_0, u) \\
&\leq \sum_{m=1}^{\infty} 2^{-m+1} \lim_{n\to\infty} \int_{\mathbb{Y}} \left| \int_{\mathbb{X}} f_m(x_1) z_1(z_0^n, u_n, y_1)(dx_1) \right. \\
&\qquad \left. - \int_{\mathbb{X}} f_m(x_1) z_1(z_0, u, y_1)(dx_1) \right| P(dy_1|z_0, u) = 0.
\end{aligned}$$

This establishes (ii), which completes the proof together with (i).

Proof of Theorem 6.3.4

We first show (i); that is, $P(dy_1|z_0, u_0)$ is continuous total variation. Let $(z_0^n, u_n) \to (z_0, u)$. Then, we have

$$\sup_{A\in\mathcal{B}(\mathbb{Y})} |P(A|z_0^n, u_n) - P(A|z_0, u)|$$

$$= \sup_{A\in\mathcal{B}(\mathbb{Y})} \Big| \int_{\mathbb{X}}\int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0,u_n)z_0^n(dx_0) - \int_{\mathbb{X}}\int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0,u)z_0(dx_0)\Big|.$$

For each $A\in\mathcal{B}(\mathbb{Y})$ and $n\geq 1$, we define

$$f_{n,A}(x_0) = \int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0,u_n)$$

and

$$f_A(x_0) = \int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0,u).$$

Then, for all $x_0^n\to x_0$, we have

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{A\in\mathcal{B}(\mathbb{Y})}|f_{n,A}(x_0^n)-f_A(x_0)|\\
&=\lim_{n\to\infty}\sup_{A\in\mathcal{B}(\mathbb{Y})}\Big|\int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0^n,u_n)-\int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0,u)\Big|\\
&\leq \lim_{n\to\infty}\|\mathcal{T}(dx_1|x_0^n,u_n)-\mathcal{T}(dx_1|x_0,u)\|_{TV}=0
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{A\in\mathcal{B}(\mathbb{Y})}|f_{A}(x_0^n)-f_A(x_0)|\\
&=\lim_{n\to\infty}\sup_{A\in\mathcal{B}(\mathbb{Y})}\Big|\int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0^n,u)-\int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0,u)\Big|\\
&\leq \lim_{n\to\infty}\|\mathcal{T}(dx_1|x_0^n,u)-\mathcal{T}(dx_1|x_0,u)\|_{TV}=0.
\end{aligned}$$

Then, by Lemma 6.6.1, we have

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{A\in\mathcal{B}(\mathbb{Y})}\Big|\int_{\mathbb{X}} f_{n,A}(x_0)z_0^n(dx_0)-\int_{\mathbb{X}} f_A(x_0)z_0(dx_0)\Big|\\
&=\lim_{n\to\infty}\sup_{A\in\mathcal{B}(\mathbb{Y})}\Big|\int_{\mathbb{X}}\int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0,u_n)z_0^n(dx_0)\\
&\qquad\qquad -\int_{\mathbb{X}}\int_{\mathbb{X}} Q(A|x_1)\mathcal{T}(dx_1|x_0,u)z_0(dx_0)\Big|\\
&=0.
\end{aligned}$$

Hence, $P(dy_1|z_0,u_0)$ is continuous in total variation.

Now, we show (ii); that is, for any $(z_0^n,u_n)\to(z_0,u)$, we have

$$\lim_{n\to\infty}\int_{\mathbb{Y}}\rho(z_1(z_0^n,u_n,y_1),z_1(z_0,u,y_1))P(dy_1|z_0,u)=0.$$

From the proof of Theorem 6.3.3, it suffices to show that

$$\lim_{n\to\infty}\int_{\mathbb{X}}\int_{I_+^{(n)}} f_m(x_1)Q(dy_1|x_1)\mathcal{T}(dx_1|z_0^n,u_n) - \int_{\mathbb{X}}\int_{I_+^{(n)}} f_m(x_1)Q(dy_1|x_1)\mathcal{T}(dx_1|z_0,u)=0. \tag{6.44}$$

Indeed, we have

$$\begin{aligned}
&\Big|\int_{\mathbb{X}}\int_{I_+^{(n)}} f_m(x_1)Q(dy_1|x_1)\mathcal{T}(dx_1|z_0^n,u_n)-\int_{\mathbb{X}}\int_{I_+^{(n)}} f_m(x_1)Q(dy_1|x_1)\mathcal{T}(dx_1|z_0,u)\Big|\\
&\leq\Big|\int_{\mathbb{X}^2} f_m(x_1)Q(I_+^{(n)}|x_1)\mathcal{T}(dx_1|x_0,u_n)z_0^n(dx_0)
\end{aligned}$$

$$
\begin{aligned}
&\qquad - \int_{\mathbb{X}^2} f_m(x_1) Q(I_+^{(n)}|x_1) \mathcal{T}(dx_1|x_0,u) z_0^n(dx_0) \Big| \\
&+ \Big| \int_{\mathbb{X}^2} f_m(x_1) Q(I_+^{(n)}|x_1) \mathcal{T}(dx_1|x_0,u) z_0^n(dx_0) \\
&\qquad - \int_{\mathbb{X}^2} f_m(x_1) Q(I_+^{(n)}|x_1) \mathcal{T}(dx_1|x_0,u) z_0(dx_0) \Big| \\
&\leq \int_{\mathbb{X}} \|\mathcal{T}(dx_1|x_0,u_n) - \mathcal{T}(dx_1|x_0,u)\|_{TV} z_0^n(dx_0) \\
&+ \Big| \int_{\mathbb{X}^2} f_m(x_1) Q(I_+^{(n)}|x_1) \mathcal{T}(dx_1|x_0,u) z_0^n(dx_0) \\
&\qquad - \int_{\mathbb{X}^2} f_m(x_1) Q(I_+^{(n)}|x_1) \mathcal{T}(dx_1|x_0,u) z_0(dx_0) \Big|,
\end{aligned}
$$

where we have used $\sup_{n\geq 1} \sup_{x_1 \in \mathbb{X}} \left| f_m(x_1) Q(I_+^{(n)}|x_1) \right| \leq 1$ in the last inequality. If we define $r_n(x_0) = \|\mathcal{T}(dx_1|x_0,u_n) - \mathcal{T}(dx_1|x_0,u)\|_{TV}$, then $r_n(x_0^n) \to 0$ whenever $x_0^n \to x_0$. Then, the first term converges to 0 by Lemma 6.6.1 as $z_0^n \to z_0$ weakly. The second term also converges to 0 by Lemma 6.6.1, since $\{\int_{\mathbb{X}} f(x_1) Q(I_+^{(n)}|x_1) \mathcal{T}(dx_1|\cdot,u) : n \geq 1\}$ is a family of uniformly bounded and equicontinuous functions by total variation continuity of $\mathcal{T}(dx_1|x_0,u)$. This proves (ii) and completes the proof together with (i).

$\diamond$

6.7 Exercises

Exercise 6.7.1 *Consider a linear system with the following dynamics:*

$$x_{t+1} = ax_t + u_t + w_t,$$

and let the controller have access to the observations given by:

$$y_t = p_t(x_t + v_t).$$

Here $\{w_t, v_t, t \in \mathbb{Z}\}$ are independent, zero-mean, Gaussian random variables, with variances $E[w^2]$ and $E[v^2]$. The controller at time $t \in \mathbb{Z}$ has access to $I_t = \{y_s, u_s, p_t \quad s \leq t-1\} \cup \{y_t\}$. Here p_t is an i.i.d. Bernoulli process such that $p_t = 1$ with probability p.

The initial state has a Gaussian distribution, with zero mean and variance $E[x_0^2]$, which we denote by ν_0. We wish to find for some $r > 0$:

$$\inf_{\gamma} J(x_0, \gamma) = E_{\nu_0}^{\gamma}[\sum_{t=0}^{3} x_t^2 + r u_t^2],$$

Compute the optimal control policy and the optimal cost. It suffices to provide a recursive form.

Hint: Show that the optimal control has a separation structure. Compute the conditional estimate through a revised Kalman Filter due to the presence of p_t.

Exercise 6.7.2 *Let X, Y be $\mathbb{R}^n$ and $\mathbb{R}^m$ valued zero-mean random vectors defined on a common probability space, which have finite covariance matrices. Suppose that their probability measures are given by P_X and P_Y respectively.*

Find

$$\inf_{K} E[(X - KY)^T (X - KY)],$$

that is find the best linear estimator of X given Y and the resulting estimation error.

Hint: You may pose the problem as a Projection Theorem problem.

Exercise 6.7.3 (Optimal Machine Repair) *Consider a POMDP given by the following description. Let there be two possible states that a machine can take: $\mathbb{X} = \{0, 1\}$, where 0 is the bad ('system is down') state and 1 is the good state. Let $\mathbb{U} = \{0, 1\}$, where 0 is the 'do nothing' control and 1 is the 'repair' control. Suppose that the transition probabilities are given by:*

$$\begin{aligned}
&P(X_{t+1} = 1|X_t = 1, U_t = 0) = 1 - \eta_1, \qquad P(X_{t+1} = 0|X_t = 1, U_t = 0) = \eta_1 > 0 \\
&P(X_{t+1} = 1|X_t = 1, U_t = 1) = 1 - \eta_2, \qquad P(X_{t+1} = 0|X_t = 1, U_t = 1) = \eta_2 > 0 \\
&P(X_{t+1} = 1|X_t = 0, U_t = 0) = 0, \qquad P(X_{t+1} = 0|X_t = 0, U_t = 0) = 1 \\
&P(X_{t+1} = 1|X_t = 0, U_t = 1) = \alpha > 0, \qquad P(X_{t+1} = 0|X_t = 0, U_t = 1) = 1 - \alpha
\end{aligned} \tag{6.45}$$

Thus, η_1 is the failure probability when the state is good (and no repair) and η_2 is the failure probability when the state is good (and when there is repair) with $\eta_1 > \eta_2$, and α is the success probability in the event of a repair.

The controller has access only to $\{0, 1\}$-valued measurement variables $Y_0, \cdots, Y_t$ and $U_0, \cdots, U_{t-1}$, at time t, where the measurements are generated by a binary symmetric channel:

$$(Y = x|X = x) = 1 - \epsilon, \qquad P(Y = 1 - x|X = x) = \epsilon,$$

for all $x \in \{0,$.

The per-stage cost function $c(x, u)$ is given by $c(0, 0) = C, c(1, 0) = 0, c(0, 1) = c(1, 1) = R$ with $0 < R < C$. Show that there exists an optimal control policy for both finite-horizon as well as infinite horizon discounted cost problems.

Exercise 6.7.4 (Zero-Delay Source Coding) *Let $\{x_t\}_{t\geq 0}$ be an $\mathbb{X}$-valued discrete-time Markov process where $\mathbb{X}$ can be a finite set or $\mathbb{R}^n$. Let there be an encoder which encodes (quantizes) the source samples and transmits the encoded versions to a receiver over a discrete noiseless channel with input and output alphabet $\mathcal{M} = \{1, 2, \ldots, M\}$, where M is a positive integer. The encoder policy γ is a sequence of functions $\{\kappa_t\}_{t\geq 0}$ with $\kappa_t : \mathcal{M}^t \times (\mathbb{X})^{t+1} \to \mathcal{M}$. At time t, the encoder transmits the $\mathcal{M}$-valued message*

$$q_t = \kappa_t(I_t)$$

with $I_0 = x_0$, $I_t = (q_{[0,t-1]}, x_{[0,t]})$ for $t \geq 1$, where. The collection of all such zero-delay encoders is called the set of admissible quantization policies and is denoted by Γ_A. A zero-delay receiver policy is a sequence of functions $\gamma^d = \{\gamma_t^d\}_{t\geq 0}$ of type $\gamma_t^d : \mathcal{M}^{t+1} \to \mathbb{U}$, where $\mathbb{U}$ denotes the finite reconstruction alphabet. Thus

$$u_t = \gamma_t^d(q_{[0,t]}), \qquad t \geq 0.$$

For the finite horizon setting the goal is to minimize the average cumulative cost (distortion)

$$J_{\pi_0}(\gamma, \gamma^d, T) = E_{\pi_0}^{\gamma,\gamma^d}\left[\frac{1}{T}\sum_{t=0}^{T-1} c_0(x_t, u_t)\right], \tag{6.46}$$

for some $T \geq 1$, where $c_0 : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is a nonnegative cost (distortion) function, and $E_{\pi_0}^{\gamma,\gamma^d}$ denotes expectation with initial distribution π_0 for x_0 and under the quantization policy γ and receiver policy γ^d.

a) Show that an optimal encoder uses a sufficient statistic, in particular, it uses $P(dx_t|q_{[0,t-1]})$ and the time information, for optimal performance.

b) Show that, when $\{x_t\}$ is i.i.d., any encoder and decoder pair can be replaced with one which only uses x_t, that is:

$$q_t = \kappa_t(x_t)$$

and the decoder only uses

$$u_t = \gamma_t^d(q_t), \qquad t \geq 0.$$

See [330], [321], [305] for finite sources and [343] for real sources and further relevant discussions, among many other recent references.

Exercise 6.7.5 *Let there be two decision makers, DM*1 *and DM*2*. Suppose that DMi (*$i = 1, 2$*) has access to:*

$$Y^i = X + V^i$$

where X, V^1, V^2 *are independent Gaussian random variables with unit variance and zero mean.*

a) Find $E[(X - E[X|Y^i])^2]$ *for* $i = 1, 2$*.*

*b) Suppose that DM*1 *and DM*2 *share their data* Y^1 *and* Y^2*. Find*

$$E[(X - E[X|Y^1, Y^2])^2]$$

*c) Suppose that DM*1 *and DM*2 *share with each other their estimates* $E[X|Y^i]$*. That is, DM*1 *has access to* Y^1 *and* $E[X|Y^2]$*; and DM*2 *has access to* Y^2 *and* $E[X|Y^1]$*. Find*

$$E\left[\left(X - E\left[X\middle|Y^1, E[X|Y^2]\right]\right)^2\right] \quad \text{and} \quad E\left[\left(X - E\left[X\middle|Y^2, E[X|Y^1]\right]\right)^2\right]$$

7

The Average Cost Problem

In this chapter, we consider the following average cost problem of finding

$$J^*_\infty(x) := \inf_\gamma J_\infty(x,\gamma) = \inf_{\gamma\in\Gamma_A} \limsup_{T\to\infty} \frac{1}{T} E^\gamma_x[\sum_{t=0}^{T-1} c(x_t,u_t)] \tag{7.1}$$

This is an important problem in applications where one is concerned about the long-term behaviour, unlike the discounted cost setup where the primary interest is in the short-term time stages.

For the study of the average cost problem, we will follow three distinct approaches; the first two will be based on the arrival at what we will call as the *average cost optimality equation*. The third approach will be based on the properties of expected (or sample path) occupation measures and their limit behaviours, leading to a linear program involving the space of probability measures. These approaches are related (e.g. via a dual optimization analysis [166, Chapter 12, p. 221], or a more direct stochastic analysis [161, Theorem 5.3]), however the conditions leading to solutions under these approaches are not identical, therefore, the corresponding conditions of existence and structural results for optimal policies are slightly different. As such, it will be instructive to study both approaches separately, as we do in the following.

7.1 Average Cost and the Average Cost Optimality Equation (ACOE) or Inequality (ACOI)

To study the average cost problem, one approach is to establish the existence of an Average Cost Optimality Equation (ACOE), and an associated verification theorem.

Definition 7.1.1 *The collection of functions $g:\mathbb{X}\to\mathbb{R}, h:\mathbb{X}\to\mathbb{R}, f:\mathbb{X}\to\mathbb{U}$ is a canonical triplet if for all $x\in\mathbb{X}$,*

$$g(x) = \inf_{u\in\mathbb{U}} \int g(x')\mathcal{T}(dx'|x,u)$$

$$g(x)+h(x) = \inf_{u\in\mathbb{U}} \left(c(x,u) + \int h(x')\mathcal{T}(dx'|x,u)\right)$$

with

$$g(x) = \int g(x')\mathcal{T}(dx'|x,f(x))$$

$$g(x)+h(x) = \left(c(x,f(x)) + \int h(x')\mathcal{T}(dx'|x,f(x))\right)$$

We will refer to these relations as the Average Cost Optimality Equation (ACOE).

Theorem 7.1.1 *[Verification Theorem] Let g, h, f be a canonical triplet. a) If g is a constant and $E_x^\gamma[|h(x_n)|] < \infty$ with*

$$\limsup_{n\to\infty} \frac{1}{n} E_x^\gamma[h(x_n)] = 0, \tag{7.2}$$

for all x and under every policy γ, then the stationary deterministic policy $\gamma^* = \{f, f, f, \cdots\}$ is optimal so that

$$g = J_\infty(x, \gamma^*) = \inf_{\gamma \in \Gamma_A} J_\infty(x, \gamma)$$

where

$$J(x, \gamma) = \limsup_{T\to\infty} \frac{1}{T} E_x^\gamma[\sum_{k=0}^{T-1} c(x_t, u_t)].$$

Furthermore, if $\limsup_{n\to\infty} \frac{1}{n}|E_x^\gamma[h(x_n)]| = 0$,

$$\lim_{n\to\infty} \left| \frac{1}{n} E_x^{\gamma^*} \sum_{t=1}^{n} [c(x_{t-1}, u_{t-1})] - g \right| \leq \limsup_{n\to\infty} \frac{1}{n} \left(|E_x^{\gamma^*}[h(x_n)] - h(x)| \right) = 0 \tag{7.3}$$

b) If g, considered above, is not a constant and depends on x, then under any policy γ

$$\limsup_{N\to\infty} \frac{1}{N} E_x^{\gamma^*}[\sum_{t=0}^{N-1} g(x_t)] \leq \inf_\gamma \limsup_{N\to\infty} \frac{1}{N} E_x^\gamma[\sum_{t=0}^{T-1} c(x_t, u_t)],$$

provided that (7.2) holds. Furthermore, $\gamma^* = \{f\}$ is optimal.

Proof: We prove (a); (b) follows from a similar reasoning. For any admissible policy γ,

$$\begin{aligned}
E^\gamma[h(x_t)|x_{[0,t-1]}, u_{[0,t-1]}] &= \int_y h(y) P(x_t \in dy | x_{t-1}, u_{t-1}) && (7.4)\\
&= c(x_{t-1}, u_{t-1}) + \int_y h(y) P(dy|x_{t-1}, u_{t-1}) - c(x_{t-1}, u_{t-1}) && (7.5)\\
&\geq \min_{u_{t-1}\in\mathbb{U}} \left(c(x_{t-1}, u_{t-1}) + \int_y h(y) P(dy|x_{t-1}, u_{t-1}) \right) - c(x_{t-1}, u_{t-1}) && (7.6)\\
&= g + h(x_{t-1}) - c(x_{t-1}, u_{t-1}) && (7.7)
\end{aligned}$$

Observe that with $h^M \to h$ a monotonically increasing sequence of bounded functions h^M which pointwise converges to h, we have that

$$\begin{aligned}
&E^\gamma[h(x_t)] = \lim_{M\to\infty} E^\gamma[h^M(x_t)]\\
&= \lim_{M\to\infty} E[E^\gamma[h^M(x_t)|x_{[0,t-1]}, u_{[0,t-1]}]] = E[\lim_{M\to\infty} E^\gamma[h^M(x_t)|x_{t-1}, u_{t-1}]]\\
&\geq E[g + h(x_{t-1}) - c(x_{t-1}, u_{t-1})] && (7.8)
\end{aligned}$$

Hence, for any admissible policy γ, $x \in \mathbb{X}$, by re-arranging the terms, we have that for all $n \in \mathbb{N}$

$$0 \leq \frac{1}{n} E_x^\gamma \left[\sum_{t=1}^{n} h(x_t) - g - h(x_{t-1}) + c(x_{t-1}, u_{t-1}) \right]$$

and thus

$$g \leq \frac{1}{n} E_x^\gamma[h(x_n)] - \frac{1}{n} E_x^\gamma[h(x_0)] + \frac{1}{n} E_x^\gamma[\sum_{t=1}^{n} c(x_{t-1}, u_{t-1})].$$

Taking the limit and using (7.2), we observe that g is a lower bound on the cost under any policy.

The above hold with equality if $\gamma^* = \{f\}$ is adopted since γ^* provides the pointwise minimum. Thus, equality holds under γ^* so that

$$g = \frac{E^{\gamma^*}[h(x_n)]}{n} - \frac{E_x^{\gamma^*}[h(x_0)]}{n} + \frac{1}{n}E_x^{\gamma^*}\sum_{t=1}^{n}[c(x_{t-1}, u_{t-1})].$$

Under (7.2),

$$g = \lim_{n\to\infty}\frac{1}{n}E_x^{\gamma^*}[\sum_{t=1}^{n}c(x_{t-1}, u_{t-1})].$$

and

$$\left|\frac{1}{n}E_x^{\gamma^*}[\sum_{t=1}^{n}c(x_{t-1}, u_{t-1})] - g\right| \leq \frac{1}{n}\Big(|E_x^{\gamma^*}[h(x_n)]| + |h(x)|\Big) \to 0,$$

as $n \to \infty$ ◇

Theorem 7.1.2 *[Optimality Through Finite Horizon Limits] If $\gamma^* = \{f, f, f, \cdots\}$ is so that*

$$J_\infty(x, \gamma^*) = \limsup_{T\to\infty}\inf_{\gamma\in\Gamma_A} J^T(x, \gamma)$$

with

$$J^T(x, \gamma) = \frac{1}{T}E_x^{\gamma}[\sum_{t=0}^{T-1}c(x_t, u_t)], \tag{7.9}$$

then γ^* is optimal.

Proof. The proof follows from the observation (by Lemma 5.5.1)

$$J_\infty^*(x) \geq \limsup_{T\to\infty}\inf_{\gamma\in\Gamma_A} J^T(x, \gamma),$$

and that γ^* achieves this lower bound. ◇

Remark 7.1. Recall that we had utilized the argument used in the proof of Theorem 7.1.2 while studying average cost LQG problems in Theorem 5.12)(iii).

Remark 7.2. Note that if we have that, in Theorem 7.1.1,

$$g = \limsup_{T\to\infty}\inf_{\gamma\in\Gamma_A} J^T(x, \gamma),$$

then it suffices to check (7.2) only for the policy γ^* to certify its optimality as this would ensure that the condition in Theorem 7.1.2, on the achievability of g by $g = J(x, \gamma^*)$, is attained. Note the analogy with Lemma 5.5.4.

Definition 7.1.2 *Let g be a constant and $h : \mathbb{X} \to \mathbb{R}$, $f : \mathbb{X} \to \mathbb{U}$ be so that for all $x \in \mathbb{X}$,*

$$g + h(x) \geq \inf_{u\in\mathbb{U}}\left(c(x, u) + \int h(x')\mathcal{T}(dx'|x, u)\right) \tag{7.10}$$

Alternatively, let

$$g + h(x) \leq \inf_{u\in\mathbb{U}}\left(c(x, u) + \int h(x')\mathcal{T}(dx'|x, u)\right) \tag{7.11}$$

with, in either case

$$\inf_{u\in\mathbb{U}}\left(c(x,u)+\int h(x')\mathcal{T}(dx'|x,u)\right)=\left(c(x,f(x))+\int h(x')\mathcal{T}(dx'|x,f(x))\right)$$

We will refer to (7.10) as the Average Cost Optimality Inequality (ACOI).

See, e.g., [14, Theorem 6.6] for the following:

Theorem 7.1.3 *[Verification Theorem]*

(i) Let (7.11) hold. If

$$\limsup_{n\to\infty}\frac{1}{n}E_x^{\gamma}[h(x_n)]\leq 0, \tag{7.12}$$

for all x and under every policy γ. Then g is a lower bound under any policy.

(ii) On the other hand, with (7.10) so that

$$g+h(x)\geq c(x,f(x))+\int h(x')\mathcal{T}(dx'|x,f(x))$$

and

$$\liminf_{n\to\infty}\frac{1}{n}E_x^{\gamma^*}[h(x_n)]\geq 0, \tag{7.13}$$

holding with $\gamma^*=\{f,f,f,\cdots\}$. Then the stationary deterministic policy $\gamma^*=\{f,f,f,\cdots\}$ satisfies

$$g\geq J(x,\gamma^*)$$

Proof: For (i): for any policy γ,

$$\begin{aligned}
E^{\gamma}[h(x_t)|x_{[0,t-1]},u_{[0,t-1]}] &= \int_y h(y)P(x_t\in dy|x_{t-1},u_{t-1}) && (7.14)\\
&= c(x_{t-1},u_{t-1})+\int_y h(y)P(dy|x_{t-1},u_{t-1})-c(x_{t-1},u_{t-1}) && (7.15)\\
&\geq \min_{u_{t-1}\in\mathbb{U}}\left(c(x_{t-1},u_{t-1})+\int_y h(y)P(dy|x_{t-1},u_{t-1})\right)-c(x_{t-1},u_{t-1}) && (7.16)\\
&\geq g+h(x_{t-1})-c(x_{t-1},u_{t-1}), && (7.17)
\end{aligned}$$

where the last inequality is due to (7.11).

As in (7.8), we have that

$$E^{\gamma}[h(x_t)]\geq E[g+h(x_{t-1})-c(x_{t-1},u_{t-1})]$$

Hence, for any policy γ

$$0\leq\frac{1}{n}E_x^{\gamma}\sum_{t=1}^{n}[h(x_t)-g-h(x_{t-1})+c(x_{t-1},u_{t-1})]$$

and

$$g\leq\frac{1}{n}E_x^{\gamma}[h(x_n)]-\frac{1}{n}E_x^{\gamma}[h(x_0)]+\frac{1}{n}E_x^{\gamma}[\sum_{t=1}^{n}c(x_{t-1},u_{t-1})].$$

Taking the limit, we observe that g is a lower bound on the cost under any policy under (7.13).

For (ii): if we start the analysis above leading to (7.17) with γ^*, we have

$$E_x^{\gamma^*}\left[\sum_{t=1}^{n} h(x_t) - E^{\gamma^*}[h(x_t)|x_{[0,t-1]}, u_{[0,t-1]}]\right] = 0$$

and

$$E^{\gamma^*}[h(x_t)|x_{[0,t-1]}, u_{[0,t-1]}] = \int_y h(y)P(x_t \in dy|x_{t-1}, u_{t-1}) \tag{7.18}$$

$$= c(x_{t-1}, f(x_{t-1})) + \int_y h(y)P(dy|x_{t-1}, f(x_{t-1})) - c(x_{t-1}, f(x_{t-1})) \tag{7.19}$$

$$\leq g + h(x_{t-1}) - c(x_{t-1}, f(x_{t-1})) \tag{7.20}$$

Iterating the above and dividing by n, we arrive at

$$g - \frac{1}{n}E_x^{\gamma^*}[h(x_n)] + \frac{1}{n}E_x^{\gamma^*}[h(x_0)] \geq \frac{1}{n}E_x^{\gamma^*}[\sum_{t=1}^{n} c(x_{t-1}, u_{t-1})].$$

Taking the limsup on both sides (and replacing $\limsup$ with $\liminf$ by reversing the negative sign on the left), and (7.13) holding for $\gamma^* = \{f, f, f, \cdots\}$, we establish the desired bound. ◇

7.2 The Value Iteration and Contraction Approach to the Average Cost Problem

Fix $z \in \mathbb{X}$ and consider the space of measurable and bounded functions h with the restriction that $h(z) = 0$. Let (g, h, f) be a canonical triplet with $g \equiv \rho \in \mathbb{R}$ so that

$$\rho + h(x) = \inf_{u \in \mathbb{U}} \left(c(x, u) + \int h(x')\mathcal{T}(dx'|x, u) \right)$$

7.2.1 Contraction under the span semi-norm

Consider the following assumption.

Assumption 7.2.1 *For some $\alpha \in [0, 1)$, and for all $x, x' \in \mathbb{X}$ and $u, u' \in \mathbb{U}$*

$$\|P(\cdot|x, u) - P(\cdot|x', u')\|_{TV} \leq 2\alpha$$

A sufficient condition for the above is the following minorization condition.

Assumption 7.2.2 *There exists a positive measure μ' with $\mathcal{T}(B|x, u) \geq \mu'(B)$, for all $B \in \mathcal{B}(\mathbb{X})$ and all $(x, u) \in \mathbb{X} \times \mathbb{U}$.*

Observe that Assumption 7.2.1 is weaker than Assumption 7.2.2.

Consider the following *span* semi-norm:

$$\|u\|_{sp} = \sup_x u(x) - \inf_x u(x)$$

The space of measurable bounded functions that satisfy $h(z) = 0$ under the semi-norm $\|u\|_{sp}$ is a Banach space (and hence the semi-norm becomes a norm in this space since $\|u\|_{sp} = 0$ implies $u \equiv 0$).

Define

$$\mathbb{T}(h)(x) = \inf_{u \in \mathbb{U}} \left(c(x,u) + \int h(x') \mathcal{T}(dx'|x,u) \right) \tag{7.21}$$

Let

$$(\mathbb{T}_z(h))(x) = (\mathbb{T}(h))(x) - (\mathbb{T}(h))(z)$$

Note that $\mathbb{T}_z$ maps the aforementioned Banach space to itself under the measurable selection conditions reviewed in *Chapter 5*. Under Assumption 7.2.2, and the measurable selection conditions reviewed in *Chapter 5*, we will show (through similar steps as those in *Chapter 5*) that the map is a contraction:

First note that for pairs (x,u) and (x',u'), with $\mu(dx_1) := P(dx_1|x,u) - P(dx_1|x',u')$ defining a signed measure, by the Jordan-Hahn decomposition theorem [181, Theorem 2.8] there exists A with $\mu(A) = -\mu(A^c) \geq 0$ so that the restriction of μ to A (i.e., $\mu_A(B) := \mu(B \cap A)$ for every Borel B) defines a non-negative measure and the restriction of $-\mu$ to A^c defines a non-negative measure with $\mu(A) - \mu(A^c) = \|\mu\|_{TV} \leq 2\alpha$ and thus $\mu(A) \leq \alpha$. Thus, for any x, x', u, u;

$$\begin{aligned}
&\int h(x_1)P(dx_1|x,u) - h(x_1)P(dx_1|x',u') \\
&= \int_A h(x_1)(P(dx_1|x,u) - P(dx_1|x',u')) + \int_{A^c} h(x_1)(P(dx_1|x,u) - P(dx_1|x',u')) \\
&= \int_A h(x_1)(P(dx_1|x,u) - P(dx_1|x',u')) - \int_{A^c} h(x_1)(P(dx_1|x',u') - P(dx_1|x,u)) \\
&\leq \int_A \left(\sup_{x_1} h(x_1) \right)(P(dx_1|x,u) - P(dx_1|x',u')) - \int_{A^c} \left(\inf_{x_1} h(x_1) \right)(P(dx_1|x',u') - P(dx_1|x,u)) \\
&\leq \left(\sup_{x_1} h(x_1) \right)\mu(A) - \left(\inf_{x_1} h(x_1) \right)((-\mu)(A^c)) \\
&\leq \alpha \|h\|_{sp}
\end{aligned}$$

Then, note that for v_1 and v_2 bounded and with $\mathbb{T}(v_i)(x)$ achieved with control u_i^x at x, we have that for any x, x'

$$\begin{aligned}
&\left((\mathbb{T}(v_1))(x) - (\mathbb{T}(v_2))(x) \right) - \left((\mathbb{T}(v_1))(x') - (\mathbb{T}(v_2))(x') \right) \\
&\leq \int (v_1(x_1) - v_2(x_1)) \left(P(dx_1|x,u_2^x) - P(dx_1|x',u_1^{x'}) \right) \leq \alpha \|v_1 - v_2\|_{sp},
\end{aligned}$$

(where for the first term $(\mathbb{T}(v_1))(x) - (\mathbb{T}(v_2))(x)$we upper bound the difference by applying u_2^x for the right hand side of (7.21) involving v_1, and for the second term $(\mathbb{T}(v_1))(x') - (\mathbb{T}(v_2))(x')$ we apply $u_1^{x'}$ for bounding $(\mathbb{T}(v_2))(x'))$ and thus, since x, x' are arbitrary, we have

$$\|\mathbb{T}(v_1) - \mathbb{T}(v_2)\|_{sp} \leq \alpha \|v_1 - v_2\|_{sp}.$$

Furthermore, since $(\mathbb{T}(v_i))(z)$ only serves as a shift term in the following, we have that

$$\|\mathbb{T}_z(v_1) - \mathbb{T}_z(v_2)\|_{sp} = \|\mathbb{T}(v_1) - \mathbb{T}(v_2)\|_{sp} \leq \alpha \|v_1 - v_2\|_{sp}.$$

We can then state the following.

Theorem 7.2.1 *[162, Lemma 3.5] The iterations*

$$h_{n+1} = \mathbb{T}_z(h_n),$$

with $h_0 \equiv 0$ converges to a fixed point: $\mathbb{T}_z(h) = h$, which leads to the ACOE triplet in Definition 7.1.1. In particular, if the cost is bounded, under Assumption 7.2.1, and the controlled kernel satisfies the measurable selection conditions given in Assumption 5.2.1 or 5.2.2, there exists a solution to the ACOE, which in turn leads to an optimal policy.

7.2.2 Contraction under sup norm via minorization by equivalence with a discounted cost problem

We now present a more direct approach. Under the slightly stronger Assumption 7.2.2, we have that with

$$\mathcal{T}'(\cdot|x,u) = \mathcal{T}(\cdot|x,u) - \mu'(\cdot)$$

a positive measure, the map

$$(\mathbb{T}'(h))(x) = \min_{u\in U}\left(c(x,u) + \int h(x_1)\mathcal{T}'(dx_1|x,u)\right)$$

is a contraction (see [162, p.61] for a historical review on this approach). With this approach, one can avoid the use of the span semi-norm approach. Accordingly, one can apply the standard value iteration algorithm using $\mathbb{T}'$, following the proof of Theorem 5.5.2. The limit equation

$$h(x) = \min_{u\in U}\left(c(x,u) + \int h(x_1)\mathcal{T}'(dx_1|x,u)\right) = \min_{u\in U}\left(c(x,u) + \int h(x_1)\mathcal{T}(dx_1|x,u)\right) - \int h(x_1)\mu'(dx_1)$$

is the desired ACOE in Definition 7.1.1 with $g \equiv \int h(x_1)\mu'(dx_1)$.

7.3 The Vanishing Discounted Cost Approach to the Average Cost Problem

7.3.1 Finite state and action spaces

Average cost emphasizes the asymptotic values of the cost function whereas the discounted cost emphasizes the short-term cost functions. However, under technical restrictions, one can show that the limit as the discounted factor converges to 1, one can obtain a solution for the average cost optimization. We now state one such condition below.

Theorem 7.3.1 *[102] [47] [14, Theorem 4.3] Consider a controlled Markov chain where the state and action spaces are finite, and suppose that under any stationary and deterministic policy the entire state space is a recurrent set. Let*

$$J_\beta(x) = \inf_{\gamma\in\Gamma_A} J_\beta(x,\gamma) = \inf_{\gamma\in\Gamma_A} E_x^\gamma[\sum_{t=0}^{\infty}\beta^t c(x_t,u_t)]$$

and suppose that γ_n^* is an optimal deterministic policy for $J_{\beta_n}(x)$. Then, there exists some $\gamma^* \in \Gamma_{SD}$ which is optimal for every β sufficiently close to 1, and is also optimal for the average cost

$$J(x) = \inf_{\gamma\in\Gamma_A} \limsup_{T\to\infty} \frac{1}{T} E_x^\gamma[\sum_{t=0}^{T-1} c(x_t,u_t)]$$

Proof. First note that for every stationary and deterministic policy f, by a slight change in notation from what was considered earlier in the notes, $J_\beta(x,f) := (1-\beta)E_x^f[\sum_{k=0}^{\infty}\beta^k c(x,f(x_k))]$ is a continuous function on $[0,1]$ (in β). Let $\beta_n \uparrow 1$. For each β_n, J_{β_n} is achieved by a stationary and deterministic policy. Since there are only finitely many such policies, there exists at least one policy f^* which is optimal for infinitely many β_n; call such a sequence β_{n_k}. We will show that this policy is optimal for the average cost problem also.

It follows that $(1-\beta_{n_k})J_{\beta_{n_k}}(x,f^*) \le (1-\beta_{n_k})J_{\beta_{n_k}}(x,\gamma)$ for all γ. Then, infinitely often for every deterministic stationary policy f:

$$(1-\beta_{n_k})J_{\beta_{n_k}}(x,f^*) - (1-\beta_{n_k})J_{\beta_{n_k}}(x,f) \le 0$$

We now claim that for some $\beta^* < 1$, $J_\beta(x,f^*) \le J_\beta(x,\gamma)$ for all $\beta \in (\beta^*,1)$. The function $(1-\beta_{n_k})J_{\beta_{n_k}}(x,f^*) - (1-\beta_{n_k})J_{\beta_{n_k}}(x,f)$ is continuous in β and uniformly bounded, therefore if the claim were not correct, the function must have infinitely many zeros. On the other hand, one can write the equation

$$J_\beta(x,f) = c(x,f) + \beta \sum_{x'} P(x'|x,f(x)) J_\beta(x',f)$$

in matrix form to obtain $J_\beta(\cdot,f) = (I - \beta P(\cdots|\cdot,f(\cdot)))^{-1} c(\cdot,f(\cdot))$. It follows that, $(1-z)(J_z(x,f^*) - J_z(x,f))$ is a rational function (that is, ratio of two polynomials with finite order) on the open unit disk (in the complex region) $|z| < 1$, such a function can only have finitely many zeros (unless it is identically zero): this follows by studying the inverse matrix $(I - zP)^{-1}$ which is analytic inside the unit disk and if it is non-zero on the boundary of the unit disk at $z = 1$, it has to be bounded away from zero in a neighborhood of $z = 1$ inside the unit disk. Therefore, it must be that for some $\beta^* < 1$, $J_\beta(x,f^*) \leq J_\beta(x,\gamma)$ for all $\beta \in (\beta^*, 1)$. We note here that such a policy is called a *Blackwell-Optimal Policy*. Now,

$$(1-\beta_{n_k}) J_{\beta_{n_k}}(x,f^*) \leq (1-\beta_{n_k}) J_{\beta_{n_k}}(x,\gamma) \tag{7.22}$$

for any γ and thus,

$$\begin{aligned} J(x,f^*) &= \liminf_{T\to\infty} \frac{1}{T} E_x^{f^*}[\sum_{k=0}^{T-1} c(x_k,u_k)] \leq \liminf_{n_k\to\infty} (1-\beta_{n_k}) J_{\beta_{n_k}}(x,f^*) = \limsup_{n_k\to\infty} (1-\beta_{n_k}) J_{\beta_{n_k}}(x,f^*) \\ &\leq \limsup_{n_k\to\infty} (1-\beta_{n_k}) J_{\beta_{n_k}}(x,\gamma) \leq \limsup_{T\to\infty} \frac{1}{T} E_x^{\gamma}[\sum_{k=0}^{T-1} c(x_k,u_k)] \end{aligned} \tag{7.23}$$

In the first equality, we use the fact that the limit exists. In the above, the sequence of inequalities follow from the following *Abelian* inequalities (see [165, Lemma 5.3.1]): Let a_n be a sequence of non-negative numbers and $\beta \in (0,1)$. Then,

$$\begin{aligned} &\liminf_{N\to\infty} \frac{1}{N} \sum_{m=0}^{N-1} a_m \leq \liminf_{\beta\uparrow 1} (1-\beta) \sum_{m=0}^{\infty} \beta^m a_m \\ &\leq \limsup_{\beta\uparrow 1} (1-\beta) \sum_{m=0}^{\infty} \beta^m a_m \leq \limsup_{N\to\infty} \frac{1}{N} \sum_{m=0}^{N-1} a_m \end{aligned} \tag{7.24}$$

As a result, f^* is optimal. The optimal cost does not depend on the initial state by the recurrence condition and irreducibility of the chain under the optimal policy. ◇

In the following, we consider more general state spaces and generalize the results presented above.

7.3.2 Standard Borel state and action spaces, ACOE and ACOI

Consider the value function for a discounted cost problem as discussed in Section 5.5:

$$J_\beta(x) = \min_{u\in\mathbb{U}} \left\{ c(x,u) + \beta \int_{\mathbb{X}} J_\beta(y) \mathcal{T}(dy|x,u) \right\}, \quad x \in \mathbb{X}. \tag{7.25}$$

Let x_0 be an arbitrary state and for all $x \in \mathbb{X}$ consider

$$\begin{aligned} &J_\beta(x) - J_\beta(x_0) \\ &= \min_{u\in\mathbb{U}} \left(c(x,u) + \beta \int \mathcal{T}(dx'|x,u)(J_\beta(x') - J_\beta(x_0)) - (1-\beta) J_\beta(x_0) \right) \end{aligned}$$

As discussed in Section 5.5, this has a solution for every $\beta \in (0,1)$ under measurable selection conditions.

Arriving at the Average Cost Optimality Equation. We recall that a family of functions F mapping a metric space $\mathbb{S}$ to $\mathbb{R}$ is said to be *equicontinuous at a point* $x_0 \in \mathbb{S}$ if, for every $\epsilon > 0$, there exists a $\delta > 0$ such that $d(x,x_0) \leq \delta \implies |f(x) - f(x_0)| \leq \epsilon$ for all $f \in F$. The family F is said to be *equicontinuous* if it is equicontinuous at each $x \in \mathbb{S}$.

Now suppose that $h_\beta(x) := J_\beta(x) - J_\beta(x_0)$ is equicontinuous (over β) and $\mathbb{X}$ is compact. By the Arzela-Ascoli Theorem (Theorem 7.3.2), taking $\beta \uparrow 1$ along some sequence, for some subsequence, $J_{\beta_{n_k}}(x) - J_{\beta_{n_k}}(x_0) \to \eta(x)$ uniformly. If the cost is bounded, then, along a further subsequence,

$$(1 - \beta_{n_k}) J_{\beta_{n_k}}(x_0) \to \zeta^* \tag{7.26}$$

for some ζ^* (which is to be shown to be independent of x_0). If we could also exchange the order of the minimum and the limit, one obtains the Average Cost Optimality Equation (ACOE):

$$\eta(x) = \min_{u \in \mathbb{U}} \left(c(x,u) + \int \mathcal{T}(dx'|x_t, u_t)\eta(x') - \zeta^* \right), \tag{7.27}$$

which has the form of the equations in Definition 7.1.1.

We now make this observation formal (and relax the compactness assumption on the state space).

Assumption 7.3.1

(a) The one stage cost function c is bounded and continuous.

(b) The stochastic kernel $\mathcal{T}(\,\cdot\,|x,u)$ is weakly continuous in $(x,u) \in \mathbb{X} \times \mathbb{U}$, i.e., if $(x_k, u_k) \to (x,u)$, then $\mathcal{T}(\,\cdot\,|x_k,u_k) \to \mathcal{T}(\,\cdot\,|x,u)$ weakly.

(c) $\mathbb{U}$ is compact.

(d) $\mathbb{X}$ is σ-compact, that is, $\mathbb{X} = \cup_n S_n$ where $S_n \subset S_{n+1}$ and each S_n is compact.

In addition to Assumption 7.3.1, we impose the following assumption in this section.

Assumption 7.3.2
There exists $\alpha \in (0,1)$ and $N \geq 0$, and a state $z_0 \in \mathbb{X}$ such that,

(e) $-N \leq h_\beta(z) \leq N$ for all $z \in \mathbb{X}$ and $\beta \in [\alpha, 1)$, where

$$h_\beta(z) = J_\beta(z) - J_\beta(z_0),$$

for some fixed $z_0 \in \mathbb{X}$.

(f) The sequence $\{h_{\beta(k)}\}$ is equicontinuous, where $\{\beta(k)\}$ is a sequence of discount factors converging to 1 which satisfies $\lim_{k\to\infty}(1-\beta(k))J^*_{\beta(k)}(z) = \rho^*$ for all $z \in \mathbb{X}$ for some $\rho^* \in [0, L]$.

Note that when the one stage cost function c is bounded by some $L \in \mathbb{R}_+$, we must have

$$|(1-\beta)J^*_\beta(z)| \leq L$$

for all $\beta \in (0,1)$ and $z \in \mathbb{X}$. Let us recall the Arzela-Ascoli theorem.

Theorem 7.3.2 *[111] Let F be an equicontinuous family of functions on a compact space $\mathbb{X}$ and let h_n be a sequence in F such that the range of f_n is compact. Then, there exists a subsequence h_{n_k} which converges uniformly to a continuous function. If $\mathbb{X}$ is σ-compact, that is $\mathbb{X} = \cup_n K_n$ with $K_n \subset K_{n+1}$ with K_n compact, the same result holds where h_{n_k} converges pointwise to a continuous function, and and the convergence is uniform on compact subsets of $\mathbb{X}$.*

Theorem 7.3.3 *Under Assumptions 7.3.1 and 7.3.2, there exist a constant $\rho^* \geq 0$, a continuous and bounded h from $\mathbb{X}$ to $\mathbb{R}$ with $-N \leq h(\,\cdot\,) \leq N$, and $\{f^*\} \in \Gamma_S$ such that (ρ^*, h, f^*) satisfies the ACOE; that is,*

$$\begin{aligned} \rho^* + h(z) &= \min_{u \in \mathbb{U}} (c(z,u) + \int_{\mathbb{X}} h(y)\mathcal{T}(dy|z,u)) \\ &= c(z, f^*(z)) + \int_{\mathbb{X}} h(y)\mathcal{T}(dy|z, f^*(z)), \end{aligned}$$

for all $z \in \mathbb{X}$. Moreover, $\{f^*\}$ is optimal and ρ^* is the value function, i.e.,

$$\inf_{\varphi} J(\varphi, z) =: J^*(z) = J(\{f^*\}, z) = \rho^*,$$

for all $z \in \mathbb{X}$.

Proof. By (7.26), we have that $(1-\beta_{n_k})J_\beta(x_0) \to \rho^*$ for some subsequence n_k as $\beta_{n_k} \uparrow 1$ and some ρ^*. Observe that for any $x \in \mathbb{X}$

$$(1-\beta_{n_k})J_{\beta_{n_k}}(x) = (1-\beta_{n_k})(J_{\beta_{n_k}}(x) - J_{\beta_{n_k}}(x_0)) + (1-\beta_{n_k})J_{\beta_{n_k}}(x_0),$$

which, by the uniform boundedness of $J_{\beta_{n_k}}(x) - J_{\beta_{n_k}}(x_0)$, implies that the limit ρ^* does not depend on x. By Assumption 7.3.2-(f) and Theorem 7.3.2, there exists a further subsequence of h_{n_k}, $\{h_{\beta(k_l)}\}$, which converges (uniformly on compact sets) to a continuous and bounded function h. Take the limit in (7.36) along this subsequence, i.e., consider

$$\begin{aligned}
\rho^* + h(z) &= \lim_l \min_{\mathbb{U}} [c(z,u) + \beta(k_l) \int_{\mathbb{X}} h_{\beta(k_l)}(y) \mathcal{T}(dy|z,u)] \\
&= \min_{\mathbb{U}} \lim_l [c(z,u) + \beta(k_l) \int_{\mathbb{X}} h_{\beta(k_l)}(y) \mathcal{T}(dy|z,u)] \\
&= \min_{\mathbb{U}} [c(z,u) + \int_{\mathbb{X}} h(y) \mathcal{T}(dy|z,u)].
\end{aligned}$$

Here, somewhat similar to Lemma 5.2.2, the exchange of limit and minimum follows from writing (using the compactness of $\mathbb{U}$, the continuity of $[c(z,u) + \beta(k_l) \int_{\mathbb{X}} h_{\beta(k_l)}(y) \mathcal{T}(dy|z,u)]$ on $\mathbb{U}$, and the equicontinuity of $\{h_{\beta(k)}\}$):

$$\min_{\mathbb{U}} [c(z,u) + \beta(k_l) \int_{\mathbb{X}} h_{\beta(k_l)}(y) \mathcal{T}(dy|z,u)] = c(z,u_l) + \beta(k_l) \int_{\mathbb{X}} h_{\beta(k_l)}(y) \mathcal{T}(dy|z,u_l)$$

$$\min_{\mathbb{U}} [c(z,u) + \int_{\mathbb{X}} h(y) \mathcal{T}(dy|z,u)] = c(z,u^*) + \int_{\mathbb{X}} h(y) \mathcal{T}(dy|z,u^*)$$

and showing that

$$\max \left(\left| \int_{\mathbb{X}} (\beta(k_l) h_{\beta(k_l)}(y) - h(y)) \mathcal{T}(dy|z,u_l) \right|, \left| \int_{\mathbb{X}} (\beta(k_l) h_{\beta(k_l)}(y) - h(y)) \mathcal{T}(dy|z,u^*) \right| \right) \to 0. \tag{7.28}$$

The last item follows from a contrapositive argument. Suppose that the term does not converge to zero, implying that for some subsequence it remains above some $\epsilon > 0$ As in the proof of Lemma 5.2.2, for any such subsequence we would have the following. By compactness of the action space $\mathbb{U}$, we would have a further subsequence so that $u_n \to u$ (ignoring the subscripts) for some u along this further subsequence. By weak continuity of the kernel Under Assumption 7.3.1, we then have that $\mathcal{T}(dy|z,u_n) \to \mathcal{T}(dy|z,u)$. Since $h_{\beta(k_l)}$ converges uniformly on compact sets, convergence to zero follows from Theorem D.3.1(i), concluding the argument (a more direct argument would be as follows: Since for every $\{u_n \to u\}$, the set of probability measures $\mathcal{T}(dy|z,u_n)$ is *tight*, for every $\epsilon > 0$ (by weak continuity in Assumption 7.3.1), one can find a compact set $K_n \subset \mathbb{X}$ so that $\int_{\mathbb{X} \setminus K_n} h_{\beta(k_l)}(y) \mathcal{T}(dy|z,u)] \leq \epsilon$ (here, by Assumption 7.3.2-(e), uniform boundedness of $h_{\beta(k_l)}$ is critical). Since on K_n, $h_{\beta(k_l)} \to h$ uniformly and h is bounded, the result follows). ◇

Remark 7.3. One can also consider (the slightly stronger condition of) Assumptions 4.2.1 and 5.5.1 of [165]; see e.g. [165, Theorem 5.5.4])); see also [285, Theorem 3.8]. Further conditions also appear in the literature; see Hernandez-Lerma and Lasserre [166] for a detailed analysis for the unbounded cost setup, and [91] for such results and a detailed literature review. Further conditions are available in [147] [317], among other references.

Corollary 7.3.1 (Beyond Minorization) *[100, Lemma 2.4] Consider Assumption 5.5.1 with $K_2 < 1$ and that $\mathbb{U}$ and $\mathbb{X}$ are compact. Then, Theorem 7.3.3 is applicable.*

Proof. The proof follows since the equicontinuity condition is satisfied in view of (5.38) (see [270, Theorem 4.37]) for all $\beta \in (0,1]$. ◇

For an explicit proof, see [100, Lemma 2.4 and Lemma 2.5]. The above is particularly useful for belief-MDPs; see Theorem 6.3.8. If compactness does not hold in Theorem 7.3.1 but $\mathbb{X}$ is σ-compact, then by Lipschitz regularity for each compact restriction, uniform boundedness would apply and convergence of h_β along a subsequence to a limit, uniformly on compact sets, would apply. To account for the non-boundedness of the limit h, an integrability condition leading to (7.28) would be sufficient to arrive at the optimality equation.

In the following, we obtain two partial generalizations of Theorem 7.3.1 to the standard Borel space setup: First, we observe that under the conditions of Theorem 7.3.3, since a solution to ACOE exists; every subsequential limit in (7.26) will need to be identical. This leads to the following.

Theorem 7.3.4 *Under the conditions of Theorem 7.3.3, (7.26) can be refined to*

$$\lim_{\beta\uparrow 1}(1-\beta)J_\beta(x_0) \to \zeta^*, \tag{7.29}$$

where J_β is defined in (7.25) and ζ^* is the optimal average cost. That is, we will have sequential convergence (and not just subsequential convergence).

The argument is via contraposition; suppose the limit of two subsequences were different. Each subsequence would then have further subsequences which would satisfy the ACOE and via the verification theorem would lead to the (same) optimal cost ζ^*. Hence the limits must be identical.

As a note, there is no claim that the limit of the relative value functions, h, is unique. If minorization holds (Assumption 7.2.2) then indeed h is unique up to a constant; in the absence of minorization, we do not have such a claim; see also [353].

We showed in Theorem 7.3.4 that the value functions of discounted cost criteria converges to the value under the average cost criterion under the conditions of Theorem 7.3.3. In practice, we would like to also see whether the policies solving the discounted cost problem (or that are at least near optimal for the discounted cost problem) are (also) near optimal for the average cost criterion. Such a result has significant implications for numerical and reinforcement learning theoretic methods (e.g, [92, Theorem 5]).

Theorem 7.3.5 *Let Assumptions 7.3.1 and 7.3.2 hold. Let γ_β solve the discounted cost optimality equation (7.25). Then, for every $\epsilon > 0$, there exists β large enough such that*

$$J(x,\gamma_\beta) - \zeta^* < \epsilon,$$

where ζ^* is the optimal average cost. That is, the discounted cost optimal policy is near-optimal for the average cost criterion.

Proof. Write

$$\begin{aligned}
&h_\beta(x)\\
&= \min_{u\in\mathbb{U}}\left(c(x,u) + \beta\int \mathcal{T}(dx'|x,u)(h_\beta(x')) - (1-\beta)J_\beta(x_0)\right)\\
&= \left(c(x,\gamma_\beta(x)) + \beta\int \mathcal{T}(dx'|x,\gamma_\beta(x))(h_\beta(x')) - (1-\beta)J_\beta(x_0)\right)
\end{aligned} \tag{7.30}$$

It follows that

$$\begin{aligned}
&h_\beta(x) + \zeta^* - (\zeta^* - (1-\beta)J_\beta(x_0)) + (1-\beta)\int \mathcal{T}(dx'|x,\gamma_\beta(x))(h_\beta(x'))\\
&= \left(c(x,\gamma_\beta(x)) + \int \mathcal{T}(dx'|x,\gamma_\beta(x))(h_\beta(x'))\right)
\end{aligned} \tag{7.31}$$

We have by assumption that h_β is uniformly bounded. Now, take β sufficiently close to 1 so that $|\zeta^* - (1-\beta)J_\beta(x_0)| \leq \epsilon$ by (7.29), and that $|(1-\beta)\int \mathcal{T}(dx'|x, \gamma_\beta(x))(h_\beta(x'))| \leq \epsilon$. Then, we have

$$\begin{aligned} & h_\beta(x) + \zeta^* + 2\epsilon \\ & \geq \left(c(x, \gamma_\beta(x)) + \int \mathcal{T}(dx'|x, \gamma_\beta(x))(h_\beta(x')) \right) \end{aligned} \tag{7.32}$$

Via Theorem 7.1.3(ii), since h_β is bounded, the above implies that γ_β achieves an average cost not larger than $\zeta^* + 2\epsilon$. $\diamond$

We now establish near optimality of near optimal discounted solutions for the average cost setup.

Theorem 7.3.6 *Let Assumptions 7.3.1 and 7.3.2 hold. Let β_ϵ be taken as in Theorem 7.3.5, be such that*

$$|\zeta^* - (1-\beta_\epsilon)J_{\beta_\epsilon}(x_0)| \leq \frac{\epsilon}{2}$$

and

$$(1-\beta_\epsilon)\|h_{\beta_\epsilon}\|_\infty \leq \frac{\epsilon}{2},$$

so that γ_{β_ϵ} is ϵ-optimal. Suppose that $\gamma^\delta_{\beta_\epsilon}$ is such that

$$J_{\beta_\epsilon}(x, \gamma^\delta_{\beta_\epsilon}) - J_{\beta_\epsilon}(x) < \delta.$$

Then,

$$J(x, \gamma^\delta_{\beta_\epsilon}) - \zeta^* < \epsilon + \delta$$

That is, the near optimal discounted cost optimal policy is near-optimal for the average cost criterion as well.

Proof. We follow the proof of Theorem 7.3.5 with a minor variation as follows: Let us write the cost attained by the policy $\gamma^\delta_{\beta_\epsilon}$ as:

$$J_{\beta_\epsilon}(x, \gamma^\delta_{\beta_\epsilon}) = c(x, \gamma^\delta_{\beta_\epsilon}) + \beta_\epsilon E[J_{\beta_\epsilon}(x_1, \gamma^\delta_{\beta_\epsilon})|x_0 = x, u_0 = \gamma^\delta_{\beta_\epsilon}(x)]$$

This follows from the fact that the cost is bounded and the arguments used in *Chapter* 5 on the verification theorem; alternatively see Section 8.1.2.

With x_0 as in the proof of Theorem 7.3.5, write

$$\bar{h}_{\beta_\epsilon}(x) := J_{\beta_\epsilon}(x, \gamma^\delta_{\beta_\epsilon}) - J_{\beta_\epsilon}(x_0)$$

Then, we have that, by substitution,

$$\bar{h}_{\beta_\epsilon}(x) = c(x, \gamma_{\beta^\delta_\epsilon}) + \beta_\epsilon E[\bar{h}_{\beta_\epsilon}(x_1)|x_0 = x, u_0 = \gamma^\delta_{\beta_\epsilon}(x)] - (1-\beta_\epsilon)J_{\beta_\epsilon}(x_0)$$

Then,

$$\bar{h}_{\beta_\epsilon}(x) + (1-\beta_\epsilon)J_{\beta_\epsilon}(x_0) + (1-\beta_\epsilon)\int \mathcal{T}(dx'|x, \gamma^\delta_{\beta_\epsilon}(x))(\bar{h}_{\beta_\epsilon}(x_1)) = c(x, \gamma^\delta_{\beta_\epsilon}) + E[\bar{h}_{\beta_\epsilon}(x_1)|x_0 = x, u_0 = \gamma^\delta_{\beta_\epsilon}(x)]$$

We have that $|(1-\beta_\epsilon)J_{\beta_\epsilon}(x_0) - \zeta^*| \leq \frac{\epsilon}{2}$ and that $\|\bar{h}_{\beta_\epsilon}\|_\infty \leq \sup_{x\in\mathbb{X}} |J_{\beta_\epsilon}(x, \gamma^\delta_{\beta_\epsilon}) - J_{\beta_\epsilon}(x)| + \|h_{\beta_\epsilon}\|_\infty \leq \delta + \|h_{\beta_\epsilon}\|_\infty$. This implies that

$$\bar{h}_{\beta_\epsilon}(x) + \frac{\epsilon}{2} + (1-\beta_\epsilon)\delta + \frac{\epsilon}{2} \geq c(x, \gamma_{\beta^\delta_\epsilon}) + E[\bar{h}_{\beta_\epsilon}(x_1)|x_0 = x, u_0 = \gamma^\delta_{\beta_\epsilon}(x)]$$

This has the same form as (7.32) and thus Theorem 7.1.3(ii) implies that $\gamma^\delta_{\beta_\epsilon}$ achieves an average cost of no larger than $\zeta^* + \epsilon + \delta$

$\diamond$

Average Cost Optimality Inequality. If one cannot verify the equicontinuity assumption or the boundedness conditions, the following holds; note that the condition of strong continuity in actions for every fixed state is required here. The result essentially follows from [165, Theorem 5.4.3] with some variations in the conditions.

Theorem 7.3.7 *Let for every measurable and bounded g, the integral $\int g(x_{t+1})P(dx_{t+1}|x_t = x, u_t = u)$ be continuous in u for every x, and there exist $N < \infty$ and a function $b(x)$ with*

$$-N \leq h_\beta(x) \leq b(x), \quad \beta \in (0,1), x \in \mathbb{X} \tag{7.33}$$

and for all $\beta \in [\alpha, 1)$ for some $\alpha < 1$ and $M \in \mathbb{R}_+$:

$$(1-\beta)J^*_\beta(z) \leq M. \tag{7.34}$$

Under these conditions, the Average Cost Optimality Inequality (ACOI) holds for appropriate η, ζ^*, f:

$$\begin{aligned} \eta(x) &\geq \min_{u\in\mathbb{U}} \left(c(x,u) + \int \mathcal{T}(dx'|x_t,u_t)\eta(x') - \zeta^* \right) \\ &= \left(c(x,f(x)) + \int \mathcal{T}(dx'|x,f(x))\eta(x') - \zeta^* \right) \end{aligned} \tag{7.35}$$

In particular, the stationary and deterministic policy $\gamma = \{f, f, f, \cdots\}$ is optimal.

Proof. By (7.34) and (7.26), we have that $(1-\beta_{n_k})J_{\beta_{n_k}}(x_0) \to \zeta^*$ for some subsequence n_k and $\beta_{n_k} \uparrow 1$. On the other hand by the optimality of J_β, and by Abelian inequalities, under any policy γ, and any sequence $\beta \uparrow 1$

$$\limsup_{\beta\to 1}(1-\beta)J_\beta(x_0) \leq \limsup_{\beta\to 1}(1-\beta)E^\gamma_{x_0}[\sum_{k=0}^{\infty}\beta^k c(x_k,u_k)] \leq \limsup_{T\to\infty}\frac{1}{T}E^\gamma_{x_0}[\sum_{k=0}^{T-1}c(x_k,u_k)]$$

thus ζ^* is a lower bound under any admissible policy. We now make the argument that this applies for any initial condition: Consider some arbitrary state $x \in \mathbb{X}$,

$$(1-\beta_{n_k})J_{\beta_{n_k}}(x) = (1-\beta_{n_k})(J_{\beta_{n_k}}(x) - J_{\beta_{n_k}}(x_0)) + (1-\beta_{n_k})J_{\beta_{n_k}}(x_0)$$

By (7.33), $(1-\beta_{n_k})(J_{\beta_{n_k}}(x) - J_{\beta_{n_k}}(x_0)) \to 0$. Thus, for any x, $(1-\beta_{n_k})J_{\beta_{n_k}}(x) \to \zeta^*$.

We now show that (7.35) holds. For this, consider again:

$$\begin{aligned} &J_\beta(x) - J_\beta(x_0) \\ &= \min_{u\in\mathbb{U}} \left(c(x,u) + \beta \int \mathcal{T}(dx'|x,u)(J_\beta(x') - J_\beta(x_0)) - (1-\beta)J_\beta(x_0) \right) \end{aligned} \tag{7.36}$$

Observe the following along the subsequence n_k, with $h_\beta(x) = J_\beta(x) - J_\beta(x_0)$, $(1-\beta_{n_k})J_{\beta_{n_k}}(x_0) \to \zeta^*$ and

$$\begin{aligned} &\liminf_{n_k\to\infty} J_{\beta_{n_k}}(x) - J_{\beta_{n_k}}(x_0) \\ &= \liminf_{n_k\to\infty} \min_{u\in\mathbb{U}} \left(c(x,u) + \beta_{n_k} \int \mathcal{T}(dx'|x,u)h_{\beta_{n_k}}(x') \right) - (1-\beta_{n_k})J_{\beta_{n_k}}(x_0) \\ &= \lim_{n_k\to\infty} \inf_{n_m > n_k} \min_{u\in\mathbb{U}} \left(c(x,u) + \beta_{n_m} \int \mathcal{T}(dx'|x,u)h_{\beta_{n_m}}(x') \right) - (1-\beta_{n_k})J_{\beta_{n_k}}(x_0) \\ &\geq \lim_{n_k\to\infty} \min_{u\in\mathbb{U}} \left(c(x,u) + \beta_{n_k} \int \mathcal{T}(dx'|x,u)H_{n_k}(x') \right) - (1-\beta_{n_k})J_{\beta_{n_k}}(x_0) \\ &= \min_{u\in\mathbb{U}} \left(c(x,u) + \int \mathcal{T}(dx'|x,u)\eta(x') - \zeta^* \right) \end{aligned}$$

where $H_{n_k}(x) := \min(\inf_{n_m > n_k} h_{\beta_{n_m}}(x), n_k)$ (so that this is a bounded function) and $\eta(x) = \lim_{m\to\infty} H_m(x)$ (so that $\eta(x) = \liminf_{n_k\to\infty} J_{\beta_{n_k}}(x) - J_{\beta_{n_k}}(x_0)$ which is the left hand term of the equation above). The last equality holds since $H_k \uparrow \eta$ as shown in Lemma 5.5.2 (though with an inequality sign since η is not necessarily bounded), and bounded from below and that $\int \mathcal{T}(dx'|x,u)H_k(x')$ is continuous on $\mathbb{U}$ (this is where we use the strong continuity in actions property given as a hypothesis).

We now show that the stationary policy $\gamma = f^\infty =: \{f, f, f, \cdots\}$ is optimal, via Theorem 7.1.3(ii): Using (7.35) repeatedly, we have that

$$E_x^{f^\infty}[\sum_{k=0}^{T-1} c(x_k, u_k)] \leq T\zeta^* + \eta(x) - E_x^{f^\infty}[\eta(x_T)] \leq T\zeta^* + \eta(x) + N.$$

Dividing by T and taking the $\limsup$, leads to the result that

$$\limsup_{T\to\infty} \frac{1}{T} E_x^{f^\infty}[\sum_{k=0}^{T-1} c(x_k, u_k)] \leq \zeta^*.$$

This completes the proof. ◇

Further sufficient conditions exist in the literature for ACOE or ACOI to hold (see [166], [318]). These conditions typically have the form of Assumption 5.5.2 or 5.5.3 together with geometric ergodicity conditions with condition (5.40) replaced with conditions of the form:

$$\sup_{u\in\mathbb{U}} \int_{\mathbb{X}} w(y)\mathcal{T}(dy|x,u) \leq \alpha w(x) + K\phi(x,u),$$

where $\alpha \in (0,1)$, $K < \infty$ and ϕ a positive function. In some approaches, ϕ and w needs to be continuous, in others it does not. For example if $\phi(x,u) = 1_{\{x\in C\}}$ for some small set C, then we recover a condition similar to (4.29) leading to geometric ergodicity.

We also note that for the above arguments to hold, there does not need to be a single invariant distribution. Here in (7.36), the pair x and x_0 should be picked as a function of the reachable set under a given sequence of policies. The analysis for such a condition is tedious in general since for every β a different optimal policy will typically be adopted; however, for certain applications the reachable set from a given point may be independent of the control policy applied.

7.4 The Convex Analytic Approach to Average Cost Markov Decision Problems

The convex analytic approach (typically attributed to Manne [222] and Borkar [59] (see also [165])) is a powerful approach to the optimization of infinite-horizon problems. It is particularly effective in proving results on the optimality of stationary policies, which can lead to a linear program. This approach is particularly effective for constrained optimization problems and infinite horizon average cost optimization problems. It avoids the use of dynamic programming or iterative contraction methods.

We are interested in the minimization

$$\inf_{\gamma\in\Gamma_A} \limsup_{T\to\infty} \frac{1}{T} E_{x_0}^{\gamma}[\sum_{t=1}^{T} c(x_t, u_t)], \tag{7.37}$$

where, as before, $E_{x_0}^{\gamma}[\cdot]$ denotes the expectation over all sample paths with initial state given by x_0 under some admissible policy γ.

7.4.1 Finite state/action setup

We first consider the finite space setting where both $\mathbb{X}$ and $\mathbb{U}$ are finite sets. We study the limit distribution of the following *empirical occupation measures* (and their expected values), under any policy γ in Γ_A. Let for $T \geq 1$

$$v_T(D) = \frac{1}{T} \sum_{t=0}^{T-1} 1_{\{(x_t,u_t)\in D\}}, \quad D \in \mathcal{B}(\mathbb{X} \times \mathbb{U}).$$

Consider policy γ in Γ_A, $x_0 \sim \eta$, and let for $T \geq 1$, the *expected* empirical occupation measures be given with

$$\mu_T(D) = E^\gamma_\eta[v_T(D)] = E^\gamma_\eta \frac{1}{T}\Big[\sum_{t=0}^{T-1} 1_{\{x_t,u_t)\in D\}}\Big], \quad D \in \mathcal{B}(\mathbb{X} \times \mathbb{U})$$

Let for $\eta \in \mathcal{P}(\mathbb{X} \times \mathbb{U})$,

$$\eta\mathcal{T}(A \times \mathbb{U}) := \sum_{x\in\mathbb{X},u\in\mathbb{U}} \mathcal{T}(A|x,u)\eta(x,u).$$

We then have

$$\begin{aligned}
&(\mu_T\mathcal{T})(A \times \mathbb{U}) \\
&= \sum_{x\in\mathbb{X},u\in\mathbb{U}} \mathcal{T}(A|x,u)\mu_T(x,u) \\
&= \sum_{x\in\mathbb{X},u\in\mathbb{U}} E^\gamma[1_{\{x_1\in A\}}|x_0 = x, u_0 = u]\mu_T(x,u) \\
&= \frac{1}{T}\sum_{k=0}^{T-1} E^\gamma\Big[\sum_{x\in\mathbb{X},u\in\mathbb{U}} E[1_{\{x_{k+1}\in A\}}|x_k = x, u_k = u]1_{\{x_k=x,u_k=u\}}\Big] \\
&= \frac{1}{T}\sum_{k=0}^{T-1} \sum_{x\in\mathbb{X},u\in\mathbb{U}} E^\gamma[1_{\{x_{k+1}\in A\}}|x_k = x, u_k = u]P(x_k = x, u_k = u) \\
&= \frac{1}{T}\sum_{k=0}^{T-1} E^\gamma[1_{\{x_{k+1}\in A\}}] \qquad (7.38)
\end{aligned}$$

Then, through what is often referred to as a *Krylov-Bogoliubov-type* argument, for every $A \subset \mathbb{X}$,

$$\begin{aligned}
&|\mu_T(A \times \mathbb{U}) - \mu_T\mathcal{T}(A \times \mathbb{U})| \\
&= E^\gamma_{\mu_0} \frac{1}{T}\Big[\sum_{t=0}^{T-1} 1_{\{x_t,u_t)\in(A\times\mathbb{U})\}} - \sum_{t=0}^{T-1} 1_{\{x_{t+1},u_{t+1})\in(A\times\mathbb{U})\}}\Big] \qquad (7.39) \\
&\leq \frac{1}{T} \to 0,
\end{aligned}$$

as $T \to \infty$. Notice that the above applies for any policy $\gamma \in \Gamma_A$.

Now, if we can ensure that for some subsequence, $\mu_{t_k} \to \mu$ for some probability measure μ, it would follow that $\mu_{t_k}\mathcal{T}(A \times \mathbb{U}) \to \mu\mathcal{T}(A \times \mathbb{U})$.

Now define

$$\mathcal{G} = \Big\{v \in \mathcal{P}(\mathbb{X} \times \mathbb{U}) : v(B \times \mathbb{U}) = \sum_{x,u} P(x_{t+1} \in B|x_t = x, u_t = u)v(x,u), \quad B \in \mathcal{B}(\mathbb{X})\Big\}$$

Further define

$$\mathcal{G}_{\mathbb{X}} = \left\{ v \in \mathcal{P}(\mathbb{X} \times \mathbb{U}) : \exists \gamma \in \Gamma_S, v(A) = \sum_{x,u} P^{\gamma}((x_{t+1}, u_{t+1}) \in A | x_t = x, u_t = u) v(x,u), \quad A \in \mathcal{B}(\mathbb{X} \times \mathbb{U}) \right\} \tag{7.40}$$

We can establish the equivalence of these sets of measures: It is evident that $\mathcal{G}_{\mathbb{X}} \subset \mathcal{G}$ since there are (seemingly) fewer restrictions for $\mathcal{G}$. We can show that these two sets are indeed equal: For $v \in \mathcal{G}$, if we write: $v(x,u) = \pi(x)\eta(u|x)$ for some η, then, we can construct a consistent $v \in \mathcal{G}_{\mathbb{X}}$: $v(B \times C) = \sum_{x \in B} \eta(C|x)\pi(x)$. The set $\mathcal{G}$ is called the set of *invariant occupation measures* (or, as is used more commonly in the literature: *ergodic occupation measures*).

Thus, every *converging subsequence* μ_{t_k} will converge to $\mathcal{G}$. And hence, any sequence $\{\mu_k\}$ will have a converging subsequence whose limit will be in the set $\mathcal{G}$. This is where finiteness is helpful: If the state space were countable, there would be no guarantee that every sequence of occupation measures would have a converging subsequence. The following has thus been established.

Lemma 7.4.1 *Under any admissible policy, any converging subsequence* $\{\mu_{t_k}\}$ *will converge to the set* $\mathcal{G}$.

Let $\langle \mu, c \rangle := \sum \mu(x,u)c(x,u)$. Let us again write that

$$J^*(x) := \inf_{\gamma \in \Gamma_A} J(x, \gamma),$$

with

$$J(x,\gamma) := \limsup_{T \to \infty} \frac{1}{T} E_x^{\gamma}[\sum_{t=0}^{T-1} c(x_t, u_t)],$$

or

$$J(x,\gamma) := \limsup_{T \to \infty} \langle \mu_T, c \rangle,$$

where μ_T is the expected empirical occupation measure under γ. Now, we have that, for any policy γ,

$$\begin{aligned} &\limsup_{T \to \infty} \langle \mu_T, c \rangle \geq \liminf_{T \to \infty} \langle \mu_T, c \rangle \\ &= \lim_{T_k \to \infty} \langle \mu_{T_k}, c \rangle = \langle \lim_{T'_k \to \infty} \mu_{T'_k}, c \rangle \geq \delta^* \end{aligned} \tag{7.41}$$

where

$$\delta^* = \inf_{v \in \mathcal{G}} \sum v(x,u)c(c,u)$$

In the above, μ_{T_k} is a subsequence for which $\langle \mu_{T_k}, c \rangle$ converges to the liminf value, there is a further subsequence $\mu_{T'_k}$ (due to the compactness of the space of expected empirical occupation measures) which has a limit, and this limit is in $\mathcal{G}$.

That is,

$$\lim_{T_k \to \infty} \langle \mu_{T_k}, c \rangle = \langle \lim_{T'_k \to \infty} \mu_{T'_k}, c \rangle \geq \delta^*.$$

Thus, we have established that

$$J^*(x) \geq \delta^*$$

If the initial state, or measure on the initial state, can be selected appropriately, or if the controlled Markov chain under an optimal policy is positive Harris recurrent the above also becomes an equality. The solution to this problem then gives us the optimal cost (under any policy). Thus, an optimal policy can be obtained through the following *linear program*:

Linear Program For Finite Models.

Given a cost function c and transition kernel $\mathcal{T}$, find the minimum

$$\min_{\nu \in \mathcal{G}} \sum_{\mathbb{X}\times\mathbb{U}} \nu(x,u)c(x,u). \tag{7.42}$$

over all probability measures ν that satisfy

$$\nu \in \mathcal{G} = \left\{ \mu \in \mathcal{P}(\mathbb{X}\times\mathbb{U}) : \mu(z,\mathbb{U}) = \sum_{\mathbb{X}\times\mathbb{U}} \mathcal{T}(z|(x,u))\mu(x,u), \quad z \in \mathbb{X} \right\}.$$

where the constraint set can also be written as

$$\sum_j \mu(z,j) = \sum_{\mathbb{X}\times\mathbb{U}} \mathcal{T}(z|(x,u))\mu(x,u), \quad z \in \mathbb{X}$$

with

$$\mu(x,u) \geq 0, \qquad x \in \mathbb{X}, u \in \mathbb{U}$$

$$\sum_{x\in\mathbb{X}, u\in\mathbb{U}} \mu(x,u) = 1$$

All of these are linear/affine constraints.

If under any stationary policy the induced Markov chain would be irreducible, then the solution of the problem (7.42) above equals the optimal cost $J^*_\infty(x)$. Let μ^* be the optimal occupation measure (this exists since the state space is finite, and thus $\mathcal{G}$ is compact, and $\sum_{\mathbb{X}\times\mathbb{U}} \mu(x,u)c(x,u)$ is continuous in μ). This induces an optimal policy $\gamma^*(u|x)$ as (defined almost surely, i.e., for x with $\sum_{\mathbb{U}} \mu^*(x,u) > 0$):

$$\gamma^*(u|x) = \frac{\mu^*(x,u)}{\sum_{\mathbb{U}} \mu^*(x,u)}.$$

Thus, we can find the optimal policy through a linear program.

7.4.2 General state/action spaces under weak continuity

The arguments presented above apply to general spaces as well. However, for the more general case considered here, we need to ensure that the set of expected occupation measures is tight, and that the set $\mathcal{G}$ is closed. In the following, we follow the presentation in [16].

We first study the weakly continuous setup, studied in [14, 14, 59, 161, 165, 201], on the existence of an optimal $\mu \in \mathcal{G}$ under the hypothesis that the transition kernel $\mathcal{T}$ is weakly continuous (Assumption 5.2.1(i)).

(H1) The transition kernel $\mathcal{T}$ is *weakly continuous*, that is

$$\mathcal{T}(f)(x) := \int_{\mathbb{X}} f(z)\mathcal{T}(dz|x,u)$$

is continuous in x, u for all $f \in C_b(\mathbb{X})$. Recall that this is the same as Assumption 5.2.1(i).

Continuing, for $T \geq 1$, we let

$$v_T(D) \;=\; \frac{1}{T}\sum_{t=0}^{T-1} 1_D(X_t, U_t), \quad D \in \mathcal{B}(\mathbb{X}\times\mathbb{U})\,.$$

Consider any policy γ in Γ_A, $X_0 \sim \nu$, and let for $T \geq 1$,

$$\mu_T^\gamma(D) \;=\; E_\nu^\gamma[v_T(D)] \;=\; \frac{1}{T}E_\nu^\gamma\left[\sum_{t=0}^{T-1} 1_D(X_t,U_t)\right], \quad D\in\mathcal{B}(\mathbb{X}\times\mathbb{U})\,.$$

We refer to $\{\mu_T^\gamma\}_{T>0}$ as the family of *mean empirical occupation measures* under the policy $\gamma\in\Gamma_A$, and with initial distribution ν. Again, through a *Krylov-Bogoliubov-type* argument, for every $A\in\mathcal{B}(\mathbb{X})$, we have

$$\begin{aligned}|\mu_T^\gamma(A\times\mathbb{U}) - \mu_T^\gamma\mathcal{T}(A)| \;&=\; \frac{1}{T}\,|E_\nu^\gamma\left[\sum_{t=0}^{T-1} 1_{A\times\mathbb{U}}(X_t,U_t) - \sum_{t=1}^{T} 1_{A\times\mathbb{U}}(X_t,U_t)\right]| \\ &\le\; \frac{1}{T}\;\to\;0 \quad \text{as } T\to\infty\,.\end{aligned} \tag{7.43}$$

Observe that (7.43) holds for any policy $\gamma\in\Gamma_A$. Suppose that, along some subsequence $\{t_k\}\subset\mathbb{N}$, μ_t^γ converges weakly to some $\mu\in(\{(x,u)\colon\ x\in\mathbb{X},\,u\in\mathbb{U}(x)\})$, which we denote as $\mu_{t_k}^\gamma\Rightarrow\mu$. We write the triangle inequality

$$|\mu(f)-\mu\mathcal{T}(f)| \;\le\; |\mu(f)-\mu_{t_k}^\gamma(f)| + |\mu_{t_k}^\gamma(f)-\mu_{t_k}^\gamma\mathcal{T}(f)| + |\mu_{t_k}^\gamma\mathcal{T}(f)-\mu\mathcal{T}(f)| \tag{7.44}$$

for $f\in C_b(\mathbb{X})$. This notation is consistent since f may be viewed also as an element of $C_b(\{(x,u)\colon\ x\in\mathbb{X},\,u\in\mathbb{U}(x)\})$. Suppose that Assumption 5.2.1(i)) holds. The first term on the right hand side of (7.44) vanishes as $k\to\infty$ by weak convergence, while the second term does the same by (7.43). Since

$$\mu_{t_k}^\gamma\mathcal{T}(f) = \mu_{t_k}^\gamma(\mathcal{T}f), \tag{7.45}$$

and $\mathcal{T}f\in C_b(\{(x,u)\colon\ x\in\mathbb{X},\,u\in\mathbb{U}(x)\})$ by Assumption 5.2.1(i)), it follows that the third term also vanishes as $k\to\infty$ by the weak convergence $\mu_{t_k}^\gamma\Rightarrow\mu$. Since the class $C_b(\mathbb{X})$ distinguishes points in $\mathcal{P}(\mathbb{X})$, this shows that $\mu(A,\mathbb{U})=\mu\mathcal{T}(A)$ for all $A\in\mathcal{B}(\mathbb{X})$, which implies that $\mu\in\mathcal{G}$ by the definition of the latter. Thus we have shown the following.

Lemma 7.4.2 *Under Assumption 5.2.1(i)), the limit of any weakly converging subsequence of mean empirical occupation measures is in $\mathcal{G}$.*

This expected average cost can be written as

$$J(x,\gamma) = \limsup_{T\to\infty}\,\langle\mu_T^\gamma, c\rangle\,,$$

where μ_T^γ is the mean empirical occupation measure under γ. Let $\{t_k\}\subset\mathbb{N}$ be a subsequence along which $\langle\mu_{t_k}^\gamma,c\rangle$ converges to $J(x,\gamma)$ and suppose that $\mu_{t_k}\Rightarrow\mu\in\mathcal{G}$. Then

$$J(x,\gamma) \;=\; \liminf_{t_k\to\infty}\langle\mu_{t_k}^\gamma,c\rangle \;\ge\; \left\langle\lim_{t_k\to\infty}\mu_{t_k}^\gamma, c\right\rangle \;=\; \langle\mu,c\rangle \;\ge\; \delta^*\,, \tag{7.46}$$

where for the first inequality we use the fact that, since c is lower semi-continuous (l.s.c.) and bounded from below, the map $\mu\to\langle\mu,c\rangle$ is lower semi-continuous. The above shows that $J^*(x)\;\ge\;\delta^*$. We now establish conditions for which the above is indeed an equality.

Assumption 7.4.1 *(A) The state and action spaces $\mathbb{X}$ and $\mathbb{U}$ are Polish. The set-valued map $\mathbb{U}\colon\mathbb{X}\to\mathcal{B}(\mathbb{U})$ is upper semi-continuous and closed-valued.*

(A')The state and action spaces $\mathbb{X}$ and $\mathbb{U}$ are compact. The set-valued map $\mathbb{U}\colon\mathbb{X}\to\mathcal{B}(\mathbb{U})$ is upper semi-continuous and closed-valued.

(B) The non-negative running cost function $c(x,u)$ is is l.s.c. and $c\colon\{(x,u)\colon\ x\in\mathbb{X},\,u\in\mathbb{U}(x)\}\to\mathbb{R}$ is inf-compact, i.e. $\{(x,u)\in\{(x,u)\colon\ x\in\mathbb{X},\,u\in\mathbb{U}(x)\}:c(x,u)\le\alpha\}$ is compact for every $\alpha\in\mathbb{R}_+$.

(B')The cost function c is bounded and l.s.c..

(C) There exists a policy and an initial state leading to a finite cost $\eta \in \mathbb{R}_+$.

(D) Assumption 5.2.1(i)) holds.

(E) Under every stationary policy, the induced Markov chain is Harris recurrent.

Before we present a theorem, we recall the discussion in Section 3.4.1 concerning ergodic properties of (control-free) Markov chains: Let $c \in L_1(\mu) := \{f : \mathbb{X} \to \mathbb{R}, \int |f(x)|\mu(dx) < \infty\}$. Suppose that μ is an invariant probability measure for an $\mathbb{X}$-valued Markov chain X_k. Then, by the individual ergodic theorem for μ almost everywhere $x \in \mathbb{X}$: $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} c(X_t) = \int c(x)\mu(dx)$, P_x almost surely (that is conditioned on $x_0 = x$, with probability one, the above holds). Furthermore, again with $c \in L_1(\mu)$, for μ almost everywhere $x \in \mathbb{X}$

$$\lim_{T\to\infty} \frac{1}{T} E_x\left[\sum_{t=1}^{T} c(X_t)\right] = \int c(x)\mu(dx), \tag{7.47}$$

On the other hand, the positive Harris recurrence property allows the almost sure convergence to take place for every initial condition: If μ is the invariant probability measure for a *positive Harris recurrent* Markov chain, it follows that for all $x \in \mathbb{X}$ and for every $c \in L_1(\mu)$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} c(X_t) = \int c(x)\mu(dx), \tag{7.48}$$

almost surely. However, as discussed earlier, it is not generally true that

$$\lim_{T\to\infty} \frac{1}{T} E_x[\sum_{t=1}^{T} c(X_t)] = \int c(x)\mu(dx),$$

for all $x \in \mathbb{X}$. Thus, we can not in general relax the boundedness condition for the convergence of the expected costs. With c bounded, for all $x \in \mathbb{X}$

$$\lim_{T\to\infty} \frac{1}{T} E_x\left[\sum_{t=1}^{T} c(X_t)\right] = \int c(x)\mu(dx) \tag{7.49}$$

In the following, we follow the arguments in [16].

Theorem 7.4.1 *a) Under (7.4.1) A, B, C, D there exists an optimal measure in $\mathcal{G}$. b) Under (7.4.1) A', B', D, E, there exists a policy in which is optimal for the control problem given in (7.64) for every initial condition.*

Proof. a) Consider (7.4.1) A, B, C, D. By (B, C) we have that the set of policies γ which lead to a finite cost is so that $\langle \mu_T^\gamma, c\rangle < \infty$ for all T, which implies that $\{\mu_T^\gamma, T > 0\}$ is tight. Thus along some subsequence $\mu_{t_k} \to \mu \in (\{(x,u)\colon\ x \in \mathbb{X},\ u \in \mathbb{U}(x)\})$. As shown in the paragraph preceding (7.4.2), $\mu \in \mathcal{G}$.

Furthermore, under hypothesis (A), the set $\mathbb{K} = \{(x, \mathbb{U}(x)), x \in \mathbb{X}\}$ is closed by [165, Lemma D.3]. Thus, by the Portmanteau theorem that every weak limit of a converging sequence of probability measures $\mathbb{K}) = 1$ is also supported on $\mathbb{K}$.

Consider a sequence $\{\mu_k\}_{k\in\mathbb{N}} \subset \mathcal{G}$ such that $\langle \mu_k, c\rangle \to \delta^*$ as $k \to \infty$, the sequence μ_{t_k} is tight by inf-compactness, and any limit point μ_* of this sequence is in $\mathcal{G}$ with $\mu_*(\{(x, \mathbb{U}(x)), x \in \mathbb{X}\}) = 1$. Thus, by [165, Prop. D.8] we have an optimal control policy ϕ. Taking limits as in (7.46), we obtain $\langle \mu_*, c\rangle = \delta^*$. This establishes the first part of the theorem.

Define a stationary policy γ via the disintegration

$$\mu_*(dx, \mathrm{d}u) \;=\; \gamma_*(\mathrm{d}u|x)\, \pi_*(dx) \tag{7.50}$$

μ_* almost surely. Note that via this disintegration the control γ_* is defined π_*-a.e. Let $\phi \in$ be any policy that agrees with γ_* on the support of π_*.

b) Under (A', B', D), via (7.45) and that $\mathcal{T}f \in C_b(\{(x,u)\colon\ x \in \mathbb{X},\ u \in \mathbb{U}(x)\})$ by Assumption 5.2.1(i)), we have that $\mathcal{G}$ is compact; we also have that Portmanteau theorem applies as in part a). By hypothesis (E), since the chain under an optimal ϕ, is Harris recurrent, π_* is its unique invariant probability measure. Optimality of ϕ, for every initial condition, then follows by positive Harris recurrence given that c is bounded via (7.49) under hypothesis (B'). Thus, $J(x,\phi) = \langle \mu_*, c\rangle$ and μ_* is optimal.

$\diamond$

Theorem 7.4.1 can be stated under weaker assumptions. See, for example, [13, Theorem 2.1] among other references in the literature.

In general, in the absence of (7.4.1) (E), there is a question of reachability. Suppose that the chain under the policy ϕ as defined in the proof of Theorem 7.4.1 is a T model (see [309]). Then, as asserted in [309, Theorem 6.1], the Doeblin decomposition of the state space contains, in general, a countable collection of maximal Harris sets. In particular, we have a decomposition into the disjoint union $\mathbb{X} = \left(\cup_{i\in\mathbb{N}} H_i\right) \cup E$, where each H_i is a maximal Harris set with invariant measure π_i, and E is transient. Now, by part (ii) of Theorem 6.1 in [309], only a finite number of the sets H_i may have a nonempty intersection with any given compact set. This implies that π_* can always be expressed as a convex combination of finitely many ergodic invariant measures. Thus, if the Markov Chain is not recurrent, the stationary policy defined above, in general, is only optimal in a restricted set of initial conditions. On implications related to insensitivity to such initial state dependence, the reader is referred to [214] and [166, Prop. 11.4.4(c) and Lemma 11.4.5(a)], among other references, for further results on sample path average cost optimality and expected average cost optimality. See [14] for further discussions.

Following the above, there exists an optimal expected empirical occupation measure, say v. This defines the optimal stationary control policy by the decomposition:

$$\mu(\cdot|u) = \frac{dv(\cdot, du)}{d\int_{u\in\mathbb{U}} v(\cdot, du)}(u),$$

v almost surely, where $\frac{d\cdot}{d\cdot}$ denotes the Radon-Nikodym derivative.

7.4.3 General state/action spaces under strong continuity in actions

There are many important applications where the kernel $\mathcal{T}$ is not weakly continuous. For example, consider dynamics described by a stochastic difference equation on $\mathbb{R}^d$ of the form

$$X_{n+1} \;=\; F(X_n, U_n) + W_n\,, \qquad n = 0, 1, 2, \dots\,,$$

where $\mathbb{X} = \mathbb{R}^n$ and the W_n's are independent and identically distributed (i.i.d.) random vectors whose distribution has a bounded and continuous density function. We assume that F is bounded and $u \mapsto F(x,u)$ is continuous for all $x \in \mathbb{X}$. It is clear that the transition kernel $\mathcal{T}$ is not, in general, weakly continuous. However, it satisfies the following hypothesis.

(H2) The transition kernel $\mathcal{T}$ satisfies the following:

(a) For any $x \in \mathbb{X}$, the map $u \mapsto \int f(z)\mathcal{T}(dz|x,u)$ is continuous for every bounded measurable function f. That is, Assumption 5.2.2(i)) holds.

(b) There exists a finite measure ν majorizing $\mathcal{T}$, that is

$$\mathcal{T}(dy|x,u) \;\le\; \nu(dy)\,, \qquad x \in \mathbb{X},\ u \in \mathbb{U}\,. \tag{7.51}$$

If in addition, the distribution of W_n has a continuous, bounded, and *a strictly positive* probability density function (a non-degenerate Gaussian distribution satisfies this condition), then positive Harris recurrence can be established by Lebesgue-irreducibility and *a uniform countable additivity condition* for compact sets following [310, Condition A], which leads to the presence of accessible compact petite sets (where one can take $V(x) = x^2$ as the Lyapunov function). For more details see [269, Example 3.1].

Assumption 7.4.2 *The following hold:*

(A) The state and action spaces $\mathbb{X}$ and $\mathbb{U}$ are Polish. The set $\mathbb{K} = \{(x, \mathbb{U}(x)), x \in \mathbb{X}\}$ is measurable (see [165, Lemma D.3] for conditions) and the set-valued map $\mathbb{U}\colon \mathbb{X} \to \mathcal{B}(\mathbb{U})$ is compact-valued.

(A') The state and action spaces $\mathbb{X}$ and $\mathbb{U}$ are compact. The set $\mathbb{K}$ is measurable and set-valued map $\mathbb{U}\colon \mathbb{X} \to \mathcal{B}(\mathbb{U})$ is compact-valued.

(B) The non-negative running cost function $c(x, u)$ is continuous in $u \in \mathbb{U}(x)$ for every $x \in \mathbb{X}$ and $c\colon \big\{(x, u)\colon\ x \in \mathbb{X},\ u \in \mathbb{U}(x)\big\} \to \mathbb{R}$ is inf-compact.

(B') The cost function c is bounded, and continuous in $u \in \mathbb{U}(x)$ for every $x \in \mathbb{X}$.

(C) There exists a policy and an initial state leading to a finite cost $\eta \in \mathbb{R}_+$.

(D) (H2) holds.

(E) Under every stationary policy, the induced Markov chain is Harris recurrent.

We again recall the w-s topology studied in Appendix (Section D.4).

By Theorem D.4.1 (see [284, Theorem 3.10] or [27, Theorem 2.5]), (7.51) implies that setwise sequential pre-compactness of marginal measures on the state ensures that every weakly converging sequence of mean empirical occupation measures also converges in the w-s sense. (7.51) implies setwise sequential pre-compactness by [269, Proposition 3.2], which in turn builds on [168, Corollary 1.4.5]; see also [153, Theorem 4.17].

First, note the following counterpart to Lemma 7.4.2.

Lemma 7.4.3 *[16] Under (H2), the limit of any w-s converging subsequence of mean empirical occupation measures is in $\mathcal{G}$.*

Proof. We follow the notation used in the discussion leading to Lemma 7.4.2. Suppose that, along some subsequence $\{t_k\} \subset \mathbb{N}$, μ_t^γ converges to some $\mu \in (\big\{(x, u)\colon\ x \in \mathbb{X},\ u \in \mathbb{U}(x)\big\})$ in the w-s sense, which we denote as $\mu_{t_k}^\gamma \Rightarrow \mu$. As in (7.44) we have the triangle inequality

$$\begin{aligned}\mu(f) - \mu\mathcal{T}(f) \ \le\ & \mu(f) - \mu_{t_k}^\gamma(f) + \mu_{t_k}^\gamma(f) - \mu_{t_k}^\gamma\mathcal{T}(f) \\ & + \mu_{t_k}^\gamma\mathcal{T}(f) - \mu\mathcal{T}(f)\end{aligned} \tag{7.52}$$

for $f \in \mathcal{M}_b(\mathbb{X})$. If (H2) holds, the first term on the right hand side of (7.52) vanishes as $k \to \infty$ by w-s convergence, while the second term does so by (7.43). We have

$$\mu_{t_k}^\gamma\mathcal{T}(f) = \mu_{t_k}^\gamma(\mathcal{T}f), \tag{7.53}$$

Since $\mathcal{T}f$ is continuous in u for every fixed x, by (H2), it follows that the third term also vanishes as $k \to \infty$ by the w-s convergence $\mu_{t_k}^\gamma \Rightarrow \mu$. This shows that $\mu(A, \mathbb{U}) = \mu\mathcal{T}(A)$ for all $A \in \mathcal{B}(\mathbb{X})$, which implies that $\mu \in \mathcal{G}$. $\diamond$

Theorem 7.4.2 *[16] a) Under (7.4.2) A, B, C, D, there exists an optimal measure in $\mathcal{G}$. b) Under (7.4.2) A', B', D, E, there exists a policy in which is optimal for the control problem given in (7.64) for every initial condition.*

7.4.4 Optimality of deterministic stationary policies

In this section, we will present conditions on the optimality of deterministic policies via the convex analytic method; see [16] for a detailed review and [232, Proposition 9.2.5], [59, Lemma 2.4] and [161, Corollary 5.4(b)] for related results on the optimality of stationary and deterministic policies arrived at via different approaches.

Hernandez-Lermá [161, Theorem 5.3] shows that an average cost optimal randomized policy ϕ, with invariant measure π_ϕ satisfies the ACOI π_ϕ almost everywhere:

$$g + h(x) \geq c(x, \phi(x)) + \int h(x')\mathcal{T}(dx'|x, \phi(x)) \tag{7.54}$$

where h is bounded from below. If one can ensure that the above holds for all $x \in \mathbb{X}$ (and not just π_ϕ almost everywhere) [161, Prop. 5.2] shows that under this condition on h, (7.54) implies that such a policy is indeed optimal. Again, if the above holds for all $x \in \mathbb{X}$, by utilizing Blackwell's theorem 5.1.1 on optimality of deterministic policies we can replace ϕ with a deterministic $f \in \Gamma_{SD}$, which will then be optimal [161, Corollary 5.4(b)].

This approach can be generalized to the case where the induced Markov chain is not positive Harris recurrent, but when the action space is finite [16]: Accordingly, one can relax the condition of (7.54) holding for every x. Let g be a constant and $h : \mathbb{X} \to \mathbb{R}_+$, $f : \mathbb{X} \to \mathcal{P}(\mathbb{U})$ be so that for all $x \in B$ for some Borel set $B \subset \mathbb{X}$,

$$g + h(x) \geq \left(c(x, f(x) + \int h(x')\mathcal{T}(dx'|x, f(x)) \right) := \int \left(c(x, u) + \int h(x')\mathcal{T}(dx'|x, u) \right) f(du|x) \tag{7.55}$$

Lemma 7.4.4 *Let (7.55) hold with*

$$\liminf_{n\to\infty} \frac{1}{n} E_x^{\gamma^*}[h(X_n)] \geq 0, \tag{7.56}$$

for all $x \in B$ where $\gamma^* = \{f, f, f, \cdots\}$ and $P^{\gamma^*}(x, B) = 1$ for all $x \in B$. Then the stationary (possibly randomized) policy $\gamma^* = \{f, f, f, \cdots\}$ satisfies

$$g \geq J(x, \gamma^*),$$

for all $x \in B$.

Proof. We have, as in the proof of Theorem 7.1.1,

$$E^\gamma[h(X_t)|x_{[0,t-1]}, u_{[0,t-1]}] = \int_y h(y)P(X_t \in dy|x_{t-1}, u_{t-1}) \tag{7.57}$$

$$= c(x_{t-1}, u_{t-1}) + \int_y h(y)P(dy|x_{t-1}, u_{t-1}) - c(x_{t-1}, u_{t-1}) \tag{7.58}$$

By iterated expectations,

$$E_x^{\gamma^*}\left[\sum_{t=1}^{n} h(X_t) - E^{\gamma^*}[h(X_t)|X_{[0,t-1]}, U_{[0,t-1]}] \right] = 0$$

Now, under γ^* we have that B is an absorbing set and thus, by (7.55) holding on the absorbing set, the following will apply almost surely with $X_0 = x$ where $x \in B$:

$$E^{\gamma^*}[h(X_t)|x_{[0,t-1]}, u_{[0,t-1]}] = \int_y h(y)P(X_t \in dy|x_{t-1}, u_{t-1}) \tag{7.59}$$

$$= c(x_{t-1}, f(x_{t-1})) + \int_y h(y)P(dy|x_{t-1}, f(x_{t-1})) - c(x_{t-1}, f(x_{t-1})) \tag{7.60}$$

$$\leq g + h(x_{t-1}) - c(x_{t-1}, f(x_{t-1})) \tag{7.61}$$

Iterating the above and dividing by n, we arrive at

$$g - \frac{1}{n}E_x^{\gamma^*}[h(X_n)] + \frac{1}{n}E_x^{\gamma^*}[h(X_0)] \geq \frac{1}{n}E_x^{\gamma^*}[\sum_{t=1}^{n} c(X_{t-1}, X_{t-1})].$$

Taking the limsup on both sides (and replacing $\limsup$ with $\liminf$ by reversing the negative sign on the left), and (7.56) holding for $\gamma^* = \{f, f, f, \cdots\}$, we establish the desired bound. $\diamond$

In particular if we have that g is a lower bound on the optimal cost (say via the convex analytic method), we can claim that γ^* is optimal for all initializations $X_0 = x$ where $x \in B$. Now, the analysis in [161, Theorem 5.3] shows that if we have an optimal invariant measure, then this leads to (7.54) for some randomized ϕ on a set of measure 1 under π_ϕ with h bounded from below. Building on [48, 51], via [161, (5.7)], this implies the existence of a deterministic control policy k which is defined on B and which satisfies

$$g + h(x) \geq \left(c(x, k(x)) + \int h(x')\mathcal{T}(dx'|x, k(x)) \right) \tag{7.62}$$

However, with $\kappa^* = \{k, k, k, \cdots\}$, to be able to claim the optimality of k over B via Lemma 7.1.3, we need to show $P^{\kappa^*}(x, B) = 1$ for all $x \in B$; that is an absorbing set under k should be a subset of the absorbing set under ϕ when $X_0 = x$ with $x \in B$. If the induced Markov chain under ϕ is positive Harris recurrent, then [161, Theorem 5.3(b)] shows that 7.54 holds everywhere (that is, for all $x \in \mathbb{X}$), and the result follows. Additionally, when $\mathbb{U}$ is countable, this result also follows via the following argument: By Blackwell's theorem 5.1.1 and by the measurable selection theorem of Blackwell and Ryll-Nardzewski [51], k can be (without loss) constructed such that for all $x : k(x) \in \{u : (c(x, u) + \int h(x')\mathcal{T}(dx'|x, u)) \leq c(x, \phi(x)) + \int h(x')\mathcal{T}(dx'|x, \phi(x))\} \cap \{u : \phi(u|x) > 0\}$. In this case, it follows by expressing the transition probabilities in terms of the countable collection of control realizations, we will have that $P^{\kappa^*}(x, B) = 1$ for all $x \in B$. This leads to the following result.

Theorem 7.4.3 *Assume that either Theorem 7.4.1 or Theorem 7.4.2 apply. Let μ_* be an optimal invariant measure. Define a stationary policy γ via the disintegration*

$$\mu_*(dx, du) \; = \; \gamma_*(du|x)\, \pi_*(dx) \tag{7.63}$$

μ_* almost surely. Take $\phi \in \Gamma_S$ be any policy that agrees with γ_* on the support of π_*.

(i) [161] If the induced Markov chain under ϕ is positive Harris recurrent, then the optimal policy can be assumed deterministic.

(ii) [16] If the induced Markov chain under an optimal policy is not positive Harris recurrent, then with $\mathbb{U}$ countable, on the support of π_*, ϕ can be assumed to be deterministic. This would lead to an optimal policy for all initial states x with $X_0 = x$ where $x \in \text{supp}\,(\pi_*)$.

There exist alternative arguments in the literature: [232, Proposition 9.2.5], [59, Lemma 2.4] focus on the properties of return sets for countable state/action space models and characterize conditions under which an optimal policy is deterministic, and the analysis in [59, Section 3.2] builds on Schauder's fixed point theorem under restrictive regularity conditions on a continuous space model. Furthermore, it can be shown that, under mild ergodicity conditions, deterministic policies are dense in the sense that the performance under stationary deterministic policies is dense in the set of performance values under randomized stationary policies [16].

Remark 7.4 (ACOI through duality with the convex analytic method). Though not directly related to the discussion above, one could note that there is a further duality relationship between ACOI (see Definition 7.1.2) and the convex analytic method, see [166, Chapter 12, p. 221].

7.4.5 Sample-path optimality

The above optimality arguments also apply in the somewhat stronger *sample-path* sense, rather than only in expectation.

Finite state/action setup

Consider the following:

$$\inf_{\gamma\in\Gamma_A}\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}[c(x_t,u_t)], \tag{7.64}$$

where there is no expectation. The above is known as the sample path cost. Let the state and action spaces be finite. Let $\mathcal{F}_t$ be the $\sigma-$field generated by $\{x_s,u_s,s\le t\}$, under any given admissible policy. Define a $\mathcal{F}_t$ measurable process with $A\in\mathcal{B}(\mathbb{X})$:

$$F_t(A)=\Big(\sum_{s=1}^{t}1_{\{x_s\in A\}}-t\sum_{\mathbb{X}\times\mathbb{U}}P(A|x,u)v_t(x,u)\Big),$$

where

$$v_t(x,u)=\sum_{s=0}^{t-1}1_{\{x_s=x,u_t=u\}},$$

is the empirical occupation measure. Note that the above can also be written as

$$F_t(A)=\Big(\sum_{s=0}^{t-1}\Big(1_{\{x_{s+1}\in A\}}-\sum_{\mathbb{X}\times\mathbb{U}}P(A|x,u)1_{\{x_s=x,u_t=u\}}\Big)\Big)$$

Thus, for $t\ge 1$,

$$\begin{aligned}
&E[F_t(A)|\mathcal{F}_{t-1}]\\
&=E\Big[\sum_{s=1}^{t}1_{\{x_s\in A\}}-\sum_{s=0}^{t-1}\sum_{\mathbb{X}\times\mathbb{U}}P(x_{s+1}\in A|x_s=x,u_s=u)1_{\{(x_s,u_s)=(x,u)\}}|\mathcal{F}_{t-1}\Big]\\
&=E\Big[\Big(1_{\{x_t\in A\}}-\sum_{\mathbb{X}\times\mathbb{U}}P(x_{s+1}\in A|x_s=x,u_s=u)1_{\{(x_{t-1},u_{t-1})=(x,u)\}}\Big)|\mathcal{F}_{t-1}\Big]\\
&+\Big(\sum_{s=1}^{t-1}1_{\{x_s\in A\}}-\sum_{s=0}^{t-2}\sum_{\mathbb{X}\times\mathbb{U}}P(x_{s+1}\in A|x_s=x,u_s=u)1_{\{(x_s,u_s)=(x,u)\}}\Big)\\
&=0 \qquad (7.65)\\
&+\Big(\sum_{s=1}^{t-1}1_{\{x_s\in A\}}-\sum_{s=0}^{t-2}\sum_{\mathbb{X}\times\mathbb{U}}P(x_{s+1}\in A|x_s=x,u_s=u)1_{\{(x_s,u_s)=(x,u)\}}\Big)|\mathcal{F}_{t-1}\Big] \qquad (7.66)\\
&=F_{t-1}(A), \qquad (7.67)
\end{aligned}$$

where (7.65) follows from the fact that $E[1_{\{x_t\in A\}}|\mathcal{F}_{t-1}]=P(x_t\in A|\mathcal{F}_{t-1})$.

We have then that

$$E[F_t(A)|\mathcal{F}_{t-1}]=F_{t-1}(A)\qquad\forall t\ge 0,$$

and $\{F_t(A)\}$ is a martingale sequence.

Furthermore, $F_t(A)$ is a bounded-increment martingale since $|F_t(A)-F_{t-1}(A)|\le 1$. Hence, for every $T>2$, $\{F_1(A),\ldots,F_T(A)\}$ forms a martingale sequence with uniformly bounded increments, and we could invoke the Azuma-Hoeffding inequality [98] to show that for all $x>0$

$$P(|\frac{F_t(A)}{t}|\ge x)\le 2e^{-2x^2t}$$

Finally, invoking the Borel-Cantelli Lemma (see Theorem B.2.1) for the summability of the estimate above, that is:

$$\sum_{n=1}^{\infty} 2e^{-2x^2 t} < \infty, \forall x > 0,$$

we deduce that

$$\lim_{t\to\infty} \frac{F_t(A)}{t} = 0 \quad a.s.$$

Thus,

$$\lim_{T\to\infty} \left(v_T(A) - \sum_{\mathbb{X}\times\mathbb{U}} P(A|x,u) v_T(x,u) \right) = 0, \quad A \subset \mathbb{X}$$

Thus, somewhat similar to the arguments in (7.39), every converging subsequence would have to be in the set $\mathcal{G}$ defined in (7.40).

Let $\langle v, c\rangle := \sum v(x,u)c(x,u)$. Now, we have that

$$\liminf_{T\to\infty} \langle v_T, c\rangle \geq \delta^*$$

since for any sequence v_{T_k} which converges to the liminf value, there exists a further subsequence $v_{T'_k}$ (due to the (weak) compactness of the space of occupation measures) which has a weak limit, and this weak limit is in $\mathcal{G}$. Then,

$$\lim_{T_k\to\infty} \langle v_{T_k}, c\rangle = \langle \lim_{T'_k\to\infty} v_{T_k}, c\rangle \geq \gamma^*.$$

Furthermore, this cost is attained by an optimal stationary policy as a consequence of positive Harris recurrence.

Note that, the above would lead to the same for the average cost problem as well (though as we studied earlier, a more direct argument is applicable for the average cost setup):

$$\liminf_{T\to\infty} E[\langle v_T, c\rangle] \geq E[\liminf_{T\to\infty} \langle v_T, c\rangle] \geq \gamma^*.$$

The standard Borel setup

As we observed, the discussion in Section 7.4.1 applies to the sample path optimality also. We now discuss a more general setting where the state and action spaces are Polish. Let $\phi : \mathbb{X} \to \mathbb{R}$ be a continuous and bounded function. Define:

$$v_T(\phi) = \frac{1}{T} \sum_{t=1}^{T} \phi(x,u).$$

Define a $\mathcal{F}_t$ measurable process, with π an admissible control policy (not necessarily stationary or Markov):

$$F_t(\phi) = \left(\sum_{s=1}^{t} \phi(x_t) \right) - t \left(\int_{\mathcal{P}\times\mathbb{U}} \int \phi(x'_t) P^{\pi}(dx'_t, du'_t)|x) v_t(dx) \right) \tag{7.68}$$

As earlier, we define $\mathcal{G}_{\mathbb{X}}$ to be the following set in this case.

$$\mathcal{G}_{\mathbb{X}} = \{\eta \in \mathcal{P}(\mathbb{X}\times\mathbb{U}) : \eta(D) = \int_{\mathbb{X}\times\mathbb{U}} P(D|z)\eta(dz), \quad \forall D \in \mathcal{B}(\mathbb{X})\}.$$

Consider the following sample-path cost:

$$\inf_{\gamma} \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} [c(x_t, u_t)], \tag{7.69}$$

where there is no expectation. Let $\langle v, c\rangle := \sum v(x,u)c(x,u)$. If one can guarantee that every sequence of empirical measures $\{v_t\}$ would have a converging subsequence to some measure v, we would have that

$$\lim_{T_k\to\infty} \langle v_{T_k}, c\rangle \geq \langle \lim_{T'_k\to\infty} v_{T_k}, c\rangle,$$

by the fact that for c continuous, non-negative if $v_k \to v$,

$$\liminf_{k\to\infty} \langle v_k, c\rangle \geq \langle v, c\rangle.$$

Since for any sequence v_{T_k} which converges to the liminf value, there exists a further subsequence $v_{T'_k}$ (due to the (weak) compactness of the space of occupation measures) which has a weak limit, and this weak limit is in $\mathcal{G}$. By Fatou's Lemma:

$$\lim_{T_k\to\infty} \langle v_{T_k}, c\rangle = \langle \lim_{T'_k\to\infty} v_{T_k}, c\rangle \geq \gamma^*.$$

To apply the convex analytic approach, we require that under any admissible policy, the set of sample path occupation measures would be tight, for almost every sample path realization. If this can be established, then the result goes through not only for the expected cost, but also the sample-path average cost, as discussed for the finite state-action setup.

Researchers in the literature have tried to establish conditions which would ensure that the set of empirical occupational measures are tight. These typically follow one of two conditions: Either cost functions are *near-monotone* type conditions [59] (this includes, but is more general than, the condition: $\lim_{|x|\to\infty} \inf_{u\in\mathbb{U}} c(x,u) = \infty$) or behave like moments [214] (when $\mathbb{X}\times\mathbb{U}$ is locally compact, there exists a sequence of compact sets K_n so that $\mathbb{X}\times\mathbb{U} = \cup_n K_n$ with$\lim_{K_n\uparrow\mathbb{X}} \inf_{(x,u)\notin K_n} c(x,u) = \infty$), or the Markov chain satisfies strong recurrence properties [59] [15, Chapter 3]. Under such conditions, the sequence of empirical occupation measures $\{v_n\}$ which give rise to a finite cost are almost surely tight, every such sequence has a convergent subsequence and thus the arguments above apply: Every expected average-cost optimal policy is also sample-path optimal provided that the initial condition belongs to the support of the invariant probability measure under an optimal policy.

Note also that one often can obtain more relaxed conditions for sample path optimality when compared with expected cost optimality as a consequence of the ergodic theorems for positive Harris recurrent Markov chains (Section 3.4.1).

Note, however, that for sample path optimality, we need to invoke weak continuity almost surely, by the martingale argument above, and accordingly the measurable selection criteria will need to be under Assumption 5.2.1.

7.5 Constrained Markov Decision Processes

Consider the following average cost problem:

$$\inf_{\gamma} J(x,\gamma) = \inf_{\gamma} \limsup_{T\to\infty} \frac{1}{T} E_x^{\gamma} \sum_{t=0}^{T-1} c(x_t, u_t) \tag{7.70}$$

subject to the constraints:

$$\limsup_{T\to\infty} \frac{1}{T} E_x^{\gamma} \sum_{t=0}^{T-1} d_i(x_t, u_t) \leq D_i \tag{7.71}$$

for $i = 1, 2, \cdots, m$ where $m \in \mathbb{N}$.

A linear programming formulation leads to the following result.

Theorem 7.5.1 *[265] [7] Let $\mathbb{X}, \mathbb{U}$ be countable. Consider (7.70-7.71). An optimal policy will randomize between at most $m+1$ deterministic policies.*

Ross also discusses a setup with one constraint where a non-stationary history-dependent policy may be used instead of randomized stationary policies.

Finally, the theory of constrained Markov Decision Processes is also applicable to Polish state and action spaces, but this requires further technicalities. If there is an accessible atom (or an artificial atom as considered earlier in *Chapter 3*) under any of the policies considered, then the randomizations can be made at the atom.

7.6 Bibliographic Notes

7.7 Exercises

Exercise 7.7.1 *Let $\mathbb{X}, \mathbb{U}$ be finite sets and consider the occupation measure corresponding to a controlled Markov chain under some arbitrary admissible control policu:*

$$v_T(A \times B) = \frac{1}{T} \sum_{t=0}^{T-1} 1_{\{(x_t,u_t) \in A \times B\}}, \quad A \subset \mathbb{X}, B \subset \mathbb{U}.$$

While proving that the limit of such a measure process lives in a specific set, the following is used, which you are asked to prove. Let γ be some arbitrary but admissible control policy and let $\mathcal{F}_t$ be the $\sigma-$field generated by $\{x_s, u_s, s \leq t\}$. Define a $\mathcal{F}_t$ measurable process

$$F_t(A) = \Big(\sum_{s=1}^{t} 1_{\{x_s \in A\}} - t \sum_{\mathbb{X} \times \mathbb{U}} P(A|x) v_t(x,u) \Big),$$

Show that, $\{F_t(A), \quad t \in \mathbb{Z}_+\}$ is a martingale sequence.

Hint: Observe that for all $t \in \{1, 2, \ldots, T\}$

$$\begin{aligned} &\Big(\sum_{s=1}^{t} 1_{\{x_s \in A\}} - t \sum_{\mathbb{X} \times \mathbb{U}} P(x_1 \in A | x_0 = x, u_0 = u) v_t(x,u) \Big) \\ &= \Big(\sum_{s=1}^{t} 1_{\{x_s \in A\}} - \sum_{s=0}^{t-1} \sum_{\mathbb{X} \times \mathbb{U}} P(x_{s+1} \in A | x_s = x, u_s = u) 1_{\{(x_s,u_s)=(x,u)\}} \Big) \end{aligned} \tag{7.72}$$

Then, show that $E[F_t(A)|\mathcal{F}_{t-1}] = F_{t-1}(A)$. You may follow Exercise 4.5.11.

Exercise 7.7.2 *a) Let, for a Markov control problem, $x_t \in \mathbb{X}, u_t \in \mathbb{U}$, where $\mathbb{X}$ and $\mathbb{U}$ are finite sets denoting the state space and the action space, respectively. Consider the optimal control problem of the minimization of*

$$\limsup_{T \to \infty} \frac{1}{T} E_x^{\gamma} [\sum_{t=0}^{T-1} c(x_t, u_t)],$$

where c is a bounded function. Further assume that under any stationary control policy, the state transition kernel $P(x_{t+1}|x_t, u_t)$ leads to an irreducible Markov Chain.

Does there exist an optimal control policy? Propose a method to find an optimal policy.

b) Is the optimal policy also sample-path optimal?

Exercise 7.7.3 *Consider a controlled Markov Chain with state space $\mathbb{X} = \{0,1\}$, action space $\mathbb{U} = \{0,1\}$, and transition kernel for $t \in \mathbb{Z}_+$:*

$$P(x_{t+1} = 1 | x_t = 0, u_t = 1) = \alpha \in (0,1)$$
$$P(x_{t+1} = 1 | x_t = 0, u_t = 0) = \beta \in (0,1)$$
$$P(x_{t+1} = 1 | x_t = 1, u_t = 0) = P(x_{t+1} = 1 | x_t = 1, u_t = 1) = \frac{1}{2}$$

Let

$$c(0,1) = \kappa \in \mathbb{R}_+, \quad c(0,0) = 1$$
$$c(1,0) = c(1,1) = 1$$

Suppose, the goal is to minimize the quantity

$$\limsup_{T\to\infty} \frac{1}{T} E_0^\gamma [\sum_{t=0}^{T-1} c(x_t, u_t)],$$

over all admissible policies $\gamma \in \Gamma_A$.

Find the optimal policy and the optimal cost, as a function of α, β, κ. Explain your answer and how you arrived at your solution.

Exercise 7.7.4 *Consider a controlled Markov chain with state space $\mathbb{X} = \{0,1\}$, action space $\mathbb{U} = \{0,1\}$, and transition kernel for $t \in \mathbb{Z}_+$:*

$$P(x_{t+1} = 1 | x_t = 0, u_t = 1) = 1$$
$$P(x_{t+1} = 1 | x_t = 0, u_t = 0) = \frac{1}{2}$$
$$P(x_{t+1} = 1 | x_t = 1, u_t = 0) = P(x_{t+1} = 1 | x_t = 1, u_t = 1) = \frac{1}{2}.$$

Let a cost function $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ be given by

$$c(0,1) = \kappa \in \mathbb{R}_+, \quad c(0,0) = 1$$
$$c(1,0) = \frac{1}{2}, \qquad c(1,1) = 1.$$

Suppose that the goal is to minimize the quantity

$$\limsup_{T\to\infty} \frac{1}{T} E_0^\gamma [\sum_{t=0}^{T-1} c(x_t, u_t)],$$

over all admissible policies $\gamma \in \gamma_A$. Recall that a policy is admissible if the controller has access to $\{x_s, s \le t; u_l, l \le t-1\}$ at time $t \in \mathbb{Z}_+$.

Find an **optimal policy** *and the* **optimal expected cost** *explicitly, as a function of κ. Explain your answer and how you arrived at your solution.*

Exercise 7.7.5 *Consider a two-state, controlled Markov Chain with state space $\mathbb{X} = \{0,1\}$, and transition kernel for $t \in \mathbb{Z}_+$:*

$$P(x_{t+1} = 0 | x_t = 0) = u_t^0$$
$$P(x_{t+1} = 1 | x_t = 0) = 1 - u_t^0$$
$$P(x_{t+1} = 1 | x_t = 1) = u_t^1$$
$$P(x_{t+1} = 0 | x_t = 1) = 1 - u_t^1.$$

Here $u_t^0 \in [0.2, 1]$ and $u_t^1 \in [0, 1]$ are the control variables. Suppose, the goal is to minimize the quantity

$$\limsup_{T\to\infty} \frac{1}{T} E_0^\gamma [\sum_{t=0}^{T-1} c(x_t, u_t)],$$

where

$$c(0, u^0) = 1 + u^0,$$
$$c(1, u^1) = 1.5, \quad \forall u^1 \in [0, 1],$$

with given $\alpha, \beta \in \mathbb{R}_+$.

Find an optimal policy and find the optimal cost.

Hint: Consider deterministic and stationary policies and analyze the costs corresponding to such policies.

Exercise 7.7.6 *[Machine repair revisited] Recall Exercise 7.7.6 with an average cost formulation. Show that there exists an optimal control policy and that this policy is stationary.*

Exercise 7.7.7 *For the model considered in Section 7.2, under Assumption 7.2.2, establish that $\mathbb{T}_z)$ is a contraction on the space of bounded functions h with $h(z) = 0$, with modulus α.*

Exercise 7.7.8 (Risk-sensitive average cost criterion) *Let $\mathbb{X}, \mathbb{U}$ be finite and consider the following risk-sensitive criterion:*

$$\lambda^* = \inf_{\gamma\in\Gamma_A} \limsup_{N\to\infty} \frac{1}{N} \log \left(E^\gamma \left[e^{\sum_{m=0}^{N-1} c(x_m, u_m)} \right] \right)$$

For this criterion, show that the verification theorem (dynamic programming equation) satisfies [67]:

$$\lambda^* V(x) = \min_{u\in\mathbb{U}} \left(e^{c(x,u)} \sum_{x_1} \mathcal{T}(x_1|x_0 = x, u_0 = u) V(x_1) \right)$$

If one views the operator on the right as a function of V, we have an eigenvalue problem with λ^* being the optimal cost. Hint. Divide both sides by λ^*, and apply iteratively the inequality under an arbitrary admissible control. Show that equality holds when an optimal policy is considered.

Exercise 7.7.9 *Study [266] as an example where an average cost optimal policy may not be stationary (and even possibly randomized stationary).*

Exercise 7.7.10 *In some problems one needs to relate the discounted cost problem, a finite horizon average cost problem and an infinite horizon average cost problem. Under what conditions do we have that*

$$\lim_{\beta\to 1} (1 - \beta) J_\beta(x_0) \to J^*(x_0)?,$$

and with J^T as given in (7.9),

$$\lim_{T\to\infty} \inf_{\gamma\in\Gamma_A} J^T(x_0, \gamma) \to J^*(x_0)?$$

Exercise 7.7.11 *a) For an infinite horizon discounted cost partially observed Markov decision problem with finite state, action and measurement spaces, suppose that we wish to restrict the policies to be stationary control policies which only are based on the most recent observation; that is $u_t = \gamma(y_t)$ for some $\gamma : \mathbb{Y} \to \mathbb{U}$ (clearly, this is suboptimal among all admissible policies, as the analysis in the Chapter shows). Given this restrictive class of policies, can one obtain an optimal policy through linear programming? b) Can you consider a setup where an optimal policy above may not be optimal among all policies (e.g., an optimal one-memory policy may not be stationary)? Hint: Consider linear systems theory (stationary output feedback vs. time-varying output feedback).*

8

Numerical and Approximation Methods

In this chapter, we will first study several computational algorithms, first in the context of finite space MDPs. After this, we will study rigorous approximation methods for continuous (standard Borel) space MDPs and POMDPs.

8.1 Value and Policy Iteration Algorithms

8.1.1 Value Iteration

Consider expected discounted cost criterion, for some $\beta \in (0,1)$,

$$J_\beta(x_0,\gamma) = E^\gamma_{x_0}[\sum_{t=0}^{\infty} \beta^t c(x_t,u_t)], \tag{8.1}$$

to be minimized. The *Value Iteration Algorithm* was presented earlier in Theorems 5.5.1 and 5.5.2; the algorithm for the bounded setup is re-stated in the following.

Theorem 8.1.1 *Suppose the cost function c is bounded, non-negative, and one of the measurable selection conditions (Assumption 5.2.1 or Assumption 5.2.2) applies. Then, there exists a unique solution to the discounted cost optimality equation*

$$v(x) = \min_{u\in\mathbb{U}} \left(c(x,u) + \beta \int_{\mathbb{X}} v(y)\mathcal{T}(dy|x,u) \right), \qquad x \in \mathbb{X}$$

Furthermore, the optimal cost (value function) is obtained by a successive iteration of policies (known as the Value Iteration Algorithm):

$$v_n(x) = \min_{u\in\mathbb{U}} \left(c(x,u) + \beta \int_{\mathbb{X}} v_{n-1}(y)\mathcal{T}(dy|x,u) \right), \qquad \forall x, n \in \mathbb{N} \tag{8.2}$$

For any $v_0 \in L_\infty(\mathbb{X})$ (or $C_b(\mathbb{X})$ under measurable selection Condition 1 in Assumption 5.2.1), the sequence converges to a unique fixed point. If $v_0(x) = 0$ for all $x \in \mathbb{X}$, then $v_n(x) \uparrow v(x)$ for all $x \in \mathbb{X}$. Under the measurable selection Assumption 5.2.1, the limit is also continuous with $v_0 \in C_b(\mathbb{X})$.

We also recall that under Assumptions 5.2.1 and 8.2.5, the value function is Lipschitz; see Theorem 5.5.3.

8.1.2 Policy Iteration

We now discuss the *Policy Iteration Algorithm*. Let $\mathbb{X}$ be countable and c be a bounded cost function. Consider again (8.1). Let $\gamma_0 := \{\gamma_0, \gamma_0, \gamma_0, \cdots, \gamma_0, \cdots\} \in \Gamma_S$ denote a deterministic stationary policy (which naturally leads to a

finite discounted expected cost here). Let this expected cost be $W_0(\cdot)$; that is,

$$W_0(x) = E_x^{\gamma_0}[\sum_{k=0}^{\infty} \beta^k c(x_k, \gamma_0(x_k))], \qquad x \in \mathbb{X}$$

Then, similar to (5.23), and using the stationarity of γ_0, we obtain

$$W_0(x) = c(x, \gamma_0(x)) + \beta \sum_{x'} W_0(x') P(x_{t+1} = x' | x_t = x, u_t = \gamma_0(x))$$

Let

$$\mathbb{T}(W_0)(x) = \min_{u \in \mathbb{U}} \left(c(x,u) + \beta \sum_{x'} W_0(x') P(x_{t+1} = x' | x_t = x, u_t = u) \right).$$

Clearly $\mathbb{T}(W_0) \leq W_0$ pointwise (in x). Now, let γ_1 be such that

$$\mathbb{T}(W_0)(x) = c(x, \gamma_1(x)) + \beta \sum_{x'} W_0(x') P(x_1 = x' | x_0 = x, u_0 = \gamma_1(x)) \tag{8.3}$$

Observe that, iterative application of (8.3) one more time (to $W_0(x')$ above with $\mathbb{T}(W_0)(x') \leq W_0(x')$) leads to

$$W_0(x) \geq E_x^{\gamma_1} \left[\sum_{k=0}^{1} \beta^k c(x_k, \gamma_1(x_k)) + \beta^2 W_0(x_2) \right].$$

and, continuing further, for any $x \in \mathbb{X}, n \in \mathbb{Z}_+$, we arrive at

$$W_0(x) \geq E_x^{\gamma_1}[\sum_{k=0}^{n-1} \beta^k c(x_k, \gamma_1(x_k)) + \beta^n W_0(x_n)]. \tag{8.4}$$

Taking the limit $n \to \infty$, this leads to the relation

$$W_0(x) \geq \mathbb{T}(W_0)(x) \geq W_1(x),$$

with

$$W_1(x) = E_x^{\gamma_1}[\sum_{k=0}^{\infty} \beta^k c(x_k, \gamma_1(x_k))]$$

so that

$$W_1(x) := c(x, \gamma_1(x)) + \beta \sum_{x'} W_1(x') P(x_{t+1} = x' | x_t = x, u_t = \gamma_1(x)).$$

We can interpret the steps of the discussion above as follows: We start with the policy, $\{\gamma_0, \gamma_0, \gamma_0, \cdots\}$ and then make the point that the policy $\{\gamma_1, \gamma_0, \gamma_0, \cdots\}$ is a better one and then $\{\gamma_1, \gamma_1, \gamma_0, \cdots\}$ is a better one and ultimately the policy $\{\gamma_1, \gamma_1, \gamma_1, \cdots, \gamma_1, \cdots\}$ is a better policy than what we started with.

We then continue this procedure for $m = 2$ by replacing W_1 with W_0 above, to arrive at

$$\begin{aligned} \mathbb{T}(W_1)(x) &= \min_{u \in \mathbb{U}} \left(c(x,u) + \beta \sum_{x'} W_1(x') P(x_{t+1} = x' | x_t = x, u_t = u) \right) \\ &= c(x, \gamma_2(x)) + \beta \sum_{x'} W_1(x') P(x_1 = x' | x_0 = x, u_0 = \gamma_2(x)) \end{aligned} \tag{8.5}$$

and ultimately $W_2(x) \leq \mathbb{T}(W_1)(x) \leq W_1(x)$, where

$$W_2(x) = E_x^{\gamma_2}[\sum_{k=0}^{\infty} \beta^k c(x_k, \gamma_1(x_k))]$$

Then, we repeat the process for $m > 2$ with

$$\begin{aligned}\mathbb{T}(W_m)(x) &= \min_{u\in\mathbb{U}} \left(c(x,u) + \beta \sum_{x'} W_m(x') P(x_{t+1} = x' | x_t = x, u_t = u) \right) \\ &= c(x, \gamma_{m+1}(x)) + \beta \sum_{x'} W_m(x') P(x_1 = x' | x_0 = x, u_0 = \gamma_{m+1}(x)) \end{aligned} \tag{8.6}$$

and ultimately $W_{m+1}(x) \leq \mathbb{T}(W_m)(x) \leq W_m(x)$, where

$$W_{m+1}(x) = E_x^{\gamma_{m+1}}[\sum_{k=0}^{\infty} \beta^k c(x_k, \gamma_{m+1}(x_k))]$$

Remark 8.1. Note that for the case where $\mathbb{X}, \mathbb{U}$ are finite, the following holds

$$W_1(x) = E_x^{\gamma_1}[\sum_{k=0}^{\infty} \beta^k c(x_k, \gamma_1(x_k))] = c(x, \gamma_1(x)) + \beta \sum W_1(x_{t+1}) \mathcal{T}(dx_{t+1} | x_t = x, u_t = \gamma_1(x))]$$

More generally, for a given stationary policy γ:

$$J_\beta(x, \gamma) = E_x^{\gamma}[\sum_{k=0}^{\infty} \beta^k c(x_k, \gamma(x_k))] = c(x, \gamma(x)) + \beta \sum_{x'} J_\beta(x', \gamma) P(x_{t+1} = x' | x_t = x, u_t = \gamma(x)) \tag{8.7}$$

can be computed by solving the following matrix equation

$$W = c_\gamma + \beta P^\gamma W,$$

leading to

$$W = (I - \beta P^\gamma)^{-1} c_\gamma,$$

where W is a column vector consisting of $\{W(x), x \in \mathbb{X}\}$; c_γ is a column vector consisting of elements $\{c(x, \gamma(x)), x \in \mathbb{X}\}$; and P^γ is a stochastic matrix with entries $P^\gamma(x, x') = P(x_{t+1} = x' | x_t = x, u_t = \gamma(x))$ (note that $(I - \beta P^\gamma)$ is always invertible for $\beta \in (0, 1)$). Thus, the implementation of the policy iteration algorithm is quite simple.

Theorem 8.1.2 *Through the policy iteration algorithm, there exists $W : \mathbb{X} \to \mathbb{R}$ such that $W_n \downarrow W$ pointwise in x, provided that for some $n \in \mathbb{N}$, $W_n(x) < \infty$ for $x \in \mathbb{X}$. If $\gamma = \{f, f, \cdots\}$ satisfies*

$$E_x^{\gamma}[\sum_{k=0}^{\infty} \beta^k c(x_k, f(x_k))] = W(x), \tag{8.8}$$

then γ is optimal among all stationary policies. And if (5.29) holds with W replacing v, γ is optimal among all policies (note that this always holds if c is bounded). For a problem with finite state and action spaces, convergence is guaranteed in a finite number of stages and the resulting policy is optimal.

Proof. By (8.4) the sequence $W_n \geq \mathbb{T}(W_n) \geq W_{n+1}$, and thus there is a limit W (since the cost per state is bounded from below) and the limit satisfies $W = \mathbb{T}(W)$, which is precisely the optimality equation (5.28). Since such a W leads to a lower bound under any stationary policy by the construction of the algorithm, an argument similar to the one in the proof of Lemma 5.5.4 leads to the result. Note that for finite models we have the condition, as noted in Lemma 5.5.4,

$$\lim_{t\to\infty} \beta^t E_x^{\gamma}[W(x_t)] = 0,$$

by (8.8), since W is bounded. ◇

8.1.3 Receding Horizon Algorithms / Model Predictive Control

Roll-out algorithms, also known as sliding-horizon or receding horizon algorithms, are practically important. Such an algorithm is provably near-optimal as the horizon length increases under some conditions. We refer the reader to [164], [79], [97] and [40] among many other papers in this direction.

8.2 Approximation through Quantization of the State and the Action Spaces

For cases where the spaces are not finite or countable, numerical methods require approximation procedures. In the following, let $\mathbb{X}, \mathbb{U}$ be standard Borel spaces with $\mathbb{U}(x) = \mathbb{U}$ compact. We will arrive at rigorously justified approximation algorithms and resulting convergence results, both via analytical as well as learning theoretic methods.

8.2.1 Finite Action Approximation to MDPs

Definition 8.2. *A measurable function $q : \mathbb{X} \to \mathbb{U}$ is called a* quantizer *from $\mathbb{X}$ to $\mathbb{U}$ if the range of q, i.e., $q(\mathbb{X}) = \{q(x) \in \mathbb{U} : x \in \mathbb{X}\}$, is finite.*

The elements of $q(\mathbb{X})$ (the possible values of q) are called the *levels* of q.

Finite Action Approximate MDP: Quantization of the Action Space

Let $d_{\mathbb{U}}$ denote the metric on $\mathbb{U}$. Since the action space $\mathbb{U}$ is compact and thus totally bounded, one can find a sequence of finite sets $\Lambda_n = \{a_{n,1}, \ldots, a_{n,k_n}\} \subset \mathbb{U}$ such that for all n,

$$\min_{i \in \{1,\ldots,k_n\}} d_{\mathbb{U}}(a, a_{n,i}) < \frac{1}{n} \text{ for all } a \in \mathbb{U}.$$

In other words, Λ_n is a $1/n$-net in $\mathbb{U}$. In the rest of Section 8.2.1, we assume that the sequence $\{\Lambda_n\}_{n\geq 1}$ is fixed. To ease the notation in the sequel, let us define the mapping

$$\Upsilon_n(f)(x) := \arg\min_{a \in \Lambda_n} d_{\mathbb{U}}(f(x), a), \tag{8.9}$$

where ties are broken so that $\Upsilon_n(f)(x)$ is measurable.

Our main objective in this section is to find conditions on the components of the MDP under which there exists a sequence of finite subsets $\{\Lambda_n\}_{n\geq 1}$ of $\mathbb{U}$ for which the following holds:

(P) If for each n, MDP$_n$ is defined as the Markov decision process having the components $\{\mathbb{X}, \Lambda_n, p, c\}$, then we would like to find conditions under which the value function of MDP$_n$ converges to the value function of the original MDP as $n \to \infty$.

Near optimality of quantized policies under weak continuity

Consider **(P)** for MDPs with weakly continuous transition probability.

Recall Assumption 5.2.1, essentially repeated for convenience of the reader:

Assumption 8.2.1 *(a) The one stage cost function c is bounded and continuous.*

(b) The stochastic kernel $\mathcal{T}(\,\cdot\,|x, a)$ is weakly continuous in $(x, a) \in \mathbb{X} \times \mathbb{U}$.

(c) $\mathbb{U}$ *is compact.*

For any real-valued measurable function u on $\mathbb{X}$, let $\mathbb{T}$ be given by

$$(\mathbb{T}v)(x) := \min_{u \in \mathbb{U}} \left(c(x,u) + \beta \int_{\mathbb{X}} v(y) \mathcal{T}(dy|x,u) \right). \tag{8.10}$$

Recall that here, $\mathbb{T}$ is the *Bellman discounted cost optimality operator* for the MDP considered earlier in (5.28). Analogously, let us define the Bellman optimality operator $\mathbb{T}_n$ of MDP$_n$ as

$$\mathbb{T}_n v(x) := \min_{u \in \Lambda_n} \left(c(x,u) + \beta \int_{\mathbb{X}} v(y) \mathcal{T}(dy|x,u) \right). \tag{8.11}$$

We have seen that both $\mathbb{T}$ and $\mathbb{T}_n$ are contraction operators. Furthermore, value functions of MDP and MDP$_n$ are fixed points of these operators; that is, $\mathbb{T}J^* = J^*$ and $\mathbb{T}_n J_n^* = J_n^*$. Let us define $v^0 = v_n^0 \equiv 0$, and $v^{t+1} = \mathbb{T}v^t$ and $v_n^{t+1} = \mathbb{T}_n v_n^t$ for $t \geq 1$; that is, $\{v^t\}_{t \geq 1}$ and $\{v_n^t\}_{t \geq 1}$ are successive approximations to the discounted value functions of the MDP and MDP$_n$, respectively (via value iteration). The following can be shown, inductively for each $t = 0, 1, 2, \cdots$:

Lemma 8.2.1 *[270, Lemma 3.19] Under Assumption 8.2.1, for any compact* $K \subset X$ *and for any* $t \geq 1$*, we have*

$$\lim_{n \to \infty} \sup_{x \in K} |v_n^t(x) - v^t(x)| = 0. \tag{8.12}$$

The following theorem states that the optimal value function of MDP$_n$ converges to the optimal value function of the original MDP. It can be proved by using Lemma 8.2.1 and taking into account that $\{v^t\}_{t \geq 1}$ and $\{v_n^t\}_{t \geq 1}$ are successive approximations to the value functions J_β^* and $J_{\beta,n}^*$, respectively.

Theorem 8.2.1 *[273] [270, Theorem 3.16] Under Assumption 8.2.1, for any compact* $K \subset X$*, we have*

$$\lim_{n \to \infty} \sup_{x \in K} |J_{\beta,n}^*(x) - J_\beta^*(x)| = 0. \tag{8.13}$$

The proof follows from a successive approximation argument applied iteratively for value iteration updates; see the proof of Theorem 12.3.2 for an explicit analysis and Section 12.5.2 for further relations between the problem considered here and the *robustness* problem considered there, where the approximation problem here is viewed as a particular instance of robustness.

We state an approximation result analogous to Theorem 8.2.1 for the average cost criterion. For average cost criteria, we impose relatively stronger ergodicity/minorization conditions on the controlled Markov chain. Note that this was utilized to establish the existence of a solution to the average cost optimality equation in Section 7.2 (see Assumption 7.2.2).

Suppose that Assumption 5.2.1 and Assumption 7.2.2 hold. This implies that, as we have seen earlier in Section 7.3, there is a solution to the average cost optimality equation (ACOE) and the stationary policy which minimizes this ACOE is an optimal policy.

Theorem 8.2.2 *[Average Cost] [273], [270, Theorem 3.22] Under Assumptions 5.2.1 and 7.2.2, the value functions (that is, the optimal expected average cost) satisfy*

$$\lim_{n \to \infty} |V_n^* - V^*| = 0,$$

where V^* *and* V_n^* *(*$n \geq 1$*) (the value functions of the true model and the approximate model sequence, respectively) do not depend on* x*.*

Remark 8.3. As we have observed earlier, when one considers partially observed MDPs (POMDPs), any POMDP can be reduced to a (completely observable) MDP whose states are the posterior state distributions or beliefs of the observer. Thus the results in this section are applicable to POMDPs as we will study in Section 9.2.1.

Near optimality of quantized policies under strong continuity in actions for each state

Consider the problem **(P)** for MDPs with strongly continuous transition probabilities in actions. Recall Assumption 5.2.2, repeated for reader's convenience:

Assumption 8.2.2

(a) The one stage cost function c is nonnegative and bounded satisfying $c(x, \cdot) \in C_b(\mathbb{U})$ for all $x \in \mathbb{X}$.

(b) The stochastic kernel $\mathcal{T}(\cdot|x,u)$ is setwise continuous in $u \in \mathbb{U}$ for every $x \in \mathbb{X}$.

(c) $\mathbb{U}$ is compact.

The following theorem states that for any $f \in \mathbb{F}$, the discounted cost function of $\Upsilon_n(f)$ converges to the discounted cost function of f as $n \to \infty$. Therefore, it implies that the discounted value function of the MDP$_n$ converges to the discounted value function of the original MDP.

Theorem 8.2.3 *[275][Discounted Cost] Consider problem **(P)** for the discounted cost under Assumption 8.2.2. Then, for any stationary policy defined with $f : \mathbb{X} \to \mathbb{U}$, $J(\Upsilon_n(f), x) \to J(f,x)$ as $n \to \infty$, for all $x \in \mathbb{X}$.*

Observe that any deterministic stationary policy f defines a stochastic kernel on $\mathbb{X}$ given $\mathbb{X}$ via

$$Q_f(\cdot|x) := \mathcal{T}(\cdot|x, f(x)). \tag{8.14}$$

Let Q_f^t denote the t-step transition probability of this Markov chain. If Q_f admits a unique invariant probability measure ν_f, then by [168, Theorem 2.3.4 and Proposition 2.4.2], there exists an invariant set $\mathbb{M}_f \in \mathcal{B}(\mathbb{X})$ with full ν_f measure such that for all x in that set we have

$$V(f,x) = \int_{\mathbb{X}} c(x, f(x)) \nu_f(dx). \tag{8.15}$$

Assumption 8.2.3 *Suppose Assumption 8.2.2 holds. In addition, we have*

(d) For any $f \in \mathbb{F}$, Q_f has a unique invariant probability measure ν_f.

(e1) The set of invariant probability measures $\Gamma_{\mathbb{F}} := \{\nu \in \mathcal{P}(\mathbb{X}) : \nu Q_f = \nu \text{ for some } f \in \mathbb{F}\}$ is relatively sequentially compact in the setwise topology.

(e2) There exists $x \in \mathbb{X}$ such that for all $B \in \mathcal{B}(\mathbb{X})$, $Q_f^t(B|x) \to \nu_f(B)$ uniformly in $f \in \mathbb{F}$.

(f) $\mathbb{M} := \bigcap_{f \in \mathbb{F}} \mathbb{M}_f \neq \emptyset$.

We note that the condition (e2) above may be slightly relaxed [347].

The following theorem states that for any $f \in \mathbb{F}$, the average cost function of $\Upsilon_n(f) \in \mathcal{Q}(\Lambda_n)$ converges to the average cost function of f as $n \to \infty$. In particular, the average value function of MDP$_n$ converges to the average value function of the original MDP.

Theorem 8.2.4 *[275][Average Cost] Let $x \in \mathbb{M}$ and $f \in \mathbb{F}$. Then, we have $V(\Upsilon_n(f), x) \to V(f,x)$ as $n \to \infty$, under Assumption 8.2.3 with either (e1) or (e2).*

One can also obtain rates of convergence results [275] [270].

8.2.2 Finite State Approximation to MDPs

In this section we study the finite-state approximation problem for MDPs, by reducing them to finite state MDPs obtained through quantization of the state space on a finite grid following [275] [270, Chapter 4]. Here two questions could be posed:

(i) **Q1** Under what conditions on the components of the MDP do the true cost functions of the policies obtained from finite models converge to the optimal value function as the number of grid points goes to infinity?

(ii) **Q2** Can we obtain bounds on the performance loss due to discretization in terms of the number of grid points if we strengthen the conditions sufficient in **(Q1)**?

We will not discuss **Q2** here, but address **Q1**. For **Q2**, the reader is referred to [270, 275]. We note that, under further explicit regularity conditions, one can indeed arrive at rates of convergence to optimality. The approach to solve problem **(Q1)** can be summarized as follows: First, we obtain approximation results for the compact-state case. We find conditions under which a compact representation leads to near optimality for non-compact state MDPs: solve the approximate MDP, and apply the optimal solution for the approximate MDP to the original MDP. We then obtain the convergence of the finite-state models to non-compact models. Consider **(Q1)** for MDPs with compact state space.

We now establish near optimality under finite state approximations. We start by choosing a collection of disjoint sets $\{B_i\}_{i=1}^M$ such that $\bigcup_i B_i = \mathbb{X}$, and $B_i \bigcap B_j = \emptyset$ for any $i \neq j$. Furthermore, we choose a representative state, $y_i \in B_i$, for each disjoint set. For this setting, we denote the new finite state space by $\mathbb{Y} := \{y_1, \ldots, y_M\}$, and the mapping from the original state space $\mathbb{X}$ to the finite set $\mathbb{Y}$ is done via

$$q(x) = y_i \quad \text{if } x \in B_i. \tag{8.16}$$

Furthermore, we choose a weighting measure $\pi^* \in \mathcal{P}(\mathbb{X})$ on $\mathbb{X}$ such that $\pi^*(B_i) > 0$ for all B_i. We now define normalized measures using the weight measure on each separate quantization bin B_i as follows:

$$\hat{\pi}^*_{y_i}(A) := \frac{\pi^*(A)}{\pi^*(B_i)}, \quad \forall A \subset B_i, \quad \forall i \in \{1, \ldots, M\}, \tag{8.17}$$

that is, $\hat{\pi}^*_{y_i}$ is the normalized weight measure on the set B_i, where y_i belongs to.

We now define the stage-wise cost and transition kernel for the MDP with this finite state space $\mathbb{Y}$ using the normalized weight measures. Indeed, for any $y_i, y_j \in \mathbb{Y}$ and $u \in \mathbb{U}$, the stage-wise cost and the transition kernel for the finite-state model are defined as

$$\begin{aligned} C^*(y_i, u) &= \int_{B_i} c(x, u)\, \hat{\pi}^*_{y_i}(dx), \\ P^*(y_j | y_i, u) &= \int_{B_i} \mathcal{T}(B_j | x, u)\, \hat{\pi}^*_{y_i}(dx). \end{aligned} \tag{8.18}$$

Having defined the finite state space $\mathbb{Y}$, the cost function C^* and the transition kernel P^*, we can now introduce the optimal value function for this finite model. We denote the optimal value function which is defined on $\mathbb{Y}$ by $\hat{J}_\beta : \mathbb{Y} \to \mathbb{R}$. Note that $\hat{J}_\beta$ satisfies the following DCOE for any $y \in \mathbb{Y}$

$$\hat{J}_\beta(y) = \inf_{u \in \mathbb{U}} \left\{ C^*(y, u) + \beta \sum_{z \in \mathbb{Y}} \hat{J}_\beta(z) P^*(z | y, u) \right\}. \tag{8.19}$$

Note that we can easily extend this function over the original state space $\mathbb{X}$ by making it constant over the quantization bins. In other words, if $y \in B_i$, then for any $x \in B_i$, we write

$$\hat{J}_\beta(x) := \hat{J}_\beta(y).$$

We further define an average loss function $L : \mathbb{X} \to \mathbb{R}$ as a result of the quantization. For some $x \in \mathbb{X}$, where x belongs to a quantization bin B_i whose representative state is y_i (i.e. $q(x) = y_i$), the average loss function $L(x)$ is defined as

$$L(x) := \int_{B_i} \|x - x'\| \, \hat{\pi}^*_{y_i}(dx') \qquad \forall x \in B_i, i = 1, \cdots, M. \tag{8.20}$$

That is, $L(x)$ can be seen as the distance of the state x to the mean of the bin B_i under the measure $\hat{\pi}^*_{y_i}$.

In the following, we present error analyses in finite state approximations defined in this section.

Finite State Approximations with Kernels Continuous in Total Variation under Expected Quantization Error Bounds

In this section, we focus on an MDP model whose transition kernel is Lipschitz continuous in x (uniform in u) under the total variation norm. This condition is somewhat different than the continuity of the transition kernel under the total variation distance. Indeed, if we have a model as in Example 5.6-(ii), then we have the required Lipschitz continuity of the transition kernel when f is Lipschitz continuous in x that is uniform in u and the density of the noise w is Lipschitz continuous. The following assumptions are imposed on the system.

Assumption 8.2.4 *(a) There exists a constant $\alpha_c > 0$ such that $|c(x,u) - c(x',u)| \leq \alpha_c \|x - x'\|$ for all $x, x' \in \mathbb{X}$ and for all $u \in \mathbb{U}$.*

(b) There exists a constant $\alpha_T > 0$ such that $\|\mathcal{T}(\cdot|x,u) - \mathcal{T}(\cdot|x',u)\|_{TV} \leq \alpha_T \|x - x'\|$ for all $x, x' \in \mathbb{X}$ and for all $u \in \mathbb{U}$.

The first result gives an error bound for the approximate value function.

Theorem 8.2.5 *[185, Theorem 4] Under Assumption 8.2.4, provided that c is bounded, we have for any initial state $x_0 \in \mathbb{X}$*

$$\left| \hat{J}_\beta(x_0) - J^*_\beta(x_0) \right| \leq \left(\alpha_c + \frac{\beta \alpha_T \|c\|_\infty}{1-\beta} \right) \sum_{t=0}^{\infty} \beta^t \sup_{\gamma \in \Gamma} E^{\mathcal{T},\gamma}_{x_0} \left[L(X_t) \right],$$

where L is defined in (8.20).

The following result provides an error bound for the approximate policy of the finite-state model when it is applied to the original model.

Theorem 8.2.6 *Under Assumption 8.2.4, we have for any initial state $x_0 \in \mathbb{X}$*

$$\left| J_\beta(x_0, \hat{\gamma}) - J^*_\beta(x_0) \right| \leq 2 \left(\alpha_c + \frac{\beta \alpha_T \|c\|_\infty}{1-\beta} \right) \sum_{t=0}^{\infty} \beta^t \sup_{\gamma \in \Gamma} E^{\mathcal{T},\gamma}_{x_0} \left[L(X_t) \right]$$

where L is defined in (8.20) and $\hat{\gamma}$ denotes the optimal policy of the finite-state approximate model given by (8.18) extended to the state space $\mathbb{X}$ via the quantization function q.

Finite State Approximation with Kernels Continuous in Wasserstein Distance under Uniform Quantization Error Bounds

In this section, we focus on the models with transition kernels that are Lipschitz continuous in x (uniform in u) under the first order Wasserstein distance. If we have a model as in Example 5.6-(i), then we have the required Lipschitz continuity of the transition kernel when f is Lipschitz continuous in x that is uniform in u. Here, instead of providing an average loss bound using (8.20) as in Theorem 8.2.5, we will provide a uniform loss bound result, and we also assume the state space to be compact. We first define

$$\bar{L} := \max_{i=1,\dots,M} \sup_{x,x' \in B_i} \|x - x'\|. \tag{8.21}$$

Here, $\bar{L}$ is the largest diameter among the quantization bins.

The following is essentially Assumption 5.5.1, re-stated for the setup considered.

Assumption 8.2.5 *(a) $\mathbb{X}$ is compact.*

(b) There exists a constant $\alpha_c > 0$ such that $|c(x,u) - c(x',u)| \leq \alpha_c \|x - x'\|$ for all $x, x' \in \mathbb{X}$ and for all $u \in \mathbb{U}$.

(c) There exists a constant $\alpha_T > 0$ such that $W_1(\mathcal{T}(\cdot|x,u), \mathcal{T}(\cdot|x',u)) \leq \alpha_T \|x - x'\|$ for all $x, x' \in \mathbb{X}$ and for all $u \in \mathbb{U}$.

Theorem 8.2.7 *[185, Theorem 5] Let Assumption 8.2.5 hold and let $\mathbb{X}$ be compact. We have*

$$\sup_{x_0 \in \mathbb{X}} \left| \hat{J}_\beta(x_0) - J^*_\beta(x_0) \right| \leq \frac{\alpha_c}{(1 - \beta\alpha_T)(1-\beta)} \bar{L}$$

where $\bar{L}$ is defined in (8.21).

The following result is similar to [270, Theorem 4.38] with a slightly improved bound.

Theorem 8.2.8 *[185, Theorem 6] Let Assumption 8.2.5 hold and let $\mathbb{X}$ be compact. We have*

$$\sup_{x_0 \in X} \left| J_\beta(x_0, \hat{\gamma}) - J^*_\beta(x_0) \right| \leq \frac{2\alpha_c}{(1-\beta)^2(1 - \beta\alpha_T)} \bar{L}.$$

where $\bar{L}$ is defined in (8.21) and $\hat{\gamma}$ denotes the optimal policy of the finite-state approximate model extended to the state space X via the quantization function q.

Finite State Approximation with Weakly Continuous Kernels and Asymptotic Convergence

In this section, we assume that $\mathbb{X}$ is σ-compact. That is, we can write $\mathbb{X} = \cup_{k=1}^{\infty} B_k$ where each B_k is compact. A finite dimensional Euclidean space is an example of such a space. Additionally, in this section, we focus on the models with transition kernels that are continuous only under the weak convergence topology. Here, instead of providing a rate of convergence, we will provide an asymptotic result. Let the quantizer be such that the M^{th} bin be the over-flow bin; that is, the first $M-1$ bins be the quantization of a compact set and the complement be assigned to B_M. Let $d_{\mathbb{X}}$ denote the metric on $\mathbb{X}$

To this end, let us define

$$L^- := \max_{i=1,\dots,M-1} \sup_{x,x' \in B_i} d_{\mathbb{X}}(x, x'). \tag{8.22}$$

Note that since $\mathbb{X}$ is σ-compact, for each M, one can find a partition $\{B_i\}_{i=1}^M$ of the state space $\mathbb{X}$ such that $L^- \to 0$ and $\bigcup_{i=1}^{M-1} B_i \nearrow \mathbb{X}$ as $M \to \infty$. Note that $B_M = \mathbb{X} \setminus (\cup_{i=1}^{M-1} B_i)$. In the following result, we assume that such a sequence of partitions is used to obtain the finite-state approximate models.

Theorem 8.2.9 *[270, Theorem 4.27] Under Assumption 8.2.1, we have for any compact $K \subset \mathbb{X}$*

$$\sup_{x_0 \in K} \left| \hat{J}_\beta(x_0) - J^*_\beta(x_0) \right| \to 0$$

and

$$\sup_{x_0 \in K} \left| J_\beta(x_0, \hat{\gamma}) - J^*_\beta(x_0) \right| \to 0$$

as $L^- \to 0$, where $\hat{\gamma}$ denotes the optimal policy of the finite-state approximate model extended to the state space $\mathbb{X}$ via the quantization function q.

We note that the result by [270, Theorem 4.27] is more general and applicable to unbounded cost functions as well. Under the bounded cost in Assumption 8.2.1, [270, Theorem 4.27] implies Theorem 8.2.9 above.

One challenge to be addressed in the proofs of the results noted above is that in the quantized models (as an intermediate step in the proof) we do not have the weak continuity condition for each of the quantized models: The issue is that the value function in the dynamic programming update iterations is not continuous (and would only be piece-wise continuous), and accordingly the finite models are not necessarily continuous in the action variables which violates the measurable selection conditions noted in Section 5.2. Nonetheless, the machinery of universally measurable policies (see Appendix C) can be utilized and the existence of optimal policies for the approximate kernels does not arise as an immediate problem (see [275, p. 6-7]) for the proof of the theorem. Alternatively, one can first quantize the action set and work on the approximate (finite-action) MDP, whose near optimality was established earlier. Note that for the finite action setup, continuity of the kernels in actions always holds. Accordingly, we assume that the action set is finite in the following analysis.

The Average Cost Case

Theorem 8.2.10 *[270] [182, Theorem 4.2] [Average Cost] Suppose that Assumption 5.2.1 and Assumption 7.2.2 hold. Then,*

$$\lim_{m \to \infty} \| J(\cdot, \hat{\gamma}) - J^* \| = 0. \tag{8.23}$$

where $\hat{\gamma}$ denotes the optimal policy of the finite-state approximate model extended to the state space $\mathbb{X}$ via the quantization function q with diameter $\frac{1}{m}$.

One can also obtain explicit rates of convergence under further regularity, similar to the discounted cost setting studied earlier.

Remark 8.4. We note that [270] imposes total variation continuity, but building on [183, Theorem 16] and adapting the arguments in the proof of [182, Theorem 4.2] (see Section 12.4), the total variation condition continuity condition on the kernel can be relaxed to weak Feller continuity, leading to the above. This will be studied in *Chapter* 12.

8.2.3 Finite Model MDP Approximation: Quantization of both the State and Action Spaces

We showed in Theorems 8.2.3 and 8.2.4 that any MDP with (infinite) compact action space and with bounded one-stage cost function can be well approximated by an MDP with finite action space, for both the discounted cost and the average cost cases.

Therefore, before discretizing the state space to compute near optimal policies, one can discretize, without loss of generality, the action space $\mathbb{U}$ in advance on a finite grid using sufficiently large number of grid points. Then, near optimality of finite models follow from the discussions in Section 8.2.2.

Remark 8.5 (An alternative direct argument via the convex analytic method for average cost criteria). For average cost criteria, an alternative argument is presented in [16, Theorem 4.2], where it is shown that under either weak or setwise continuity conditions, both finite state and action models are near optimal by considering the set of invariant occupation measures under unique ergodicity conditions. Furthermore, under the conditions noted, deterministic and stationary policies are shown to be dense among those that are randomized and stationary, in the sense that the cost

under any randomized stationary policy can be approximated arbitrarily well by deterministic and stationary policies. Furthermore, the dense set of deterministic and stationary policies can be assumed to have finite range. A further utility of this approach is that geometric ergodicity, e.g. via minorization, is not needed.

8.3 Numerical Methods for POMDPs

8.3.1 Near optimality of quantized policies under weak Feller or Wasserstein regularity of non-linear filters

As we have seen, any POMDP can be reduced to a completely observable Markov process [352], [262], whose states are the posterior state distributions or 'beliefs' of the observer; that is, the state at time t is

$$\pi_t(\,\cdot\,) := P\{X_t \in \,\cdot\,|y_0, \ldots, y_t, u_0, \ldots, u_{t-1}\} \in \mathcal{P}(\mathbb{X}).$$

We called this conditional probability measure process the *filter* process. The filter process has state space $\mathcal{P}(\mathbb{X})$ and action space $\mathbb{U}$. Here, $\mathcal{P}(\mathbb{X})$ is equipped with the Borel σ-algebra generated by the topology of weak convergence. The transition probability of the filter process is given in (6.27).

Accordingly, we have a fully observed belief-MDP. Now, by combining the approximation results in Section 8.2 and reinforcement learning theoretic results to be presented in Section 9.3, together with the the weak Feller continuity results presented in Section 6.3.2, we can conclude that the numerical methods can also be applied to POMDPs under the conditions reported in Theorems 6.3.3 and 6.3.4 [128] [184].

This has explicitly been demonstrated in [276, Theorem 3], where also methods for quantizing probability measures have been studied. This also applies under Wasserstein regularity of non-linear filter kernels, with explicit conditions given in Theorem 6.3.5 [100]; see Theorem 8.2.7.

Accordingly, due to the weak Feller property of controlled non-linear filters, we can obtain rigorous approximation results by quantizing probability measures.

In the following, we present an alternative approach.

8.3.2 Near-optimality of finite window policies under filter stability

One can also show that under filter stability (see Section 6.4), finite window policies are near optimal [188]. Consider the following:

$$J_\beta(\mu, \mathcal{T}, \gamma) = E_\mu^{\mathcal{T},\gamma}\left[\sum_{t=0}^{\infty} \beta^t c(x_t, u_t)\right], \qquad J_\beta^*(\mu, \mathcal{T}) = \inf_{\gamma \in \Gamma} J_\beta(\mu, \mathcal{T}, \gamma).$$

The question we ask is: suppose that instead of using all available history, we construct an approximate model using the finite window information variables

$$\begin{aligned} I_t^N &= \{y_{[t-N,t]}, u_{[t-N,t-1]}\}, \text{ if } t \geq N, \\ I_t^N &= \{y_{[0,t]}, u_{[0,t-1]}\}, \text{ if } 0 < t < N, \\ I_0 &= \{y_0\}, \end{aligned} \tag{8.24}$$

that is we observe the information variables through a window whose length is N. Suppose, we denote the optimal value function of the approximate model by J_β^N and the approximate policy by γ^N.

Inspired from filter stability, consider the following: For any time step $t \geq N$ and for a fixed observation realization sequence $y_{[0,t]}$ and control action sequence $u_{[0,t-1]}$, the state process can be viewed as

$$P^{\mu}(x_t \in \cdot | y_{[0,t]}, u_{[0,t-1]}) = P^{\pi_{t-N_-}}(X_t \in \cdot | y_{[t-N,t]}, u_{[t-N,t-1]})$$

where

$$\pi_{t-N_-}(\cdot) = P^{\mu}(x_{t-N} \in \cdot | y_{[0,t-N-1]}, u_{[0,t-N-1]}).$$

That is, we can view the state as the Bayesian update of π_{t-N_-}, the predictor at time $t-N$, using the observations $y_{t-N}, \ldots, y_t$. Notice that with this representation only the most recent N observation realizations are used for the update and the past information of the observations is embedded in π_{t-N_-}. Consider the following set (state space):

$$\mathcal{Z} = \left\{(\pi, y_{[0,N]}, u_{[0,N-1]}); \pi \in \mathcal{P}(\mathbb{X}), y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N\right\}$$

We place the product metric on this new space: weak convergence on the belief and usual metric on the measurements and actions.

The approach is summarized in Figure 8.1.

Now, for a fixed $\hat{\pi} \in \mathcal{P}(\mathbb{X})$, consider the quantized state space:

$$\mathcal{Z}^N = \left\{(\hat{\pi}, y_{[0,N]}, u_{[0,N-1]}); y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N\right\}$$

The idea is to quantize $\mathcal{Z}$ as follows: collapse all π to a fixed state $\hat{\pi}$, define an approximate finite MDP and establish performance bounds utilizing filter stability and the robustness approach presented earlier

In the following, we will assume that $\mathbb{X}$ is $\mathbb{R}^n$ for some n and that $\mathbb{U}, \mathbb{Y}$ are finite sets.

The actual state space and the finite approximation are:

$$\mathcal{Z} = \left\{(\pi, y_{[0,N]}, u_{[0,N-1]}); \pi \in \mathcal{P}(\mathbb{X}), y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N\right\}$$
$$\mathcal{Z}^N = \left\{(\hat{\pi}, y_{[0,N]}, u_{[0,N-1]}); y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N\right\}$$

Define the map and $F : \mathcal{Z} \to \mathcal{Z}^N$, such that for $(\pi, y_{[0,N]}, u_{[0,N-1]}) \in \mathcal{Z}$

$$F(\pi, y_{[0,N]}, u_{[0,N-1]}) = (\hat{\pi}, y_{[0,N]}, u_{[0,N-1]}).$$

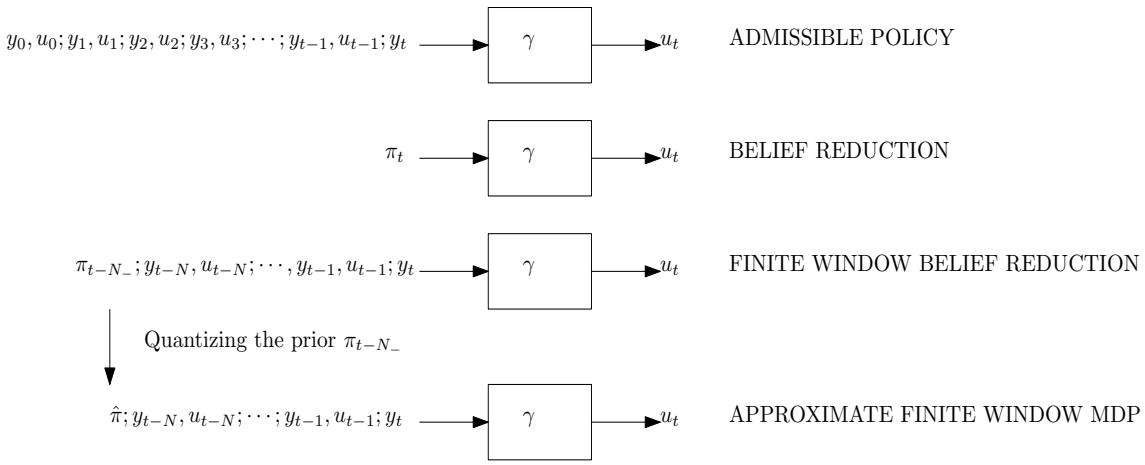

Fig. 8.1: Construction of the Finite-Window Approximate MDP from the Finite-Window Belief-MDP [189].

Using the map F and the finite set $\mathcal{Z}^N$, one can define a finite belief MDP, and construct a policy for this finite model, by extending it, we can use the policy, say $\tilde{\phi}^N$ for the original model.

The cost function for the approximate model is

$$\begin{aligned}\hat{c}(\hat{z}_t^N, u_t) = \hat{c}(\hat{\pi}, I_t^N, u_t) &:= \tilde{c}(\phi(\hat{\pi}, I_t^N), u_t) \\ &= \int_{\mathbb{X}} c(x_t, u_t) P^{\hat{\pi}}(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}).\end{aligned}$$

We define the controlled transition model for the approximate model by

$$\hat{\eta}^N(\hat{z}_{t+1}^N | \hat{z}_t^N, u_t) = \hat{\eta}^N(\hat{\pi}, I_{t+1}^N | \hat{\pi}, I_t^N, u_t) := \hat{\eta}\Big(\mathcal{P}(\mathbb{X}), I_{t+1}^N | \hat{\pi}, I_t^N, u_t\Big). \tag{8.25}$$

We will write $\mathcal{Z}_{\hat{\pi}}^N$ to make the dependence on $\hat{\pi}$ and N more explicit.

For simplicity, if we assume $N = 1$, then the transitions can be rewritten for some $I_{t+1}^N = (\hat{y}_{t+1}, \hat{y}_t, \hat{u}_t)$ and $I_t^N = (y_t, y_{t-1}, u_{t-1})$

$$\begin{aligned}\hat{\eta}^N(\hat{\pi}, \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \hat{\pi}, y_t, y_{t-1}, u_{t-1}, u_t) &= \hat{\eta}(\mathcal{P}(\mathbb{X}), \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \hat{\pi}, y_t, y_{t-1}, u_{t-1}, u_t) \\ &= 1_{\{y_t = \hat{y}_t, u_t = \hat{u}_t\}} P^{\hat{\pi}}(\hat{y}_{t+1} | y_t, y_{t-1}, u_t, u_{t-1}).\end{aligned} \tag{8.26}$$

Denoting the optimal value function for the approximate model by J_β^N, we can write the following fixed point equation

$$J_\beta^N(\hat{z}^N) = \min_{u \in \mathbb{U}} \left(\hat{c}(\hat{z}^N, u) + \beta \sum_{\hat{z}_1^N \in \hat{\mathcal{Z}}_{\hat{\pi}}^N} J_\beta^N(\hat{z}_1^N) \hat{\eta}^N(\hat{z}_1^N | \hat{z}^N, u) \right). \tag{8.27}$$

E.g., for $N = 1$, we rewrite the fixed point equation for some $\hat{z}_0^N = (\hat{\pi}, y_1, y_0, u_0)$ as

$$J_\beta^N(\hat{\pi}, y_1, y_0, u_0) = \min_{u_1 \in \mathbb{U}} \left(\hat{c}(\hat{\pi}, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in \mathbb{Y}} J_\beta^N(\hat{\pi}, y_2, y_1, u_1) P^{\hat{\pi}}(y_2 | y_1, y_0, u_1, u_0) \right). \tag{8.28}$$

We can now investigate the following approximation error terms:

$$|\tilde{J}_\beta^N(\hat{z}) - J_\beta^*(\hat{z})|, J_\beta(\hat{z}, \tilde{\phi}^N) - J_\beta^*(\hat{z}).$$

The first one is the difference between the optimal value function of the original model and that for the approximate model. The second term is the performance loss due to the policy calculated for the approximate model being applied to the true model.

Using the approaches presented in Section 8.2 and what is to be presented in *Chapter* 12 (building on [183, 187], we can show that the loss is related to the term:

$$\begin{aligned}&L_t := \\ &\sup_{\hat{\gamma} \in \hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}} \left[\| P^{\pi_t^-}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\hat{\pi}}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \|_{TV} \right]\end{aligned} \tag{8.29}$$

Notice that this term is directly related to filter stability studied in *Chapter* 6.

Theorem 8.3.1 *[188, 189] [Continuity of Value Functions] For $\hat{z}_0 = (\pi_0^-, I_0^N)$, if a policy $\hat{\gamma}$ acts on the first N step of the process which produces I_0^N, we then have*

$$E^{\hat{\gamma}}_{\pi_0^-}\left[\left|\tilde{J}^N_\beta(\hat{z}_0) - J^*_\beta(\hat{z}_0)\right| | I_0^N\right] \leq \frac{\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t$$

Theorem 8.3.2 *[188, 189] [Robustness of Approximate Finite Window Model Solution applied to Actual Model] For $\hat{z}_0 = (\pi_0^-, I_0^N)$, with a policy $\hat{\gamma}$ acting on the first N steps*

$$E^{\hat{\gamma}}_{\pi_0^-}\left[\left|J_\beta(\hat{z}_0, \tilde{\phi}^N) - J^*_\beta(\hat{z}_0)\right| | I_0^N\right] \leq \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t.$$

As one example, we now show that the term L_t can be bounded by the filter stability result presented in Theorem 6.4.1. Recall that this states that

$$E^{\mu,\gamma}\left[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}\right] \leq 2\alpha^n.$$

which holds uniformly for all $\mu \ll \nu$ where $\alpha := (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(Q))$.

Since $\tilde{\delta}(\mathcal{T})$ is a uniform Dobrushin coefficient over all control actions, the above bound is valid under any control policy. Thus we have that

$$\begin{aligned} L_t &= \sup_{\gamma \in \Gamma} E^{\hat{\gamma}}_{\pi_0^-}\left[\|P^{\pi_t^-}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\hat{\pi}}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]})\|_{TV}\right] \\ &\leq 2\alpha^N \end{aligned} \tag{8.30}$$

Theorem 8.3.3 *[188, 189] Assume the following holds:*

(i) The exponential filter stability condition applies:

$$\alpha := (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(Q)) < 1 \tag{8.31}$$

(ii) The transition kernel $\mathcal{T}$ is dominated, i.e. there exists a dominating measure $\hat{\pi} \in \mathcal{P}(\mathbb{X})$ such that for every $x \in \mathbb{X}$ and $u \in \mathbb{U}$, $\mathcal{T}(\cdot|x,u) \ll \hat{\pi}(\cdot)$.

Then, by choosing the dominating measure $\hat{\pi}$ for the approximate model,

$$E^{\hat{\gamma}}_{\pi_0^-}\left[\left|J_\beta(\hat{z}_0, \tilde{\phi}^N) - J^*_\beta(\hat{z}_0)\right| | I_0^N\right] \leq \frac{4\|c\|_\infty}{(1-\beta)^2}\alpha^N. \tag{8.32}$$

Instead of the Dobrushin based analysis, the more relaxed stability condition presented in Example 6.9 can also be adopted, though not leading to an exponential convergence rate in the memory size. In that case, $L_t \to 0$ asymptotically and thus (8.32) converges to 0 as N increases only asymptotically, without a geometric convergence rate in N.

Via a somewhat different, and more direct, derivation, [188, Section 4.2 and Theorem 17] presented the following alternative condition involving sample path-wise uniform filter stability term

$$\bar{L}^N_{TV} := \sup_{z \in \mathcal{P}(X)} \sup_{y_{[0,N]}, u_{[0,N-1]}} \left\| P^z(\cdot | y_{[0,N]}, u_{[0,N-1]}) - P^{z^*}(\cdot | y_{[0,N]}, u_{[0,N-1]}) \right\|_{TV}, \tag{8.33}$$

to show the following *uniform* error bound:

$$\sup_z \left|J_\beta(z, \gamma_N) - J^*_\beta(z)\right| \leq \frac{2(1 + (\alpha_{\mathcal{Z}} - 1)\beta)}{(1-\beta)^3(1 - \alpha_{\mathcal{Z}}\beta)} \|c\|_\infty \bar{L}^N_{TV} \tag{8.34}$$

for all $\beta \in (0,1)$ under a contraction condition, for some constant $\alpha_{\mathcal{Z}}$ defined in [188]. Additionally, [188, Theorem 9] provided conditions where the error is in the bounded-Lipschitz metric (which is equivalent to the Wasserstein-1 metric when the state space $\mathbb{X}$ is compact), however these were only applicable for a restrictive subset of the discount

parameter β. On the other hand, the bound in (8.32) is in expectation whereas the bound in (8.34) is uniform, and thus the results are complementary.

As a complementary condition, via the Birkhoff-Hopf theorem, a controlled version of a contraction via the Hilbert metric [146] can be utilized [101]: Recall that

$$F(z,y,u)(\cdot) = \Pr\{X_{k+1} \in \cdot \mid Z_k = z, Y_{k+1} = y, U_k = u\}$$

Assumption 8.3.1 *1. $Q(y|x) \geq \epsilon > 0$ for every $x \in \mathbb{X}$ and $y \in \mathbb{Y}$.*

2. The transition kernel $\mathcal{T}(.|.,u)$ is a mixing kernel (see Definition 6.4.8) for every $u \in \mathbb{U}$.

Lemma 8.3.1 *[1] Under Assumption 8.3.1, there exists a constant $r < 1$ such that*

$$h(F(\mu,y,u), F(\nu,y,u)) \leq r h(\mu,\nu) \tag{8.35}$$

for every comparable $\mu, \nu \in \mathcal{P}(\mathbb{X})$ and for every $u \in \mathbb{U}$ and $y \in \mathbb{Y}$. Here $r = \frac{1-\epsilon_u^2\epsilon}{1+\epsilon_u^2\epsilon}$, ϵ_u is the mixing constant of the kernel $\mathcal{T}(.|.,u)$.

Theorem 8.3.4 *[101] Under Assumption 8.3.1, there exists a constant $r < 1$ and K such that*

$$\bar{L}_{TV}^N \leq r^{N-1} K. \tag{8.36}$$

Here, $K = \frac{2}{\log 3} \sup h(Z_1, Z_1^*)$ and $r = \sup_{u\in\mathbb{U}} \frac{1-\epsilon_u^2\epsilon}{1+\epsilon_u^2\epsilon}$.

Corollary 8.6. *[101] Under Assumption 8.3.1, there exists a constant $r < 1$ and K such that*

$$E\left[J_\beta(\pi_N^-, \mathcal{T}, \gamma^N) - J_\beta^*(\pi_N^-, \mathcal{T}) | I_0^N\right] \leq \frac{2\|c\|_\infty}{(1-\beta)^2} r^{N-1} K. \tag{8.37}$$

Here, $K = \frac{2}{\log 3} \sup h(Z_1, Z_1^*)$ and $r = \sup_{u\in\mathbb{U}} \frac{1-\epsilon_u^2\epsilon}{1+\epsilon_u^2\epsilon}$.

Despite these consequential approximation results, implementing the above is still tedious; though possible. Can reinforcement learning be feasible? Can we view the finite history as an effective *state*? We will address this question in the following chapter.

8.4 Bibliographic Notes

For computational and learning methods, there is an extensive literature, where various approaches have been developed. A partial list of these techniques is as follows: approximate dynamic programming, approximate value or policy iteration, simulation-based techniques, neuro-dynamic programming (or reinforcement learning), state aggregation, etc. [39, 41, 88, 112].

We refer the reader to the monograph [270] for a general treatment of approximation results along what has been presented for continuous spaces; these also build on on [185, 269, 272, 273, 275]. presentation, only focus on the setup where the state spaces considered are compact. A generalization of some of the approximation results are presented in [187] in view of robustness properties.

8.5 Exercises

Exercise 8.5.1 *Consider a controlled Markov chain with state space $\mathbb{X} = \{0,1\}$, action space $\mathbb{U} = \{0,1\}$, and transition kernel for $t \in \mathbb{Z}_+$:*

$$P(x_{t+1} = 1|x_t = 0, u_t = 1) = P(x_{t+1} = 1|x_t = 1, u_t = 1) = \alpha$$

$$P(x_{t+1} = 1|x_t = 0, u_t = 0) = P(x_{t+1} = 1|x_t = 1, u_t = 0) = 1 - \alpha.$$

where $\alpha \in (0,1)$. Let a cost function $c(x,u)$, with $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ be given by

$$c(0,1) = c(0,0) = 1 \qquad\qquad c(1,0) = c(1,1) = 2.$$

Suppose that the goal is to minimize the quantity

$$E_0^\gamma[\sum_{t=0}^{\infty} \beta^t c(x_t, u_t)],$$

for a fixed $\beta \in (0,1)$, over all admissible policies $\gamma \in \Gamma_A$.

Find an optimal policy and the optimal expected cost explicitly, as a function of α, β (note that the initial condition is $x_0 = 0$).

Exercise 8.5.2 *Consider the following problem: Let $\mathbb{X} = \{1,2\}, \mathbb{U} = \{1,2\}$, where $\mathbb{X}$ denotes whether a fading channel is in a good state ($x = 2$) or a bad state ($x = 1$). There exists an encoder who can either try to use the channel ($u = 2$) or not use the channel ($u = 1$). The goal of the encoder is send information across the channel.*

Suppose that the encoder's cost (to be minimized) is given by:

$$c(x,u) = -1_{\{x=2,u=2\}} + \alpha(u-1),$$

for $\alpha = 1/2$ (if you view this as a maximization problem, you can see that the goal is to maximize information transmission efficiency subject to a cost involving an attempt to use the channel; the model can be made more complicated but the idea is that when the channel state is good, $u = 2$ can represent a channel input which contains data to be transmitted and $u = 1$ denotes that the channel is not used).

Suppose that the transition kernel is given by:

$$P(x_{t+1} = 2|x_t = 2, u_t = 2) = 0.8, \qquad P(x_{t+1} = 1|x_t = 2, u_t = 2) = 0.2$$

$$P(x_{t+1} = 2|x_t = 2, u_t = 1) = 0.2, \qquad P(x_{t+1} = 1|x_t = 2, u_t = 1) = 0.8$$

$$P(x_{t+1} = 2|x_t = 1, u_t = 2) = 0.5, \qquad P(x_{t+1} = 1|x_t = 1, u_t = 2) = 0.5$$

$$P(x_{t+1} = 2|x_t = 1, u_t = 1) = 0.9, \qquad P(x_{t+1} = 1|x_t = 1, u_t = 1) = 0.1$$

We will consider either a discounted cost criterion for some $\beta \in (0,1)$ (you can fix an arbitrary value)

$$\inf_\gamma E_x^\gamma[\sum_{t=0}^{\infty} \beta^t c(x_t, u_t)] \tag{8.38}$$

or the average cost criterion

$$\inf_\gamma \limsup_{T\to\infty} \frac{1}{T} E_x^\gamma[\sum_{t=0}^{T-1} c(x_t, u_t)]. \tag{8.39}$$

a) Using Matlab or some other program, obtain a solution to the problem given above in (9.34) through the following:

(i) Policy Iteration

(ii) Value Iteration.

b) Consider the criterion given in (9.35). Apply the convex analytic method, by solving the corresponding linear program, to find the optimal policy. In Matlab, the command linprog can be used to solve linear programming problems. See (7.42).

Exercise 8.5.3 (Convex Analytic Method for Discounted Cost) *Consider the convex analytic method for a discounted cost problem, where the expected occupation measures are defined, for a given $\gamma \in \Gamma_A$ as*

$$\eta(A \times B) = \sum_{k=0}^{\infty} \beta^k P^\gamma(X_k \in A, U_k \in B)$$

Show that the set of all such expected occupation measures η are equivalent to the set of all expected occupation measures achieved by stationary (and randomized) γ, that is, by $\gamma \in \Gamma_{SR}$.

Hint: Obtain a recursive equation involving

$$\begin{aligned}
&\eta(A \times \mathbb{U}) = P(X_0 \in A) + \beta \sum_{k=1}^{\infty} \beta^{t-1} P^\gamma(X_k \in A) \\
&= P(X_0 \in A) + \beta \sum_{k=1}^{\infty} \beta^{k-1} \int \mathcal{T}(A|x,u) P^\gamma(X_{k-1} \in dx, U_{k-1} \in du) \\
&= P(X_0 \in A) + \beta \int \int \mathcal{T}(A|x,u) \sum_{k=1}^{\infty} \beta^{k-1} P^\gamma(X_{k-1} \in dx, U_{k-1} \in du) \\
&= P(X_0 \in A) + \beta \int \int \mathcal{T}(A|x,u) \eta(dx, du)
\end{aligned}$$

Then, note the similarity with (8.7), in that this equation is also satisfied by selecting a stationary control κ defined almost everywhere with $\kappa(du|x) = \frac{d\eta(dx,du)}{d\eta(dx)}(x)$.

Exercise 8.5.4 *Let $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ be bounded, where $\mathbb{X}$ is the state space and $\mathbb{U}$ is the action space for a controlled stochastic system. Suppose that under a stationary policy γ, the expected discounted cost, for $\beta < 1$, is given by*

$$J_\beta(x, \gamma) := E_x^\gamma[\sum_{k=0}^{\infty} \beta^k c(x_k, \gamma(x_k))] = c(x, \gamma(x)) + \beta \int J_\beta(x_{t+1}, \gamma) \mathcal{T}(dx_{t+1}|x_t = x, u_t = \gamma(x))$$

Let f_1 and f_2 be two stationary policies. Define a third policy, g, as:

$$g(x) = f_1(x) 1_{\{x \in C\}} + f_2(x) 1_{\{x \in \mathbb{X} \setminus C\}}$$

where

$$C = \{x : J_\beta(x, f_1) \leq J_\beta(x, f_2)\}$$

and $\mathbb{X} \setminus C$ denotes the complement of this set.

Show that $J_\beta(x, g) \leq J_\beta(x, f_1)$ and $J_\beta(x, g) \leq J_\beta(x, f_2)$ for all $x \in \mathbb{X}$.

Exercise 8.5.5 *Consider a controlled Markov chain with state space $\mathbb{X} = \{0, 1\}$, action space $\mathbb{U} = \{0, 1\}$, and transition kernel for $t \in \mathbb{Z}_+$:*

$$P(x_{t+1} = 1 | x_t = 0, u_t = 1) = 1$$

$$P(x_{t+1} = 1 | x_t = 0, u_t = 0) = \frac{1}{2}$$

$$P(x_{t+1} = 1|x_t = 1, u_t = 0) = P(x_{t+1} = 1|x_t = 1, u_t = 1) = \frac{1}{2}.$$

Let a cost function $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ *be given by*

$$c(0,1) = \frac{1}{2}, \qquad c(0,0) = 1$$

$$c(1,0) = \frac{5}{4}, \qquad c(1,1) = 2$$

Suppose that the goal is to minimize the quantity

$$E_0^{\gamma}[\sum_{t=0}^{\infty}(\frac{1}{2})^t c(x_t, u_t)],$$

over all admissible policies $\gamma \in \Gamma_A$.

Find an **optimal policy** *using* **Policy Iteration**.

Note. *Please note that you can only use a pen and paper for this problem. Note that for an invertible 2x2 matrix*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

we have

$$A^{-1} = \frac{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}{ad - bc}$$

9

Reinforcement Learning

In this chapter, we will study stochastic learning and reinforcement learning methods, first in the context of finite space MDPs and later on for general continuous (standard Borel) space MDPs and POMDPs.

9.1 Stochastic Learning Algorithms and the Q-Learning Algorithm

In some Markov Decision Problems (MDPs), one does not know the true transition kernel or the cost function, and may wish to use past data to obtain an asymptotically optimal solution (that is, via *learning* from past data). In some problems, this may be used as an efficient numerical method to obtain approximately optimal solutions. There may also be setups where a prior probabilistic knowledge on the system dynamics may be used to learn the true system. In particular, one may apply Bayesian (probabilistically driven given some prior information) or non-Bayesian (primarily empirical, without assuming a prior probabilistic model) methods.

An important class of non-Bayesian methods are known as stochastic approximation algorithms: such approximation methods are used extensively in many application areas. A typical stochastic approximation algorithm has the following form

$$x_{t+1} = x_t + \alpha_t(F(x_t) - x_t + w_t) \tag{9.1}$$

where w_t is a zero-mean noise variable, x_t is a stochastic process and w_t is some driving noise. The goal is to arrive at a point x^* which satisfies $x^* = F(x^*)$, where F may correspond to an optimality condition.

Exercise 9.1.1 (Stochastic gradient descent) *We revisit the stochastic gradient descent algorithm discussed in Theorem 4.3.4 noting the similarity with (9.1). Consider a convex function $f : \mathbb{R}^n \to \mathbb{R}$ and denote the set of minima of f by X^*. Let a sequence of iterates be given by*

$$x_{k+1} = x_k - \gamma_k\Big(\nabla_x f(x_k) + n_k\Big), \quad x_0 \in \mathbb{R}^n, \tag{9.2}$$

where the random variables n_k are zero-mean, orthogonal to one another, and have their second moments uniformly bounded. If $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$, the sequence of iterates (9.2) converges almost surely to some element $x^* \in X^*$

9.1.1 Q-Learning

$Q-$learning [26, 41, 300, 303, 306, 323] is a stochastic approximation algorithm used for fully observed finite space MDPs that does not require the knowledge of the transition kernel, or even the cost (or reward) function for its implementation. In this algorithm, the incurred per-stage cost variable is observed through simulation of a single sample path.

In this context, one may reflect on the way humans respond to experience, for example a baby learning to experiment with gravity without knowing physics and therefore physical models at all!

When the state and action spaces are finite, under mild conditions regarding infinitely often hits for all state-action pairs, this algorithm is known to converge to the optimal cost. We now discuss this algorithm.

Consider a Markov Decision Problem with finite state and action sets with the criterion given in (8.1) for some $\beta \in (0,1)$.

Let $Q : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ denote the *Q-factor* of the controller. Let us assume that the decision maker applies an arbitrary admissible policy γ and updates its Q-factors as follows for $t \geq 0$,

$$Q_{t+1}(x,u) = Q_t(x,u) + \alpha_t(x,u)\left(c(x,u) + \beta \min_v Q_t(X_{t+1},v) - Q_t(x,u)\right) \tag{9.3}$$

where the initial condition Q_0 is given, $\alpha_t(x,u)$ is the step-size for (x,u) at time t, u_t is chosen arbitrarily -e.g., according to some random exploration policy γ- as long as some technical conditions noted below hold, and the random state $X_{t+1} \sim P(X_{t+1} \in \cdot\,|X_t = x, U_t = u)$. It is assumed that, for all (x,u), $t \geq 0$, the following hold

Assumption 9.1.1 *For all* (x,u), $t \geq 0$,

(i) $\alpha_t(x,u) \in [0,1]$

(ii) $\alpha_t(x,u) = 0$ *unless* $(x,u) = (x_t,u_t)$

(iii)$\alpha_t(x,u)$ *is a (deterministic) function of* $(x_0,u_0),\ldots,(x_t,u_t)$.

(iv)$\sum_{t\geq 0} \alpha_t(x,u) = \infty$, *almost surely*

(v) $\sum_{t\geq 0} \alpha_t^2(x,u) \leq C$, *almost surely, for some (deterministic) constant* $C < \infty$.

A common way to select α coefficients in the algorithm is to take for every (x,u) pair:

$$\alpha_t(x,u) = \frac{1}{1 + \sum_{k=0}^{t} 1_{\{X_k = x, U_k = u\}}}$$

The selection of the control actions for each state can be arbitrary, as long as the assumptions above are guaranteed to hold.

Let F be an operator acting on the Q factors defined by:

$$F(Q)(x,u) = c(x,u) + \beta \sum_{x'} \mathcal{T}(x'|x,u) \min_v Q(x',v), \tag{9.4}$$

where, as before, $\mathcal{T}(x'|x,u) = P(x_1 = x'|x_0 = x, u_0 = u)$ is the transition kernel. Consider the following fixed point equation.

$$Q^*(x,u) = F(Q^*)(x,u) = c(x,u) + \beta \sum_{x'} \mathcal{T}(x'|x,u) \min_v Q^*(x',v) \tag{9.5}$$

whose existence and uniqueness follow essentially identically from arguments used in the contraction analysis utilized in *Chapter 5* (see Theorem 5.5.2), by using the norm $\|Q\|_\infty = \max_{(x,u)} |Q(x,u)|$: Since for every realization x_{t+1} of X_{t+1}, $|\min_v Q_t(x_{t+1},v) - \min_{v'} Q_t^*(x_{t+1},v')| \leq \max_v |Q_t(x_{t+1},v) - Q_t^*(x_{t+1},v)|$, we have that

$$|F(Q_t)(x,u) - F(Q^*)(x,u)| \leq \beta \|Q_t - Q^*\|_\infty := \beta \max_{x,u} |Q_t(x,u) - Q^*(x,u)| \tag{9.6}$$

Now, note that we can write (9.3) as

$$Q_{t+1}(x,u) = Q_t(x,u) + \alpha_t(x,u)\Bigg(F(Q_t)(x,u) - Q_t(x,u) + \Bigg(c(x,u) + \beta \min_v Q_t(x_{t+1},v) - F(Q_t)(x,u)\Bigg)\Bigg) \tag{9.7}$$

which is in the same form as (9.1) since

$$w_t := \Bigg(c(x_t,u_t) + \beta \min_v Q_t(x_{t+1},v) - F(Q_t)(x_t,u_t)\Bigg), \tag{9.8}$$

conditioned on the filtration generated by $\{x_t, u_t\}$ up to time t, is a zero-mean random variable.

Let us write (9.7) as

$$Q_{t+1}(x,u) = (1-\alpha_t(x,u))Q_t(x,u) + \alpha_t(x,u)\left(F(Q_t)(x,u) + \left(c(x,u) + \beta \min_v Q_t(x_{t+1},v) - F(Q_t)(x,u)\right)\right)$$

and by (9.5)

$$\begin{aligned} &Q_{t+1}(x,u) - Q^*(x,u) \\ &= (1-\alpha_t(x,u))(Q_t(x,u) - Q^*(x,u)) + \alpha_t(x,u)\Bigg(F(Q_t)(x,u) - F(Q^*)(x,u) \\ &\qquad + \Bigg(c(x,u) + \beta \min_v Q_t(x_{t+1},v) - F(Q_t)(x,u)\Bigg)\Bigg) \end{aligned} \tag{9.9}$$

Theorem 9.1.1 *(i) Under Assumption 9.1.1, the algorithm (9.3) converges almost surely to Q^*.*

(ii) A stationary policy f^* which satisfies $\min_u Q^*(x,u) = Q^*(x, f^*(x))$ is an optimal policy.

Proof. (i) From (9.9), the process Q_t satisfies the following form, with $S_t = Q_t - Q_t^*$:

$$S_{t+1}(x,u) = (1-\alpha_t(x,u))S_t(x,u) + \alpha_t(x,u)\Bigg((F(Q_t)(x,u) - F(Q^*)(x,u)) + w_t\Bigg),$$

where $\{\alpha_t\}$ satisfies Assumption 9.1.1 and w_t is given in (9.8).

We will consider the following two parallel dynamics, as in [178, Theorem 1]:

$$S^a_{t+1}(x,u) = (1-\alpha_t(x,u))S^a_t(x,u) + \alpha_t(x,u)w_t, \tag{9.10}$$

$$S^b_{t+1}(x,u) = (1-\alpha_t(x,u))S^b_t(x,u) + \alpha_t(x,u)\Bigg(F(Q_t)(x,u) - F(Q^*)(x,u)\Bigg). \tag{9.11}$$

We have $S_t(x,u) = S^a_{t+1}(x,u) + S^b_{t+1}(x,u)$. We will study each of these two additive terms separately.

The first step is to show that $S^a_{t+1}(x,u) \to 0$ almost surely. We will show this further below.

Assume then for now that $S^a_t(x,u) \to 0$ almost surely. We then focus on $S^b_{t+1}(x,u) + S^a_{t+1}(x,u)$, using the fact that, by (9.6)

$$\|(F(Q_t)(\cdot,\cdot) - F(Q^*)(\cdot,\cdot))\|_\infty \le \beta\|S_t\|_\infty \le \beta\|S^a_{t+1}\|_\infty + \beta\|S^b_{t+1}\|_\infty$$

Almost surely for sample paths ω, for every $\epsilon > 0$, there exists $N(\omega)$ such that for $t \ge N(\omega)$, $\|S^a_{t+1}(\omega)\|_\infty \le \epsilon$ (where, in the following, we suppress the sample path dependence and thus omit ω). In the following, we assume that $t \ge N$. Now, for some M large enough, let $\hat{\beta} := \beta(1 + \frac{1}{M}) < 1$ and for $\|S^b_t\|_\infty > M\epsilon$ note that

$$\beta\|S^b_t(x,u) + \epsilon\| \le \hat{\beta}\|S^b_t\|_\infty$$

and

$$
\begin{aligned}
|S^b_{t+1}(x,u)| &\leq (1-\alpha_t(x,u))|S^b_t(x,u)| + \left|\alpha_t(x,u)\left(F(Q_t)(x,u) - F(Q^*)(x,u)\right)\right| \\
&\leq (1-\alpha_t(x,u))|S^b_t(x,u)| + \alpha_t(x,u)(\beta\|S^a_{t+1}\| + \beta\|S^b_{t+1}\|) \qquad (9.12)\\
&\leq (1-\alpha_t(x,u))|S^b_t(x,u)| + \alpha_t(x,u)\hat{\beta}\|S^b_t\|_\infty \qquad (9.13)\\
&< \|S^b_t\|_\infty \qquad (9.14)
\end{aligned}
$$

Hence $\max_{x,u}(S^b_t(x,u))$ monotonically decreases for $\|S^b_t\|_\infty > M\epsilon$ leading to two possibilities: it either gets below $M\epsilon$ or it never gets below $M\epsilon$ in which case by the monotone non-decreasing property it will converge to some number, say M_1 with $M_1 \geq M\epsilon$.

Now, if the former is the case: once $\|S^b_t\|_\infty \leq M\epsilon$ we can show via (9.12) and $\beta(M+1)/M < 1$ that it will remain there thereafter.

We now show that the latter, that is with the limit being $M_1 \geq M\epsilon$, is not possible. The relation

$$|S^b_{t+1}(x,u)| \leq (1-\alpha_t(x,u))|S^b_t(x,u)| + \alpha_t(x,u)\hat{\beta}\|S^b_t\|_\infty$$

implies that (via an argument similar to what is known as Grönwall's lemma, as can be inductively shown) the solution is bounded from above by the solution to the equation

$$|S^b_{t+1}(x,u)| = (1-\alpha_t(x,u))|S^b_t(x,u)| + \alpha_t(x,u)\hat{\beta}\|S^b_t\|_\infty$$

which can be shown to converge to zero. This follows from the reasoning that, for any fixed D, the iterate

$$|S^b_{t+1}(x,u)| = (1-\alpha_t(x,u))|S^b_t(x,u)| + \alpha_t(x,u)\hat{\beta}D$$

converges to $\hat{\beta}D$; this follows since the effects of the initial condition diminish by the summability of α_t (see Exercise 3.5.7) and hence there exists only one limit solution, which by inspection will be equal to $\hat{\beta}D$ (the uniqueness of the solution can be shown by subtracting $\hat{\beta}D$ from the iterates, whose limit would be zero for any initialization). Therefore, if there is an upper bound D_0 on the iterates, the bounds for future iterations eventually get smaller and smaller than $\hat{\beta}D_0 + \delta =: D_1$ for any arbitrarily small $\delta > 0$, and as time progresses by an inductive reasoning, eventually the iterates would have to converge to zero (see also [178, Proof of Lemma 3] or [306, p. 196]).

Thus, for any $\epsilon > 0$, for large enough t, we have that $\|S^b_t\|_\infty \leq M\epsilon$. Since $\epsilon > 0$ is arbitrary, the convergence result follows.

We now discuss S^a_t[1].

[1]We note that an alternative, more direct, argument, due [189, Theorem 4.1], is also possible under additional structure on the random exploration policy used to generate the actions and if the learning rate leads to averaging dynamics as in Exercise 9.1.2; see [189, Theorem 4.1] for the analysis to follow and [191, Theorem 2.1]: In particular, if the policy adopted to randomly generate actions leads to a positive Harris recurrent Markov chain, then the following more direct argument is possible to establish convergence without apriori showing boundedness of the iterates: Write

$$Q_{t+1}(x,u) = (1-\alpha_t(x,u))Q_t(x,u) + \alpha_t(x,u)\left(F(Q_t)(x,u) + \left(c(x,u) + \beta\min_v Q_t(x_{t+1},v) - F(Q_t)(x,u)\right)\right)$$

and

$$
\begin{aligned}
Q_{t+1}(x,u) - Q^*(x,u) &= (1-\alpha_t(x,u))(Q_t(x,u) - Q^*(x,u)) \\
&+\alpha_t(x,u)\left(F(Q_t)(x,u) - F(Q^*)(x,u) + \beta\min_v Q_t(x_{t+1},v) - \beta E[\min_v Q_t(x_{t+1},v)|x_t = x, u_t = u]\right)
\end{aligned} \qquad (9.15)
$$

With

Taking the square of S_t^a, we obtain:

$$E[(S_{t+1}^a(x,u))^2|\mathcal{F}_t] \leq (S_t^a(x,u))^2 - 2\alpha_t(S_t^a(x,u))^2 + \alpha_t^2(S_t^a(x,u))^2 + \alpha_t^2(x,u)w_t^2 \tag{9.19}$$

First, by an argument identical to that used in the proof of the Comparison Theorem (Theorem 4.2.3), we have that for any $T > 0$:

$$\begin{aligned} E[\sum_{t=0}^{T-1}(2\alpha_t - \alpha_t^2)(S_t^a(x,u))^2 &\leq (S_0^a(x,u))^2 + E[\sum_{t=0}^{T-1}\alpha_t^2(x,u)w_t^2] \\ &\leq \quad (S_0^a(x,u))^2 + C(\sup_t w_t^2), \end{aligned} \tag{9.20}$$

where we use Assumption 9.1.1(v). We now show that $(\sup_t w_t^2)$ is uniformly bounded: By (9.7), we have that

$$Q_{t+1}(x,u) = (1-\alpha_t(x,u))Q_t(x,u) + \alpha_t(x,u)(c(x,u) + \beta \min_v Q_t(x_{t+1},v))$$

which implies that

$$|Q_{t+1}(x,u)| \leq (1-\alpha_t(x,u))\|Q_t\|_\infty + \alpha_t(x,u)(c(x,u) + \beta\|Q_t\|_\infty)$$

Now, if $\|Q_t\|_\infty > \frac{\|c\|_\infty}{1-\beta} =: L_1$, we have that

$$|Q_{t+1}(x,u)| \leq (1-\alpha_t(x,u))\|Q_t\|_\infty + \alpha_t(x,u)(c(x,u)+\beta\|Q_t\|_\infty) = \|Q_t\|_\infty + \alpha_t(x,u)(c(x,u)+(\beta-1)\|Q_t\|_\infty) < \|Q_t\|_\infty,$$

And thus, since this holds for all (x,u) pairs, $\|Q_{t+1}\|_\infty < \|Q_t\|_\infty$ whenever $\|Q_t\|_\infty > L_1$ and hence $|Q_t\|_\infty$ would be a decreasing sequence as long as it is above L_1. On the other hand, if $\|Q_t\|_\infty \leq L_1$, then $\|Q_{t+1}\|_\infty \leq L_1$ as well. These imply that $\|Q_t\|_\infty$ is uniformly bounded almost surely. As a consequence w_t is also bounded, uniformly over time. Thus, the right hand side of (9.20) is bounded.

Furthermore, by re-writing (9.19), in the expression

$$E[(S_{t+1}^a(x,u))^2|\mathcal{F}_t] \leq (S_t^a(x,u))^2 - (2\alpha_t - \alpha_t^2)(S_t^a(x,u))^2 + \alpha_t^2(x,u)w_t^2,$$

$$r_t(x,u) := \beta \min_v Q_t(x_{t+1},v) - \beta E[\min_v Q_t(x_{t+1},v)|x_t = x, u_t = u].$$
$$r_t^*(x,u) := \beta \min_v Q^*(x_{t+1},v) - \beta E[\min_v Q^*(x_{t+1},v)|x_t = x, u_t = u]$$

$$\begin{aligned} Q_{t+1}(x,u) - Q^*(x,u) &= (1-\alpha_t(x,u))(Q_t(x,u) - Q^*(x,u)) \\ &+\alpha_t(x,u)\Big(F(Q_t)(x,u) - F(Q^*)(x,u) + r_t^*(x,u) + (r_t(x,u) - r_t^*(x,u))\Big) \end{aligned}$$

Obtain the sum:

$$R_{t+1}^a(x,u) = (1-\alpha_t(x,u))S_t^a(x,u) + \alpha_t(x,u)r^*(x,u)_t, \tag{9.16}$$

$$R_{t+1}^b(x,u) = (1-\alpha_t(x,u))S_t^b(x,u) + \alpha_t(x,u)\Big((r_t(x,u) - r_t^*(x,u)\Big), \tag{9.17}$$

$$R_{t+1}^c(x,u) = (1-\alpha_t(x,u))S_t^b(x,u) + \alpha_t(x,u)\Big(F(Q_t)(x,u) - F(Q^*)(x,u)\Big), \tag{9.18}$$

$S_t(x,u) = R_{t+1}^a(x,u) + R_{t+1}^b(x,u) + R_{t+1}^c(x,u)$. By ergodicity, $R_t^a \to 0$ almost surely by averaging and the positive Harris recurrence under the exploration policy (by the additional assumption noted). Note that

$$|(r_t(x,u) - r_t^*(x,u) + F(Q_t)(x,u) - F(Q^*)(x,u)| = |\beta \min_v Q_t(x_{t+1},v) - \beta \min_v Q^*(x_{t+1},v)| \leq \beta\|Q_t - Q_t^*\|_\infty$$

and thus,

$$R_{t+1}^b(x,u) + R_{t+1}^c(x,u) \leq \beta\|Q_t - Q_t^*\|_\infty$$

As a result, by replacing S_{t+1}^b with $R_{t+1}^b + R_{t+1}^c$ we can trace the proof steps presented above without the analysis on S_t^a to follow.

the term $\alpha_t^2(y,u)w_t^2$ is finite almost surely. This implies, by Theorem 4.3.1, that S_t^a converges to some random variable almost surely. The inequality (9.20) then implies that this limit must be zero: Suppose not; since α_t is not summable, there exists an infinite sequence of times so that each summation of α_t between the times is bounded from below by a positive constant. Through this, if $(S_t^a)^2$ were not to converge to zero (given that it does converge to something else), it would remain above a positive constant after a sufficiently large time, and then it would follow that $\sum_t(2\alpha_t - \alpha_t^2)S_t^{a\,2}$ would not remain bounded. Therefore, if this were to happen with non-zero measure, the expectation of this term would be unbounded, which in turn would, as $T \to \infty$, violate (9.20). You can also, alternatively, build on Theorem 4.3.2.

(ii) Now, consider

$$Q^*(x,u) = F(Q^*)(x,u) = c(x,u) + \beta \sum_{x'} P(x'|x,u) \min_v Q^*(x',v)$$

Note that the minimum of u, for each x, is essentially the solution to the Discounted Cost Optimality Equation studied in *Chapter 5* (see Theorem 5.5.2). Hence, the stationary policy $\{f^*\}$ is optimal. ◇

We also refer the reader to the proof of [178, Theorem 1] for an alternative proof.

Exercise 9.1.2 *To gain some further intuition, and also a more direct proof for the case where the α_k term is taken as $\frac{1}{k}$ (or more precisely; for every (x,u) pair: $\alpha_t(x,u) = \frac{1}{1+\sum_{k=0}^{t} 1_{\{x_k=x,u_k=u\}}}$), consider the following averaging dynamics: Let a_t be a sequence of scalars and define:*

$$s_T = \frac{1}{T}\sum_{k=0}^{T-1} a_k$$

Observe that for $T > 1$, $Ts_T = (T-1)s_{T-1} + a_{T-1}$ which leads to

$$s_T = s_{T-1} + \frac{1}{T}(a_{T-1} - s_{T-1})$$

In view of this observation, conclude that with α_k in Assumption 9.1.1 taken as $\frac{1}{k}$, we have an averaging dynamics. One may interpret the Q-learning algorithm and its convergence properties with this insight. Furthermore, one can see that the convergence holds under more relaxed conditions, this will be utilized in Sections 9.2.2 and 9.3.

Remark 9.1. Via studying (9.10) and (9.13) separately, one can also arrive at convergence rates as a function of the number of iterates, see [5, 121, 299].

9.1.2 Reinforcement Learning for the Average Cost Criterion

For the average cost criterion, we can follow two approaches: (i) One is to utilize near optimality of discounted criterion policies as shown in Theorem 7.3.5 and Theorem 7.3.6; and (ii) another approach is via directly applying an algorithm tailored for the average cost criterion. We note here that the average cost setup is typically more challenging due to the lack of contraction properties. Nonetheless, as we have seen earlier in *Chapter* 7, one can follow contraction updates for the average cost criterion as well. See [190] for algorithms and an extensive review.

We also note that the references [2, 150] are among the earliest studies that provide convergent learning algorithms based on relative value iteration, and the convergence of these algorithms in these studies have been established via the ODE method [68] for finite models.

9.1.3 Synchronous Q-Learning

The algorithm above is known as the *asynchronous* Q-learning algorithm: At any given time only a single (x,u) pair can be updated. In some applications, where extensive data or simulations are available, one can simultaneously update multiple or all state-action pairs, leading to a synchronous version. The analysis above is applicable to the synchronous

setup, but the synchronous setup often allows for a more direct stochastic analysis especially for the average cost criterion.

9.2 Reinforcement Learning Methods for POMDPs

For the analysis in this section, please first recall the discussion in Section 8.3.1.

9.2.1 Near optimality of quantized policies under weak Feller property of non-linear filters

As we have seen, any POMDP can be reduced to a completely observable Markov process, whose states are the posterior state distributions or 'beliefs' of the observer; that is, the state at time t is

$$\pi_t(\,\cdot\,) := P\{X_t \in \,\cdot\,|y_0, \ldots, y_t, u_0, \ldots, u_{t-1}\} \in \mathcal{P}(\mathbb{X}).$$

As discussed earlier, this conditional probability measure process is the filter process. The filter process has state space $\mathcal{P}(\mathbb{X})$ and action space $\mathbb{U}$. Here, $\mathcal{P}(\mathbb{X})$ is equipped with the Borel σ-algebra generated by the topology of weak convergence. The transition probability of the filter process is given in (6.27).

Accordingly, we have a fully observed belief-MDP. Now, by combining the approximation results in Section 8.2 and reinforcement learning theoretic results to be presented in Section 9.3, together with the the weak Feller continuity results presented in Section 6.3.2, we can conclude that the numerical methods can also be applied to POMDPs under the conditions reported in Theorems 6.3.3 and 6.3.4 [128] [184].

Accordingly, due to the weak Feller property of controlled non-linear filters, we can apply quantized Q-learning, to be introduced in Section 9.3, to also belief-based models to also arrive at near optimality of control policies. However, one should note that some subtleties with regard to unique ergodicity properties arise; see [191].

9.2.2 Near-optimality of finite window policies under filter stability and Q-learning convergence

As an alternative approach, we also saw earlier in Section 8.3.2 that finite memory policies are near optimal under filter stability conditions; see the program in Figure 8.1. We now study the reinforcement learning implementation of this approach.

Recall that under mild conditions, for finite state and action MDPs, the Q-learning algorithm given in (9.3) converges to a fixed point which leads to the Discounted Cost Optimality Equation (DCOE).

Learning in POMDPs is challenging, mainly due to the non-Markovian behavior of the observation process. For POMDPs, an attempt may be to study the iterations given by

$$\begin{aligned} Q_{k+1}(y_k, u_k) = (1 - \alpha_k(y_k, u_k))Q_k(y_k, u_k) \\ + \alpha_k(y_k, u_k)\left(C_k(y_k, u_k) + \beta \min_v Q_k(Y_{k+1}, v)\right) \end{aligned}$$

However, the observation process y_t is not a controlled Markov process and the cost that is realized is $c(x_k, u_k)$, which is not a function of y_k and u_k only. A two-part question then is the following:

(i) Would the Q-learning iterates for such a setup indeed converge?,

(ii) And, if they do converge, where do they converge to?

The answer to the first part of the question is positive under mild conditions [290] and [302]; and the answer to the second part of the question is that under filter stability conditions, the convergence is to near optimality with an explicit error bound between the performance loss and the memory window size.

Thus, to answer this, we consider a generalization using a finite window and use again filter stability [189] :

Assume that we start keeping track of the last $N+1$ observations and the last N control action variables after at least $N+1$ time steps. That is, at time t, we keep track of the information variables

$$I_t^N = \begin{cases} \{y_t, y_{t-1}, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}\} & \text{if } N > 0 \\ y_t \quad \text{if } N = 0. \end{cases}$$

We will construct the Q-value iteration using these information variables. In what follows, we will drop the N dependence on I_t^N and sometimes we will use $N = 1$ for simplicity of the notation. For these new approximate states, we follow the usual Q learning algorithm such that for any $I \in \mathbb{Y}^{N+1} \times \mathbb{U}^N$ and $u \in \mathbb{U}$

$$Q_{t+1}(I,u) = (1-\alpha_t(I,u))Q_t(I,u) + \alpha_t(I,u)\left(C_t(I,u) + \beta \min_v Q_t(I_1^t, v)\right), \tag{9.21}$$

where $I_1^t = \{Y_{t+1}, y_t, \ldots, y_{t-N+1}, u_t, \ldots, u_{t-N+1}\}$, we put the t dependence to emphasize that the distribution of Y_{t+1} and hence I_1^t are different for every t.

To choose the control actions, we use polices that choose the control actions randomly and independent of everything else such that at time t

$$u_t = u_i, \text{ w.p } \sigma_i$$

for any $u_i \in \mathbb{U}$ with $\sigma_i > 0$ for all i.

We note that for the convergence of the learning algorithm, it is sufficient for the hidden state process to converge to its invariant distribution under the exploration policy. Hence, any policy that leads the hidden state process to its invariant measure and visits every action with positive probability can be used for the exploration. For example, the control action can also be chosen to be a function of the most recent measurement and randomized (as long as all actions have positive probability of being selected for every measurement realization); this would again lead to a uniquely ergodic hidden state process under our assumptions. The algorithm is summarized in Algorithm 9.2.1.

Algorithm 9.2.1 *Set Parameters: Input: Q_0 (initial Q-function), γ^* (exploration policy), N (memory window length for I^N), L (number of data points), $\{M(I,u)\}_{(I,u)} \equiv 0$ (number of visits to finite memory state action pairs (I,u)).*

Initialize *Start with Q_0*

Iterate *If (I_t, U_t) is the current memory state-action pair $\Longrightarrow$ generate the cost $c(X_t, U_t)$ and the next state $Y_{t+1} \sim \mathcal{T}(\cdot|X_t, U_t)$, and update I_{t+1} given I_t, Y_t, U_t.*

Iterate *set*

$$M(I_t, U_t) = M(I_t, U_t) + 1.$$

Iterate *For $t = 0, \ldots, L-1$,*

Update Q-function Q_t for the inputs (I_t, U_t) as follows:

$$\begin{aligned} Q_{t+1}(I_t, U_t) = (1 - \alpha_t(I_t, U_t))\, Q_t(I_t, U_t) \\ + \alpha_t(I_t, U_t)\left(c(X_t, U_t) + \beta \min_{v \in \mathbb{U}} Q_t(I_{t+1}, v)\right), \end{aligned}$$

where

$$\alpha_t(I_t, U_t) = \frac{1}{1 + M(I_t, U_t)}.$$

43 Generate $U_{t+1} \sim \gamma^*$.

End

Return Q_L

Algorithm 9.2.1 differs from the usual Q-value iterations:

(i) The distribution of I_1^t, which is the consecutive N-window information variable when we hit the (I, u), is generally different for every t and the pair (I, u) is not a controlled Markov process.

In other words, the controlled transitions are time dependent, that is, if we assume $N = 1$ then for some $I = (y_t, y_{t-1}, u_{t-1})$ and $u = u_t$:

$$Pr(I_1^t = (y'_{t+1}, y'_t, u'_t)|z = (y_t, y_{t-1}, u_{t-1}), u_t) = 1_{\{y_t = y'_t, u_t = u'_t\}} Pr(y_{t+1}|y_t, y_{t-1}, u_t, u_{t-1})$$

is not stationary and might change at every time step t, since $Pr(y_{t+1}|y_t, y_{t-1}, u_t, u_{t-1})$ depends on the marginal distribution of x_{t-1} (x_{t-N} in the general case).

(ii) Here, we only observe the cost realizations of the underlying state process $\{x_t\}_t$ and the control actions. For example, if we assume that $N = 1$ then the cost we observe is $c(x_t, u_t)$. However, $c(x_t, u_t)$ depends on (I, u) pair randomly and in a time dependent way so that for some $I = (y_t, y_{t-1}, u_{t-1})$ and $u = u_t$:

$$C_t(I, u) = c(x_t, u_t) \in B, \quad \text{w.p. } Pr(X_t \in \{x : c(x, u_t) \in B\}|y_t, y_{t-1}, u_{t-1})$$

where $Pr(dx_t|y_t, y_{t-1}, u_{t-1})$ can be seen as some *pseudo-belief* on the underlying state variable given $I = (y_t, y_{t-1}, u_{t-1})$, the most recent $N = 1$ information variables. In other words, $Pr(dx_t|y_t, y_{t-1}, u_{t-1})$ is the Bayesian update of π_{t-1}, the marginal distribution of the true state x_{t-1} at the time step $t - 1$, using $I = (y_t, y_{t-1}, u_{t-1})$ and thus, it is time dependent. ◇

We will observe that, if one assumes that the hidden state process, $\{x_t\}_t$ is positive Harris recurrent, or at least, admits a unique invariant probability measure π^* under a stationary exploration policy γ, then the average of approximate state transitions gets closer to

$$P^*(I_{t+1}|I_t, u_t) := \hat{\eta}^N((\pi^*, I_{t+1})|(\pi^*, I_t), u_t) \tag{9.22}$$

with $\hat{\eta}^N$ is defined as in (8.25) and (8.26). In particular, if we assume $N = 1$, then we write

$$P^*(I_{t+1} = (y'_{t+1}, y'_t, u'_t)|I_t = (y_t, y_{t-1}, u_{t-1}), u_t) = \mathbb{1}_{\{y'_t = y_t, u'_t = u_t\}} P^{\pi^*}(y_{t+1}|y_t, y_{t-1}, u_t, u_{t-1}) \tag{9.23}$$

where $P^{\pi^*}(y_{t+1}|y_t, y_{t-1}, u_t, u_{t-1})$ denotes the distribution of y_{t+1} when the marginal distribution on x_{t-1} is given by the invariant measure π^*.

We also have that the sample path averages of the random cost realizations get close to,

$$C^*(I, u) = \hat{c}(\pi^*, I, u) = \int_{\mathbb{X}} c(x, u) P^{\pi^*}(dx|I)$$

where, $P^*(x|I)$ is the Bayesian update of π^*, using I. If we assume $N = 1$, we can write for some $I = (y_1, y_0, u_0)$ and $u = u_1$

$$C^*(y_1, y_0, u_0, u_1) = \hat{c}(\pi^*, (y_1, y_0, u_0), u_1) = \int_{\mathbb{X}} c(x_1, u_1) P^{\pi^*}(dx_1|y_1, y_0, u_0). \tag{9.24}$$

Now consider the following fixed point equation

$$Q^*(I, u) = C^*(I, u) + \beta \sum_{I'} P^*(I'|I, u) \min_v Q^*(I', v) \tag{9.25}$$

where P^* is defined in (9.22) and C^* is defined in (9.24).

The existence of a such fixed point follows from usual contraction arguments. The same fixed equation can also be written as, for $N = 1$, and for $I = (y_1, y_0, u_0)$ and $u = u_1$

$$Q^*\left((y_1, y_0, u_0), u_1\right) = C^*\left((y_1, y_0, u_0), u_1\right) + \beta \sum_{y_2 \in \mathbb{Y}} P^{\pi^*}(y_2 | y_1, y_0, u_1, u_0) \min_{v \in \mathbb{U}} Q^*\left((y_2, y_1, u_1), v\right). \tag{9.26}$$

We note that the stationary distribution π^* does not have to be calculated by the decision maker. The Q value iterations given in (9.21) only use the finite-memory variables I, and π^* is not used in the iterations. We will show that the algorithm naturally converges to (9.25), if the state process is positive Harris recurrent, or at least, admits a unique invariant probability measure π^* under a stationary exploration policy γ, where π^* will be the stationary distribution of the hidden state process x_t under the exploration policy. The performance loss will depend on the stationary distribution π^* that is learned via the exploration policy, however, we will establish further upper bounds that are uniform over such π^* which decrease exponentially with the window size N.

That is, one runs Q-learning algorithm by pretending that the finite window is the *state*. We first need to specify some conditions that would be needed.

Assumption 9.2.1 *(i)* $\alpha_k(I, u) = \frac{1}{k}$ *if* $I_k = I, u_k = u$.

(ii) Under the stationary {memoryless or finite memory exploration} policy, say γ*, the true state process,* $\{X_t\}_t$*, admits a unique invariant probability measure* π^*_γ.

(iii)During the exploration phase, every (I, u) *pair is visited infinitely often.*

Condition (iii) can be relaxed: However, one needs to ensure that the set of all (y, u) pairs which are visited infinitely often during exploration is so that an optimal policy is learned (visited infinitely often), and when this optimal policy (learned via the convergence of Q-learning) is implemented, the closed-loop process always remains in this set; see [191] and [92, Lemma 6 and Corollary 2].

Theorem 9.2.1 *[189] Under the previous assumption, the algorithm given by*

$$Q_{k+1}(I, u) = (1 - \alpha_k(I, u))Q_k(I, u) + \alpha_k(I, u)\left(C_k(I, u) + \beta \min_v Q_k(I_1, v)\right),$$

converges almost surely to Q^* *which satisfies*

$$Q^*(I, u) = C^*(I, u) + \beta \sum_{I'} P^*(I' | I, u) \min_v Q^*(I', v)$$

which are the Q values for the approximate belief MDP.

Corollary 9.2.1 *[189] Under the filter stability conditions, finite window Q learning is nearly optimal:*

(a) Under the exponential filter stability condition, for any policy γ^N *that satisfies* $Q^*(I, \gamma^N(I)) = \min_u Q^*(I, u)$

$$E\left[J_\beta(\pi^-_N, \mathcal{T}, \gamma^N) - J^*_\beta(\pi^-_N, \mathcal{T}) | I^N_0\right] \leq \frac{4\|c\|_\infty}{(1-\beta)^2}\alpha^N.$$

(b) If we only have asymptotic filter stability (uniform in policies) in total variation, as $N \to \infty$,

$$E\left[J_\beta(\pi^-_N, \mathcal{T}, \gamma^N) - J^*_\beta(\pi^-_N, \mathcal{T}) | I^N_0\right] \to 0.$$

Remark 9.2. Under the conditions of Corollary 8.6, the bounds via the Hilbert metric provide complementary sufficient conditions for geometric decay of the approximation error to zero in the memory length.

Remark 9.3. We caution the reader that our result assumes that the cost starts running after time N: that is the effective cost is:

$$E\left[\sum_{k=N}^{\infty}\beta^{k-N}c(x_k,u_k)\right]. \tag{9.27}$$

Of course, this criterion is also applicable if the system starts running prior to time $-N$ and the costs become in effect after time 0.

If this criterion is not applicable, and the first N stages are also crucial, (i) if β is large enough, we can conclude that the first N stages are not as critical for the analysis as their contributions will be minor in comparison with the future stages for the criterion, which can also be seen by noting that for large enough β, the contributions of the first N time stages become negligible:

$$(1-\beta)E\left[\sum_{k=0}^{\infty}\beta^{k}c(x_k,u_k)\right].$$

(ii) On the other hand, if β is not large and if the cost starts running at time 0, then, we can first run the Q-learning algorithm above to find the best N-window policies which optimizes (9.27). The remaining question would be to optimize:

$$E\left[\sum_{k=0}^{N-1}c(x_k,u_k)+V(I_k)\right] \tag{9.28}$$

as a finite-horizon optimal control problem with a terminal cost and the terminal cost V can be estimated by (9.27) via Theorem 8.3.1 and Theorem 9.2.1. The question then becomes how to select the first N actions, leading to a problem with a finite search complexity for a finite horizon problem, without knowing the system dynamics. For this, one can run a MCMC algorithm in parallel simulations to find the optimal policy for the first N time stages. Since the resulting policy minimizing (9.28) will be at least as good as the first N-window policy under the optimal (belief-MDP) policy (which is not designed to optimize (9.28) but the original cost, the bounds presented in Theorem 8.3.2 will be applicable even when the cost criterion includes the first N time stages.

See Theorem 6.4.4 for sufficient conditions for asymptotic filter stability under total variation.

9.3 Q-Learning For Continuous State and Action Spaces: Quantized Q-Learning, its Convergence and Near-Optimality

With the approach above, by quantizing the state space and viewing the quantization output as a measurement (and quantizer as a measurement kernel), and thus as a POMDP; we can arrive at an approximate finite MDP. First note that under Assumption 8.2.1, Theorem 8.2.1 leads to near-optimality of finite actions.

Let $\rho \equiv Q_m$, be a nearest-neighbour quantizer with finite range $\mathbb{X}_m$:

$$Q_m(z) := \arg\min_{z_k\in\mathbb{X}_m} d_{\mathbb{X}}(z,z_k).$$

Then, one runs the following: Using the nearest neighbour map ρ, write for any $(x,u)\in\mathbb{X}\times\mathbb{U}$

$$\begin{aligned} Q_{k+1}(\rho(x),u) &= (1-\alpha_k(\rho(x),u))Q_k(\rho(x),u) \\ &\quad + \alpha_k(\rho(x),u)\left(C_k(\rho(x),u)+\beta\min_v Q_k(\rho(X_1),v)\right) \end{aligned} \tag{9.29}$$

that is for any true value of the state, we use its representative state from the finite set $\mathbb{X}_m$: *Thus, one run Q-learning as if the quantized state is the actual state.*

For exploration, we again use polices that choose the control actions randomly and independent of everything else with positive probability for every action: the invariant measure of the state process x_t (under the exploration policy) should put positive measure on these bins and satisfy an ergodicity condition to be presented in the following:

Assumption 9.3.1 *(i) With $y = \rho(x)$, we let $\alpha_t(y,u) = 0$ unless $(Y_t, U_t) = (y,u)$. Otherwise, let*

$$\alpha_t(y,u) = \frac{1}{1 + \sum_{k=0}^{t} 1_{\{Y_k = I, U_k = u\}}}.$$

(ii) Under the exploration policy γ^*, the state process is uniquely ergodic (and thus has a unique invariant probability measure π_{γ^*}.

(iii)During the exploration phase, every observation-action pair (y,u) is visited infinitely often.

As noted in Remark 4.3, Meyn and Tweedie [233, Theorem 13.0.1] show that for an aperiodic Harris recurrent Markov chain, for each initial state $x \in \mathbb{X}$,

$$\lim_{n\to\infty} \sup_{B \in \mathcal{B}(\mathbb{X})} |P^n(x,B) - \pi(B)| = 0,$$

that is $P^n(x,\cdot)$ converges to π in *total variation*. This assumption is sufficient, but not necessary for the second item to hold.

Algorithm 9.3.1 *Set Parameters: Input: Q_0 (initial Q-function) $q : \mathbb{X} \to \mathbb{Y}$ (quantizer), γ^* (exploration policy), L (number of data points), $\{N(y,u) = 0\}_{(y,u)\in\times\mathbb{U}}$ (number of visits to state-action pairs).*

Initialize *Start with Q_0*

Iterate *If (X_t, U_t) is the current state-action pair $\Longrightarrow$ generate the cost $c(X_t, U_t)$ and the next state $X_{t+1} \sim \mathcal{T}(\cdot\,|X_t, U_t)$,*

Iterate *set*

$$N(q(X_t), U_t) = N(q(X_t), U_t) + 1.$$

Iterate *For $t = 0, \ldots, L-1$,*

Update Q-function Q_t for the inputs $(q(X_t), U_t)$ as follows:

$$\begin{aligned} Q_{t+1}(q(X_t), U_t) = (1 - \alpha_t(q(X_t), U_t))\, Q_t(q(X_t), U_t) \\ + \alpha_t(q(X_t), U_t) \left(c(X_t, U_t) + \beta \min_{v\in\mathbb{U}} Q_t(q(X_{t+1}), v) \right), \end{aligned}$$

where

$$\alpha_t(q(X_t), U_t) = \frac{1}{1 + N(q(X_t), U_t)}.$$

43 Generate $U_{t+1} \sim \gamma^*$.

End

Return Q_L

Theorem 9.3.1 *[185] Under Assumption 9.2.1, for every pair $(y_i, u) \in \mathbb{Y} \times \mathbb{U}$, the algorithm given above converges to*

$$Q^*(y_i, u) = C^*(y_i, u) + \beta \sum_{y_j \in \mathbb{Y}} P^*(y_j | y_i, u) \min_{v\in\mathbb{U}} Q^*(y_j, v).$$

Here, P^* and C^* are defined by

$$C^*(y_i,u) = \int_{B_i} c(x,u)\,\hat{\pi}^*_{y_i}(dx)$$
$$P^*(y_j|y_i,u) = \int_{B_i} \mathcal{T}(B_j|x,u)\,\hat{\pi}^*_{y_i}(dx), \tag{9.30}$$

where

$$\hat{\pi}^*_{y_i}(A) := \frac{\pi_{\gamma^*}(A)}{\pi_{\gamma^*}(B_i)}, \quad \forall A \subset B_i, \quad \forall i \in \{1,\dots,M\}, \tag{9.31}$$

and π_{γ^*} is the invariant measure of the state process under the exploration policy γ^*.

Remark 9.4. Observe the similarity with (8.18). Here, however, the weighting measures $\hat{\pi}^*_{y_i}$ are not arbitrarily selected, and are induced by the exploration policies.

9.3.1 Error Analysis for Convergence of Quantized Q-Learning for Continuous Space MDPs

The model described in (8.18) is the same model given by the equations (9.30). Hence, the results and error bounds from Section 8.2.2 can be used for the error analysis of the Q-learning algorithm given by (9.3.1). For the remainder of this section, we present a series of results for the performance of the policies learned through the approximate Q-learning algorithm in (9.3.1) building on the results from Section 8.2.2. In these corollaries, it is always assumed that $\pi_{\gamma^*}(B_i) > 0$ for all $i \in \{1,\dots,M\}$ where B_i's are the quantization bins and π_{γ^*} is the invariant measure on the state process under the exploration policy γ^*.

Error Analysis for Non-Compact MDPs

The first result is in asymptotic nature and requires very mild conditions for the convergence (i.e., continuity of the stage-wise cost and weak continuity of the transition kernel). It follows from Theorem 9.3.1 and Theorem 8.2.9.

Corollary 9.3.1 *[185] Under Assumption 9.2.1 and Assumption 8.2.1, the Q learning algorithm in (9.3.1) converges to Q^* in Theorem 9.3.1 with probability 1 and for any policy $\hat{\gamma}$ that satisfies $Q^*(x,\hat{\gamma}(x)) = \min_{u\in\mathbb{U}} Q^*(x,u)$ (i.e., greedy policy of Q^*), for any compact $K \subset \mathbb{X}$, we have*

$$\sup_{x_0\in K} \left|J_\beta(x_0,\hat{\gamma}) - J^*_\beta(x_0)\right| \to 0$$

as $L^- \to 0$, where L^- is defined in (8.22).

We recall now that the error bounds to be presented in Corollary 9.3.2 and Corollary 9.3.3 below will involve the function L and the uniform bound $\bar{L}$ which are defined as follows: for some $x \in \mathbb{X}$ where x belongs to a quantization bin B_i whose representative state is y_i (i.e. $q(x) = y_i$) and averaging measure $\hat{\pi}^*_{y_i}$, we have

$$L(x) := \int_{B_i} \|x - x'\|\,\hat{\pi}^*_{y_i}(dx')$$
$$\bar{L} := \max_{i=1,\dots,M} \sup_{x,x'\in B_i} \|x - x'\|.$$

The following result follows from Theorem 9.3.1 and Theorem 8.2.6.

Corollary 9.3.2 *[185] Under Assumption 9.2.1 and Assumption 8.2.4, the Q-learning algorithm in (9.3.1) converges to Q^* in Theorem 9.3.1 with probability 1 and for any policy $\hat{\gamma}$ that satisfies $Q^*(x, \hat{\gamma}(x)) = \min_{u\in\mathbb{U}} Q^*(x, u)$ (i.e., greedy policy of Q^*), for any initial state x_0, we have*

$$\left|J_\beta(x_0, \hat{\gamma}) - J^*_\beta(x_0)\right| \leq 2\left(\alpha_c + \frac{\beta\alpha_T\|c\|_\infty}{1-\beta}\right)\sum_{t=0}^{\infty}\beta^t \sup_{\gamma\in\Gamma} E^\gamma_{x_0}\left[L(X_t)\right].$$

Application to Models with Compact State Spaces

For the case with compact spaces, we obtain sharper bounds in the following.

The following result follows from Theorem 9.3.1 and Theorem 8.2.8.

Corollary 9.3.3 *[185] Under Assumption 9.2.1 and Assumption 8.2.5, the Q learning algorithm in (9.3.1) converges to Q^* in Theorem 9.3.1 with probability 1 and for any policy $\hat{\gamma}$ that satisfies $Q^*(x, \hat{\gamma}(x)) = \min_{u\in\mathbb{U}} Q^*(x, u)$ (i.e., greedy policy of Q^*), we have*

$$\sup_{x_0\in\mathbb{X}}\left|J_\beta(x_0, \hat{\gamma}) - J^*_\beta(x_0)\right| \leq \frac{2\alpha_c}{(1-\beta)^2(1-\beta\alpha_T)}\bar{L}.$$

where $\bar{L}$ is defined in (8.21).

Building on the results presented, we now show that for compact state spaces, the terms $L(x)$ and the uniform bound $\bar{L}$ can be explicitly bounded via cardinality of finite approximating set $\mathbb{Y}$ and dimension d of the state space. To this end, we assume that the state space $\mathbb{X} \subset \mathbb{R}^d$ is compact, and thus totally bounded. Then, for a given M, we can quantize $\mathbb{X}$ by choosing a finite subset $\mathbb{Y} = \{y_1, \ldots, y_M\}$ such that

$$\max_{x\in\mathbb{X}} \min_{y_i\in\mathbb{Y}} \|x - y_i\| \leq \alpha(1/M)^{1/d}$$

for some $\alpha > 0$, which is possible since $\mathbb{X}$ is totally bounded ([111, Theorem 2.3.1]). Using this construction, one can then write the following immediate bounds:

$$\begin{aligned} L(x) &\leq 2\alpha(1/M)^{1/d}, \text{ for all } x \in \mathbb{X}, \\ \bar{L} &\leq 2\alpha(1/M)^{1/d}. \end{aligned}$$

We can then state the following results, which follow from Corollary 9.3.2 and Corollary 9.3.3.

Corollary 9.3.4 *[185] If the state space $\mathbb{X} \subset \mathbb{R}^d$ is compact, under Assumption 9.2.1 and Assumption 8.2.4, the Q-learning algorithm in (9.3.1) converges to Q^* in Theorem 9.3.1 with probability 1 and for any policy $\hat{\gamma}$ that satisfies $Q^*(x, \hat{\gamma}(x)) = \min_{u\in\mathbb{U}} Q^*(x, u)$ (i.e., greedy policy of Q^*), for any initial state x_0, we have*

$$\left|J_\beta(x_0, \hat{\gamma}) - J^*_\beta(x_0)\right| \leq \left(\alpha_c + \frac{\beta\alpha_T\|c\|_\infty}{1-\beta}\right)\frac{4\alpha(1/M)^{1/d}}{1-\beta}$$

Corollary 9.3.5 *[185] If the state space $\mathbb{X} \subset \mathbb{R}^d$ is compact, under Assumption 9.2.1 and Assumption 8.2.5, the Q learning algorithm in (9.3.1) converges to Q^* in Theorem 9.3.1 with probability 1 and for any policy $\hat{\gamma}$ that satisfies $Q^*(x, \hat{\gamma}(x)) = \min_{u\in\mathbb{U}} Q^*(x, u)$ (i.e., greedy policy of Q^*), we have*

$$\sup_{x_0\in\mathbb{X}}\left|J_\beta(x_0, \hat{\gamma}) - J^*_\beta(x_0)\right| \leq \frac{4\alpha_c}{(1-\beta)^2(1-\beta\alpha_T)}\alpha(1/M)^{1/d}$$

9.4 A General Q-Learning Convergence Theorem

In this section, we present a generalization of the convergence results presented earlier in the chapter. In many problems including most of those in applied, health, and social sciences, and financial mathematics, one may not even know whether the problem studied can be formulated as a fully observed Markov Decision Process (MDP), or a partially observable Markov Decision Process (POMDP) or a multi-agent system where other agents are present or not. There are many practical settings where one works with data and does not know the possibly very complex structure under which the data is generated and tries to respond to the environment. For such settings, a common practical approach is to view the system as an MDP, with a perceived state and action (which may or may not define a genuine controlled Markov chain and therefore, the MDP assumption may not hold in actuality), and arrive at corresponding solutions via some learning algorithm.

Toward this end, a general convergence theorem was given in [191, Theorem 2.1], with further implications and refinements reported in [92, Section IV.B]. This is presented in the following:

Let $\{C_t\}_t$ be $\mathbb{R}$-valued, $\{S_t\}_t$ be $\mathbb{S}$-valued and $\{U_t\}_t$ be $\mathbb{U}$-valued three stochastic processes. Consider the following iteration defined for each $(s,u) \in \mathbb{S} \times \mathbb{U}$ pair

$$Q_{t+1}(s,u) = (1 - \alpha_t(s,u))\, Q_t(s,u) + \alpha_t(s,u)\,(C_t + \beta V_t(S_{t+1})) \tag{9.32}$$

where $V_t(s) = \min_{u \in \mathbb{U}} Q_t(s,u)$, and $\alpha_t(s,u)$ is a sequence of constants also called the learning rates. We note that unlike the finite setup considered earlier where $w_t \in (9.8) was a conditionally zero-mean martingale noise, such a property does not app
Markovian nature of the dynamics.

We assume that the process U_t is selected so that the following condition holds.

Assumption 9.4.1 *$\mathbb{S}, \mathbb{U}$ are finite sets, and the joint process $(S_{t+1}, S_t, U_t, C_t)_{t \geq 0}$ is asymptotically ergodic in the sense that for the given initialization random variable S_0, for any measurable bounded function f, we have that with probability one,*

$$\frac{1}{N} \sum_{t=0}^{N-1} f(S_{t+1}, S_t, U_t, C_t) \to \int f(s_1, s, u, c) \pi(ds_1, ds, du, dc)$$

for some measure π such that $\pi(\mathbb{S} \times s \times u \times \mathbb{R}) > 0$ for any $(s,u) \in \mathbb{S} \times \mathbb{U}$.

Remark 9.5. The assumption that $\pi(\mathbb{S} \times s \times u \times \mathbb{R}) > 0$ for any $(s,u) \in \mathbb{S} \times \mathbb{U}$ is in the same spirit as the standard condition for reinforcement algorithms that every state-action pair is visited infinitely often during training. We note that it is possible to relax this condition, if one is only interested in the convergence of the algorithm. In particular, we might consider a measure π such that $\pi(\mathbb{S} \times s \times u \times \mathbb{R}) > 0$ for all $(s,u) \in B \subset \mathbb{S} \times \mathbb{U}$, for some subset B, where the set B represents so called trained state-action pairs. For the learned policies to be optimal, however, one needs to make sure that the controlled process stays within the trained part of the system during the execution of an optimal policy.

The above implies Assumption 9.4.2(ii)-(iii) below:

Assumption 9.4.2 *i.* $\alpha_t(s,u) = 0$ *unless* $(S_t, U_t) = (s,u)$. *Furthermore,*

$$\alpha_t(s,u) = \frac{1}{1 + \sum_{k=0}^{t} 1_{\{S_k = s, U_k = u\}}}$$

and with probability 1, $\sum_t \alpha_t(s,u) = \infty$

ii. For C_t, we have, as $t \to \infty$,

$$\frac{\sum_{k=0}^{t} C_k 1_{\{S_k = s, U_k = u\}}}{\sum_{k=0}^{t} 1_{\{S_k = s, U_k = u\}}} \to C^*(s,u),$$

almost surely for some function C^*.

iii. For the S_t process, we have, for any function f, as $t \to \infty$,

$$\frac{\sum_{k=0}^{t} f(S_{k+1}) 1_{\{S_k=s, U_k=u\}}}{\sum_{k=0}^{t} 1_{\{S_k=s, U_k=u\}}} \to \int f(s_1) P^*(ds_1|s,u)$$

almost surely for some P^*.

Note that a stationarity assumption is not required. Under Assumption 9.4.1, we have that with $f(S_{t+1}, S_t, U_t, C_t) = C_t 1_{\{S_t=s, U_t=u\}}$, as $N \to \infty$,

$$\frac{1}{N} \sum_{t=0}^{N-1} C_t 1_{\{S_t=s, U_t=u\}} \to \int_{C \in \mathbb{R}} C\pi(S=s, U=u, dC).$$

We also have that with $f(S_{t+1}, S_1, U_t, C_t) = 1_{\{S_t=s, U_t=u\}}$, as $N \to \infty$,

$$\frac{1}{N} \sum_{t=0}^{N-1} 1_{\{S_t=s, U_t=u\}} \to \pi(S=s, U=u)$$

almost surely. Hence, we can write

$$\frac{\frac{1}{t+1}\sum_{k=0}^{t} C_k 1_{\{S_k=s, U_k=u\}}}{\frac{1}{t+1}\sum_{k=0}^{t} 1_{\{S_k=s, U_k=u\}}} \to \int C\pi(dC|S=s, U=u) =: C^*(s,u)$$

which implies Assumption 9.4.2 (ii). Similarly, one can also establish Assumption 9.4.2 (iii) under Assumption 9.4.1.

As before, let $\mathbb{S}, \mathbb{U}$ be finite sets. Consider the following equation

$$Q^*(s,u) = C^*(s,u) + \beta \sum_{s_1 \in \mathbb{S}} V^*(s_1) P^*(s_1|s,u) \tag{9.33}$$

for some functions Q^*, C^*, to be defined explicitly, and for some regular conditional probability distribution $P^*(\cdot|s,u)$, where $V^*(u) := \min_u Q^*(s,u)$.

Theorem 9.4.1 *[191, Theorem 2.1] Under Assumption 9.4.2, $Q_t(s,u) \to Q^*(s,u)$ almost surely for each $(s,u) \in \mathbb{S} \times \mathbb{U}$ pair where Q^* satisfies (9.33), for any initialization of Q_0.*

We note that this result generalizes those presented earlier in this chapter and in particular Theorem 9.2.1. See also [78, 107, 189].

9.5 Bibliographic Notes

Q-learning was introduced and studied in [323], [306], [26]. An ODE approach to Q-learning presents a rather direct proof of convergence [64] [319]. Two comprehensive resources on reinforcement learning are [301] and [231].

Quantized Q-learning and its convergence and near-optimality is studied in [185].

Q-learning for partially observed MDPs have been studied in [138, 229, 289, 302]. Our analysis here builds on [189] which also establishes convergence to near optimality.

9.6 Exercises

Exercise 9.6.1 *Consider a controlled Markov chain with state space $\mathbb{X} = \{0, 1\}$, action space $\mathbb{U} = \{0, 1\}$, and transition kernel for $t \in \mathbb{Z}_+$:*

$$P(x_{t+1} = 1|x_t = 0, u_t = 1) = P(x_{t+1} = 1|x_t = 1, u_t = 1) = \alpha$$
$$P(x_{t+1} = 1|x_t = 0, u_t = 0) = P(x_{t+1} = 1|x_t = 1, u_t = 0) = 1 - \alpha.$$

where $\alpha \in (0, 1)$. Let a cost function $c(x, u)$, with $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ be given by

$$c(0, 1) = c(0, 0) = 1 \qquad c(1, 0) = c(1, 1) = 2.$$

Suppose that the goal is to minimize the quantity

$$E_0^\gamma[\sum_{t=0}^{\infty} \beta^t c(x_t, u_t)],$$

for a fixed $\beta \in (0, 1)$, over all admissible policies $\gamma \in \Gamma_A$.

Find an optimal policy and the optimal expected cost explicitly, as a function of α, β (note that the initial condition is $x_0 = 0$).

Exercise 9.6.2 *Consider the following problem: Let $\mathbb{X} = \{1, 2\}, \mathbb{U} = \{1, 2\}$, where $\mathbb{X}$ denotes whether a fading channel is in a good state ($x = 2$) or a bad state ($x = 1$). There exists an encoder who can either try to use the channel ($u = 2$) or not use the channel ($u = 1$). The goal of the encoder is send information across the channel.*

Suppose that the encoder's cost (to be minimized) is given by:

$$c(x, u) = -1_{\{x=2,u=2\}} + \alpha(u - 1),$$

for $\alpha = 1/2$ (if you view this as a maximization problem, you can see that the goal is to maximize information transmission efficiency subject to a cost involving an attempt to use the channel; the model can be made more complicated but the idea is that when the channel state is good, $u = 2$ can represent a channel input which contains data to be transmitted and $u = 1$ denotes that the channel is not used).

Suppose that the transition kernel is given by:

$$P(x_{t+1} = 2|x_t = 2, u_t = 2) = 0.8, \qquad P(x_{t+1} = 1|x_t = 2, u_t = 2) = 0.2$$
$$P(x_{t+1} = 2|x_t = 2, u_t = 1) = 0.2, \qquad P(x_{t+1} = 1|x_t = 2, u_t = 1) = 0.8$$
$$P(x_{t+1} = 2|x_t = 1, u_t = 2) = 0.5, \qquad P(x_{t+1} = 1|x_t = 1, u_t = 2) = 0.5$$
$$P(x_{t+1} = 2|x_t = 1, u_t = 1) = 0.9, \qquad P(x_{t+1} = 1|x_t = 1, u_t = 1) = 0.1$$

We will consider either a discounted cost criterion for some $\beta \in (0, 1)$ (you can fix an arbitrary value)

$$\inf_\gamma E_x^\gamma[\sum_{t=0}^{\infty} \beta^t c(x_t, u_t)] \tag{9.34}$$

or the average cost criterion

$$\inf_\gamma \limsup_{T\to\infty} \frac{1}{T} E_x^\gamma[\sum_{t=0}^{T-1} c(x_t, u_t)]. \tag{9.35}$$

a) Using Matlab or some other program, obtain a solution to the problem given above in (9.34) through the following:

(i) Policy Iteration

(ii) Value Iteration.

(iii)Q-Learning. Note that a common way to pick α coefficients in the Q-learning algorithm is to take for every x, u pair:

$$\alpha_t(x,u) = \frac{1}{1+\sum_{k=0}^{t} 1_{\{x_k=x, u_k=u\}}}$$

b) Consider the criterion given in (9.35). Apply the convex analytic method, by solving the corresponding linear program, to find the optimal policy. In Matlab, the command linprog can be used to solve linear programming problems. See (7.42).

Exercise 9.6.3 *Revisit Exercise 9.6.3, part a). Apply Q-Learning, noting that a common approach to pick α coefficients in the Q-learning algorithm is to take for every x, u pair:*

$$\alpha_t(x,u) = \frac{1}{1+\sum_{k=0}^{t} 1_{\{x_k=x, u_k=u\}}}$$

Exercise 9.6.4 *One can apply Q-learning even when the model is not known, but for the results to be optimal, it is essential that the system we are dealing with is an actual MDP.*

a) However, imagine that we have a POMDP but we run the Q-learning algorithm as if the system is an MDP. Sections 9.2 and 9.3 have shown that Q-learning, when a finite window memory of the most recent measurements and actions is viewed as a state, converges even in this case, and even to near optimality under mild conditions related to either filter stability or appropriate approximation bounds. Study Exercise 7.7.6 and apply finite memory Q-learning for this example.

b) [Quantized Q-learning]As a further instance we have considered the finite-state quantization of a continuous space MDP and view this as a POMDP.

Consider the setup in Exercise 9.6.3. Revise the problem above with the following transition kernel so that $\mathbb{X} = [0,1]$ (thus the channel's quality is not binary) and for each Borel $A \in [0,1]$

$$P(x_{t+1} \in A | x_t = z_0, u_t = 1) = 2\int_A (1-x)dx, \qquad P(x_{t+1} \in A | x_t = z_0, u_t = 0) = 2\int_A x dx,$$

for all $z_0 \in [0,1]$, and suppose that the encoder's per-stage cost (to be minimized) is given by:

$$c(x,u) = -xu + \eta u.$$

for some $\eta \in \mathbb{R}$. Apply quantized Q-learning by quantizing the channel state with uniform quantization of increasing granularity.

Exercise 9.6.5 *Recall Exercise 7.7.6: Consider a POMDP given by the following description. Let there be two possible states that a machine can take: $\mathbb{X} = \{0,1\}$, where 0 is the bad ('system is down') state and 1 is the good state. Let $\mathbb{U} = \{0,1\}$, where 0 is the 'do nothing' control and 1 is the 'repair' control. Suppose that the transition probabilities are given by:*

$$\begin{aligned}
&P(X_{t+1}=1|X_t=1,U_t=0) = 1-\eta_1, \qquad P(X_{t+1}=0|X_t=1,U_t=0) = \eta_1 > 0\\
&P(X_{t+1}=1|X_t=1,U_t=1) = 1-\eta_2, \qquad P(X_{t+1}=0|X_t=1,U_t=1) = \eta_2 > 0\\
&P(X_{t+1}=1|X_t=0,U_t=0) = 0, \qquad P(X_{t+1}=0|X_t=0,U_t=0) = 1\\
&P(X_{t+1}=1|X_t=0,U_t=1) = \alpha > 0, \qquad P(X_{t+1}=0|X_t=0,U_t=1) = 1-\alpha
\end{aligned} \tag{9.36}$$

Thus, η_1 is the failure probability when the state is good (and no repair) and η_2 is the failure probability when the state is good (and when there is repair) with $\eta_1 > \eta_2$, and α is the success probability in the event of a repair.

The controller has access only to $\{0,1\}$-valued measurement variables $Y_0, \cdots, Y_t$ and $U_0, \cdots, U_{t-1}$, at time t, where the measurements are generated by a binary symmetric channel:

$$(Y = x|X = x) = 1 - \epsilon, \qquad P(Y = 1 - x|X = x) = \epsilon,$$

for all $x \in \{0,$*. The per-stage cost function* $c(x, u)$ *is given by* $c(0, 0) = C, c(1, 0) = 0, c(0, 1) = c(1, 1) = R$ *with* $0 < R < C$*.*

Apply Q-learning using the finite window $I_t^N = \{[y_{t-N}, t]\}, [u_{t-N}, t - 1]\}\}$ *as an effective state, as described in Section 9.2 of the lecture notes. When can you guarantee convergence to near-optimality as the window size increases? Compute the numerical performance for several* N *values. Reflect on the trade-off between performance, computational demands, and memory length for such a control design.*

10

Decentralized and Multi-Agent Stochastic Control

10.1 Introduction

In classical stochastic control problems considered so far in this document, we were given a system of the form

$$x_{t+1} = f(x_t, u_t, w_t), \quad t \in \mathbb{Z}_+,$$

where actions are to be generated using some control policy $\gamma = \{\gamma_t\}$ with

$$u_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_+,$$

where I_t is the information available at t. Here, w_t is i.i.d. noise. If $I_t = \{x_0, \cdots, x_t; u_0, \cdots, u_{t-1}\}$, we have a *fully observed* system. If the controller only has measurements

$$y_t = g(x_t, v_t),$$

$I_t = \{y_0, \cdots, y_t; u_0, \cdots, u_{t-1}\}$, we have a *partially observed* system. These have been studied extensively in the previous chapters.

As we observed earlier in these notes, given an optimality criterion (e.g. expected finite horizon cost, discounted cost, average cost, terminal cost), for such classical stochastic control setups, there are few powerful techniques to establish the existence/computation of optimal policies:

(i) The dynamic programming approach and backward induction: Weak-continuity / strong continuity properties and measurable selection conditions leads to existence / explicit computations.

(ii) The strategic measures approach (see Section 5.4).

(iii) For infinite horizon problems, linear programming/convex analytic techniques.

All of these crucially build on the fact that $I_t \subset I_{t+1}$, that is, information is expanding. In the absence of this condition, which facilitates the applicability of the iterated expectations theorem (Theorem 4.1.3), much of the standard analysis on existence/structure/recursions is no longer applicable: The reader is referred to the derivation at the beginning of *Chapter* 5.

However, a very important class of optimal stochastic control problems involve setups where a number of decentralized decision makers are present. In this context, we will consider a collection of decision makers (DMs) where each has access to some local information variable: Such a collection of decision makers who wish to minimize a common cost function and who has an agreement on the system (that is, the probability space on which the system is defined, and the policy and action spaces) is said to be a *stochastic team*. Such problems are called *decentralized stochastic control* problems.

To gain some insight, let us consider the following model.

$$x_{t+1} = f(x_t, u_t^1, \ldots, u_t^L, w_t),$$

with each decision maker DM m arriving at their action u^m at time t using only local information:

$$u_t^m = \gamma_t^m(I_t^m), t \in \mathbb{Z}_+,$$

where I_t^m denotes some information variable. Decentralized stochastic control theory requires more general approaches when compared with the classical setup that we have considered up until this chapter, primarily due to the informational subtleties, to be presented further in the following.

To study such problems in a systematic fashion, we will present a classification for decentralized stochastic control models based on the informational and dynamical relations between the decision makers in the following. Toward this goal, in the following we introduce Witsenhausen's intrinsic model.

10.2 Solution Concepts, Information Structures and Witsenhausen's Intrinsic Model

10.2.1 Witsenhausen's intrinsic model

Witsenhausen's contributions (e.g., [327, 328, 331]) to decentralized stochastic control and characterization of information structures have been crucial in our understanding of stochastic team theory. In this section, we introduce the characterizations as laid out by Witsenhausen, termed as *the Intrinsic Model* [328]; see [349] for a comprehensive overview and further characterizations and classifications of information structures. In this model (described in discrete time), any action applied at any given time is regarded as applied by an individual decision maker/agent, who acts only once. One advantage of this model, in addition to its generality, is that the characterizations regarding information structures can be compactly described.

Suppose that in the decentralized system considered below, there is a pre-defined order in which the decision makers act. Such systems are called *sequential systems* (for non-sequential teams, we refer the reader to Andersland and Teneketzis [8], [9] and Teneketzis [304], in addition to Witsenhausen [326] and [349, p. 113]). Suppose that in the following, the action and measurement spaces are standard Borel spaces, that is, Borel subsets of Polish (complete, separable and metric) spaces. In the context of a sequential system, the *Intrinsic Model* of Witsenhausen [329] is the following characterization of information structures, where we consider a decentralized stochastic control model with N decision makers (DMs) (also called agents). In the model, there exist the following:

- A collection of *measurable spaces* $\{(\Omega, \mathcal{F}), (\mathbb{U}^i, \mathcal{U}^i), (\mathbb{Y}^i, \mathcal{Y}^i), i \in \mathcal{N}\}$, specifying the system's distinguishable events, and the control and measurement spaces. Here $N = |\mathcal{N}|$ is the number of control actions taken, and each of these actions is taken by an individual (different) DM (hence, even a DM with perfect recall can be regarded as a separate decision maker every time it acts). The pair $(\Omega, \mathcal{F})$ is a measurable space (on which an underlying probability may be defined). The pair $(\mathbb{U}^i, \mathcal{U}^i)$ denotes the measurable space from which the action, u^i, of decision maker i is selected. The pair $(\mathbb{Y}^i, \mathcal{Y}^i)$ denotes the measurable observation/measurement space for DM i.
- A *measurement constraint* which establishes the connection between the observation variables and the system's distinguishable events. The $\mathbb{Y}^i$-valued observation variables are given by $y^i = \eta^i(\omega, \mathbf{u}^{[1,i-1]})$, $\mathbf{u}^{[1,i-1]} = \{u^k, k \leq i-1\}$, η^i measurable functions and u^k denotes the action of DM k. Hence, the information variable y^i induces a σ-field, $\sigma(\mathcal{I}^i)$ over $\Omega \times \prod_{k=1}^{i-1} \mathbb{U}^k$. The collection $\{\mathcal{J}^i; i = 1, \ldots, N\}$ or $\{\eta^i; i = 1, \ldots, N\}$ is called the *information structure* of the system.
- A *design constraint* which restricts the set of admissible N-tuple control laws $\underline{\gamma} = \{\gamma^1, \gamma^2, \ldots, \gamma^N\}$, also called *designs* or *policies*, to the set of all measurable control functions, so that $u^i = \gamma^i(y^i)$, with $y^i = \eta^i(\omega, \mathbf{u}^{[1,i-1]})$, and γ^i, η^i measurable functions. Let Γ^i denote the set of all admissible policies for DM i and let $\mathbf{\Gamma} = \prod_k \Gamma^k$.

We note that, the intrinsic model of Witsenhausen gives a set-theoretic characterization of information fields, however, for standard Borel spaces, the model above is equivalent to that of Witsenhausen's; see Exercise 1.6.6.

One can also introduce a fourth component.

- A *probability measure* P defined on $(\Omega, \mathcal{F})$ which describes the measures on the events in the model.

10.2.2 Solution concepts

Thus, we will assume that we are given a probability measure P on $(\Omega, \mathcal{F})$. Additionally, we have a loss (or cost) function $c : \Omega_0 \times (\mathbb{U}^1 \times \cdots \times \mathbb{U}^N) \to \mathbb{R}_+$ to be optimized where Ω_0 is an appropriate signal space.

Let

$$\underline{\gamma} = \{\gamma^1, \cdots, \gamma^N\} \in \mathbf{\Gamma}.$$

We then have,

$$J(\underline{\gamma}) = E[c(\omega_0, \mathbf{u})] = E[c(\omega_0, \gamma^1(y^1), \cdots, \gamma^N(y^N))], \tag{10.1}$$

for some non-negative measurable loss (or cost) function $c : \Omega \times \prod_k \mathbb{U}^k \to \mathbb{R}_+$. Here, we have the notation $\mathbf{u} = \{u^t, t \in \mathcal{N}\}$. Here, ω_0 may be viewed as the cost function relevant exogenous variable and is contained in ω.

Definition 10.2.1 *For a given stochastic team problem with a given information structure, $\{J; \Gamma^i, i \in \mathcal{N}\}$, a policy (strategy) N-tuple $\underline{\gamma}^* := (\gamma^{1^*}, \ldots, \gamma^{N^*}) \in \mathbf{\Gamma}$ is an optimal team decision rule (team-optimal decision rule or simply team-optimal solution) if*

$$J(\underline{\gamma}^*) = \inf_{\underline{\gamma} \in \mathbf{\Gamma}} J(\underline{\gamma}) =: J^*, \tag{10.2}$$

provided that such a strategy exists. The cost level achieved by this strategy, J^*, is the minimum (or *optimal*) *team cost.*

Definition 10.2.2 *For a given N-person stochastic team with a fixed information structure, $\{J; \Gamma^i, i \in \mathcal{N}\}$, an N-tuple of strategies $\underline{\gamma}^* := (\gamma^{1^*}, \ldots, \gamma^{N^*})$ constitutes a Nash equilibrium (synonymously, a person-by-person optimal (pbp optimal) solution) if, for all $\beta \in \Gamma^i$ and all $i \in \mathcal{N}$, the following inequalities hold:*

$$J^* := J(\underline{\gamma}^*) \leq J(\underline{\gamma}^{-i*}, \beta), \tag{10.3}$$

where we have adopted the notation

$$(\underline{\gamma}^{-i*}, \beta) := (\gamma^{1^*}, \ldots, \gamma^{i-1^*}, \beta, \gamma^{i+1^*}, \ldots, \gamma^{N^*}). \tag{10.4}$$

For notational simplicity, let for any $1 \leq k \leq N$, $\gamma^{-k} := \left\{\gamma^i, i \in \{1, \cdots, N\} \setminus \{k\}\right\}$. In the following, we will denote by bold letters the ensemble of random variables across the DMs; that is, $\mathbf{y} = \{y^i, i = 1, \cdots, N\}$ and $\mathbf{u} = \{u^i, i = 1, \cdots, N\}$.

Example 10.1. Consider the following model of a system with two decision makers [349]. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\mathcal{F}$ be the power set of Ω. Let the action space be $\mathbb{U}^1 = \{U(\text{up}), D(\text{down})\}$, $\mathbb{U}^2 = \{L(\text{left}), R(\text{right})\}$, and $\mathcal{U}^1$ and $\mathcal{U}^2$ be the power sets of $\mathbb{U}^1$ and U^2 respectively.

Suppose the probability measure P is given by $P(\omega_i) = p_i, i = 1, 2, 3$ and $p_1 = p_2 = 0.3$, $p_3 = 0.4$, and the loss function $c(\omega, u^1, u^2)$ is given by the following matrices

		u^2: L	u^2: R
u^1	U	1	0
	D	3	1

$\omega : \omega_1 \leftrightarrow 0.3$

		u^2: L	u^2: R
u^1	U	2	3
	D	2	1

$\omega_2 \leftrightarrow 0.3$

		u^2: L	u^2: R
u^1	U	1	2
	D	0	2

$\omega_3 \leftrightarrow 0.4$

Case 1. First, let us consider the case where both agents have access to the true state of nature, and hence $\mathbf{Y}^1 = \mathbf{Y}^2 = \sigma(\{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\})$, the $\sigma-$field generated by the singletons.

In this case, the unique team-optimal decision rules are:

$$\gamma^{1*}(\omega) = \begin{cases} U, \ \omega = \omega_1 \\ \\ D, \text{ else} \end{cases} \qquad \gamma^{2*}(\omega) = \begin{cases} L, \ \omega = \omega_3 \\ \\ R, \text{ else} \end{cases},$$

which we may write symbolically as

$$\underline{\gamma}^* = (UDD, RRL).$$

We point to the observation that even though the policy pair (UDD, RRL) is unique as a team-optimal solution (which is also, by definition, *pbp* optimal), it is not the unique *pbp* optimal solution. The policy pair (UUD, RLL) is also *pbp* optimal, but it is suboptimal.

Case 2. Let the information fields $\mathcal{J}^1 = \{\emptyset, \{\omega_1\}, \{\omega_2, \omega_3\}, \Omega\}$ and $\mathcal{J}^2 = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3\}, \Omega\}$.

For the above model, the unique optimal control strategy is given by

$$\gamma^{1,*}(y^1) = \begin{cases} U, & y^1 = \{\omega_1\} \\ D, & \text{else} \end{cases}$$

$$\gamma^{2,*}(y^2) = \begin{cases} R, & y^2 = \{\omega_1, \omega_2\} \\ L, & \text{else} \end{cases}$$

The development of a systematic solution approach in optimal decentralized stochastic control requires a cautious classification of such problems, primarily in view of information structures.

10.2.3 Classification of information structures

Static vs. dynamic information structures

Under the intrinsic model presented, an Information structure (IS) is *dynamic* if the information available to at least one DM is affected by the action of at least one other DM. An IS is *static*, if the information available at every decision maker is only affected by exogenous disturbances:

(i) A sequential team is *static*, if the information available at every decision maker is only affected by exogenous disturbances (Nature); that is no other decision maker can affect the information at any given decision maker.

(ii) A sequential team problem is *dynamic* if the information available to at least one DM is affected by the action of at least one other DM.

Figure 10.1 is a depiction for a static team problem.

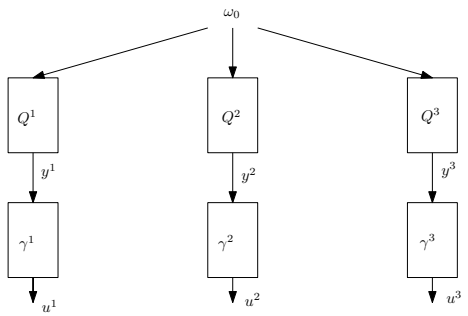

Fig. 10.1: An example of a static information structure. Here, $Q^i(y^i \in \cdot|\omega_0) := P(\eta^i(\omega) \in \cdot|\omega_0)$, $i = 1, 2, 3$.

Classical, quasi-classical (partially nested), and nonclassical information structures

(i) An IS $\{y^i, 1 \leq i \leq N\}$ is *classical* if y^i contains all of the information available to DM k for $k < i$. E.g.: Classical stochastic control problems; the material in these lecture notes up to this chapter (involving fully observed or partially observed models) has thus been with regard to *classical* information structures.

(ii) An IS is *quasi-classical* or *partially nested*, if whenever u^k, for some $k < i$, affects y^i, y^i contains y^k.

(iii)An IS which is not partially nested is *nonclassical*.

Fully and partially observable models reviewed earlier in the chapter are classical in the sense described above. Non-classical problems can be very challenging. As we will see in Section 10.4.2, even a linear system with Gaussian noise can lead to a problem which is very difficult to study and can admit an optimal solution which is not linear; this example is known as Witsenhausen's counterexample [325]. This is called a counterexample since it shows that even linear quadratic Gaussian (LQG) models can admit solutions which are not linear and hence it is a counterexample to a natural conjecture (that all solutions to such LQG problems should be linear) when the information structure is nonclassical.

Quasi-classical information structures possess useful characteristics which allow for solution methods tailored for such models, as we will see in the chapter.

10.2.4 A state space model

A sub-class of sequential teams involve setups where there is a controlled state model, where unlike the classical single-agent model, the state realizations are not available to the agents. In a state space model, one assumes that the decentralized control system has a state x_t that is evolving with time. The evolution of the state is controlled by the actions of the agents (control stations). We may assume that the system has N control stations where each control station i chooses a control action u_t^i at time t. The system considered runs in discrete time, either for a finite or an infinite horizon. In the context of Witsenhausen's intrinsic model, any decision maker applying an action at a given time stage is interpreted as a different decision maker.

Let $\mathbb{X}$ denote the space of realizations of the state x_t, and $\mathbb{U}^i$ denote the space of realization of control actions u_t^i. Let T denote the set of time for which the system runs.

The initial state x_1 is a random variable and the state of the system evolves as

$$x_{t+1} = f_t(x_t, u_t^1, \ldots, u_t^N; w_t^0), \quad t \in \mathcal{T}, \tag{10.5}$$

where $\{w_t^0, t \in T\}$ is an independent noise process that is also independent of x_1. We assume that each control station i observes the following at time t

$$y_t^i = g_t^i(x_t, w_t^i), \tag{10.6}$$

where $\{w_t^i, t \in T\}$ are measurement noise processes that are independent across time, independent of each other, and independent of $\{w_t^0, t \in T\}$ and x_1.

The above evolution does not completely describe a dynamical control system, because we have not yet specified the data available at each control station. In general, the random variable I_t^i available at control station i at time t will be a function of all the past system variables $\{x_{[1,t]}, \mathbf{y}_{[1,t]}, \mathbf{u}_{[1,t-1]}, \mathbf{w}_{[1,t]}\}$, i.e.,

$$I_t^i = \eta_t^i(x_{[1,t]}, \mathbf{y}_{[1,t]}, \mathbf{u}_{[1,t-1]}, \mathbf{w}_{[1,t]}), \tag{10.7}$$

where we use the notation $\mathbf{u} = \{u^1, \ldots, u^N\}$ and $x_{[1,t]} = \{x_1, \ldots, x_t\}$. The collection $\{I_t^i, i = 1, \ldots, N, t \in T\}$ is called the *information structure* of the system, in analogy with Witsenhausen's intrinsic model.

When T is finite, say equal to $\{1, \ldots, T\}$, the above model is a special case of the sequential intrinsic model presented above. The set $\{x_1, w_t^0, w_t^1, \ldots, w_t^N, t \in T\}$ denotes the primitive random variable with probability measure given by the product measure of the marginal probabilities; the system has $N \times T$ DMs, one for each control station at each time. DM (i, t) observes I_t^i and chooses u_t^i. The information sub-fields $\mathcal{J}^k$ are determined by $\{\eta_t^i, i = 1, \ldots, N, t \in T\}$.

Some important information structures are

1. *Complete information sharing*: In complete information sharing, each DM has access to present and past measurements and past actions of all DMs. Such a system is equivalent to a centralized system.

$$I_t^i = \{\mathbf{y}_{[1,t]}, \mathbf{u}_{[1,t-1]}\}, t \in T.$$

2. *Complete measurement sharing*: In complete measurement sharing, each DM has access to the present and past measurements of all DMs. Note that past control actions are not shared.

$$I_t^i = \{y_{[1,t]}\}, \quad t \in T.$$

3. *Delayed information sharing* In delayed information sharing, each DM has access to n-step delayed measurements and control actions of all DMs.

$$I_t^i = \begin{cases} \{y^i_{[t-n+1,t]}, u^i_{[t-n+1,t-1]}\mathbf{y}_{[1,t-n]}, \mathbf{u}_{[1,t-n]}\}, & t > n \\ \{y^i_{[1,t]}, u^i_{[1,t-1]}\}, & t \le n \end{cases} \tag{10.8}$$

4. *Delayed measurement sharing* In delayed measurement sharing, each DM has access to n-step delayed measurements of all DMs. Note that control actions are not shared.

$$I_t^i = \begin{cases} \{y^i_{[t-n+1,t]}, u^i_{[1,t-1]}, \mathbf{y}_{[1,t-n]}\}, & t > n \\ \{y^i_{[1,t]}, u^i_{[1,t-1]}\}, & t \le n \end{cases}$$

5. *Delayed control sharing* In delayed control sharing, each DM has access to n-step delayed control actions of all DMs. Note that measurements are not shared.

$$I_t^i = \begin{cases} \{y^i_{[1,t]}, u^i_{[t-n+1,t-1]}, \mathbf{u}_{[1,t-n]}\}, & t > n \\ \{y^i_{[1,t]}, u^i_{[1,t-1]}\}, & t \le n \end{cases}$$

6. *Periodic information sharing* In periodic information sharing, the DMs share their measurements and control periodically after every k time steps. No information is shared at other time instants.

$$I_t^i = \begin{cases} \{y^i_{[\lfloor t/k \rfloor k,t]}, u^i_{[\lfloor t/k \rfloor k,t]}, \mathbf{y}_{[1,\lfloor t/k \rfloor k]}, \mathbf{u}_{[1,\lfloor t/k \rfloor k]}\}, & \\ & t \ge k \\ \{y^i_{[1,t]}, u^i_{[1,t-1]}\}, & t < k \end{cases}$$

7. *Completely decentralized information* In a completely decentralized system, no data is shared between the DMs.

$$I_t^i = \{y^i_{[1,t]}, u^i_{[1,t-1]}\}, \quad t \in T.$$

In all the information structures given above, each DM has perfect recall (PR), that is, each DM has full memory of its past information. In general, a DM need not have perfect recall. For example, a DM may only have access to its current observation, in which case the information structure is

$$I_t^i = \{y_t^i\}, \quad t \in T. \tag{10.9}$$

To complete the description of the team problem, we have to specify the loss function. For some applications, one may have that the loss function is of additive form:

$$c(x_{[1,T]}, \mathbf{u}_{[1,T]}) := \sum_{t \in T} c(x_t, \mathbf{u}_t) \tag{10.10}$$

where each term in the summation is known as the *incremental* (or *stagewise*) *loss*. The objective would be to choose control policies γ_t^i such that $u_t^i = \gamma_t^i(I_t^i)$ so as to minimize the expected loss (10.10).

10.3 Solutions to Static Teams

Definition 10.3.1 *Given a static stochastic team problem $\{J; \Gamma^i, i \in \mathcal{N}\}$, a policy N-tuple $\underline{\gamma} \in \mathbf{\Gamma}$ is stationary if (i) $J(\underline{\gamma})$ is finite, (ii) the N partial derivatives in the following equations are well defined, and (iii) $\underline{\gamma}$ satisfies these equations:*

$$\left[\nabla_{u^i} E_{\omega|y^i} c(\omega_0; \underline{\gamma}^{-i}(\mathbf{y}), u^i)\right] |_{u^i=\gamma^i(y^i)} = 0, a.s. \quad i \in \mathcal{N}. \tag{10.11}$$

There is a close connection between stationarity and person-by-person-optimality, as we discuss in the following. The results to be presented below are due to Krainak et. al. [196] and [349], generalizing Radner [260]. We follow the presentation in [349], which also contains the proofs of the results.

Theorem 10.3.1 *[260] [196] Let $\{J; \Gamma^i, i \in \mathcal{N}\}$ be a static stochastic team problem where $\mathbb{U}^i = \mathbb{R}^{m_i}, i \in \mathcal{N}$, the loss function $c(\omega_0, \mathbf{u})$ is convex and continuously differentiable in $\mathbf{u}$ a.s., and $J(\underline{\gamma})$ is bounded from below on $\mathbf{\Gamma}$. Let $\underline{\gamma}^* \in \mathbf{\Gamma}$ be a policy N-tuple with a finite cost ($J(\underline{\gamma}^*) < \infty$), and suppose that for every $\underline{\gamma} \in \mathbf{\Gamma}$ such that $J(\underline{\gamma}) < \infty$, the following holds:*

$$\sum_{i \in \mathcal{N}} E\{\nabla_{u^i} c(\omega_0; \mathbf{u})|_{\mathbf{u}=\underline{\gamma}^*(\mathbf{y})[\gamma^i(y^i) - \gamma^{i*}(y^i)]\} \geq 0, \tag{10.12}$$

where $E\{\cdot\}$ denotes the total expectation and the notation $\nabla_{u^i} c(\omega_0; \underline{\gamma}^*(\mathbf{y}))$ means that the partial derivatives are evaluated under policy $\underline{\gamma}^*$. Then, $\underline{\gamma}^*$ is a team-optimal policy, and it is unique if c is strictly convex in $\mathbf{u}$.

Proof Sketch. First, by the convexity of c, we obtain

$$\frac{1}{\alpha}\left[c(\omega_0; \underline{\gamma}^*(\mathbf{y}) + \alpha[\underline{\gamma}(\mathbf{y}) - \underline{\gamma}^*(\mathbf{y})]) - c(\omega_0; \underline{\gamma}^*(\mathbf{y}))\right] \leq c(\omega_0; \underline{\gamma}(\mathbf{y})) - c(\omega_0; \underline{\gamma}^*(\mathbf{y})),$$

for all $\alpha \in (0, 1]$. Using the definition of J, this inequality can equivalently be written as (by taking the total expectation):

$$h(\alpha) := \frac{1}{\alpha}[E\{c(\omega_0; \underline{\gamma}^*(\mathbf{y}) + \alpha[\underline{\gamma}(\mathbf{y}) - \underline{\gamma}^*(\mathbf{y})])\} - J(\underline{\gamma}^*)] \leq J(\underline{\gamma}) - J(\underline{\gamma}^*),$$

where $\alpha \in (0, 1]$. Note that both $J(\underline{\gamma})$ and $J(\underline{\gamma}^*)$ are finite, by hypothesis, and the first random variable (i.e., the first loss function) also has a finite expectation for every $\alpha \in (0, 1]$ because of the bound provided by the inequality. Now, due to the convexity of c, its finite integral, $E\{c(\omega_0; \underline{\gamma}^*(\mathbf{y}) + \alpha[\underline{\gamma}(\mathbf{y}) - \underline{\gamma}^*(\mathbf{y})])\}$ is also convex in α. This leads to the conclusion that (by a property of convex functionals that $h(\alpha)$ is a monotonically nonincreasing function as $\alpha \downarrow 0$, and furthermore $h(1) = J(\underline{\gamma}) - J(\underline{\gamma}^*)$ is bounded (by hypothesis). It then follows from the monotone convergence theorem that $\lim_{\alpha \downarrow 0} h(\alpha)$ exists, and the limit and expectation operations can be interchanged. As a consequence of continuous differentiability, this then leads to the inequality

$$\sum_{i=1}^{N} E\{\nabla_{u^i} c(\omega_0; \underline{\gamma}^*(\mathbf{y}))[\gamma^i(y^i) - \gamma^{i^*}(y^i)]\} \leq J(\underline{\gamma}) - J(\underline{\gamma}^*)$$

from which team-optimality of $\underline{\gamma}^*$ follows, since the left-hand-side is nonnegative, by (10.12).

If c were strictly convex in $\mathbf{u}$, a.s., then all the inequalities above would be strict, for $\underline{\gamma} \neq \underline{\gamma}^*$, thus leading to

$$J(\underline{\gamma}^*) < J(\underline{\gamma})$$

which implies that $\underline{\gamma}^*$ is the unique team-optimal solution. ◇

Note that the conditions of *Theorem 10.3.1* above do not include the stationarity of $\underline{\gamma}^*$, and furthermore inequality (10.12) may not generally be easy to check, since they involve all permissible policies $\underline{\gamma}$ (with finite cost).

If the following N inequalities hold:

$$E\{\nabla_{u^i} c(\omega; \underline{\gamma}^*(\mathbf{y}))[\gamma^i(y^i) - \gamma^{i*}(y^i)]\} \geq 0, \quad i \in \mathcal{N}, \tag{10.13}$$

then (10.12) would also hold.

Then, either one of the following two conditions will achieve this objective [196] [349]:

(c.1) For all $\underline{\gamma} \in \Gamma$ such that $J(\underline{\gamma}) < \infty$, the following random variables are integrable

$$\nabla_{u^i} c(\omega_0; \underline{\gamma}^*(\mathbf{y}))[\gamma^i(y^i) - \gamma^{i*}(y^i)], \quad i \in \mathcal{N}$$

(c.2) Γ^i is a Hilbert space for each $i \in \mathcal{N}$, and $J(\underline{\gamma}) < \infty$ for all $\underline{\gamma} \in \Gamma$. Furthermore,

$$E_{\omega|y^i}\{\nabla_{u^i} c(\omega_0; \underline{\gamma}^*(\mathbf{y})\} \in \Gamma^i, \quad i \in \mathcal{N}.$$

Here, (c.2) can directly be obtained from (c.1) if Γ^i, $i \in \mathcal{N}$, are taken as Hilbert spaces. Here we give it as a separate condition because in some problems (such as linear quadratic—as we shall see shortly) (c.2) follows quite readily from the problem formulations (due to the condition that a finite expected cost is attained under the considered policies).

Theorem 10.3.2 *[196] [349] Let $\{J; \Gamma^i, i \in \mathcal{N}\}$ be a static stochastic team problem which satisfies all the hypotheses of Theorem 10.3.1, with the exception of the inequality (10.12). Instead of (10.12), let either (c.1) or (c.2) be satisfied. Then, if $\underline{\gamma}^* \in \mathbf{\Gamma}$ is a stationary policy it is also team optimal. Such a policy is unique if $c(\omega_0; \mathbf{u})$ is strictly convex in $\mathbf{u}$, a.s.*

What needs to be shown is that under stationarity, (c.1) or (c.2) implies Theorem 10.3.1; this follows once again from the law of the iterated expectations (Theorem 4.1.3); see [349]. If $(c.1)$ holds, then for all $i \in \mathcal{N}$,

$$\begin{aligned}
&E\left[\nabla_{u^i} c(\omega_0; \underline{\gamma}^*(\mathbf{y}))[\gamma^i(y^i) - \gamma^{i*}(y^i)]\right] \\
&= E\left[E\left[\nabla_{u^i} c(\omega_0; \underline{\gamma}^*(\mathbf{y}))[\gamma^i(y^i) - \gamma^{i*}(y^i)]\middle| y^i\right]\right] \\
&= E\left[E\left[\nabla_{u^i} c(\omega_0; \underline{\gamma}^*(\mathbf{y}))\middle| y^i\right](\gamma^i(y^i) - \gamma^{i*}(y^i))\right] \\
&= 0
\end{aligned} \tag{10.14}$$

under stationarity (where, again the order of expectation and differentiation is justified by the monotone convergence theorem) and thus Theorem 10.3.2 holds.

To appreciate some of the fine points of *Theorems 10.3.1* and *10.3.2*, let us now consider the following example, which was discussed by Radner (1962) [260], and Krainak et al. (1982) [196].

Example 10.2. Let $N = 2$, $\mathbb{U}^1 = \mathbb{U}^2 = \mathbb{R}$, $\xi = x$ be a Gaussian random variable with zero mean and unit variance ($\sim N(0, 1)$), and the loss functional be given by

$$c(x; u^1, u^2) = (u^1 - u^2)^2 e^{x^2} + 2u^1 u^2.$$

Note that c is strictly convex and continuously differentiable in (u^1, u^2) for every value of x. Hence, if the true value of x were known to both agents, the problem would admit a unique team optimal solution: $u^1 = u^2 = 0$, which is also stationary. Since this team-optimal solution does not use the precise value of x, it is certainly optimal also under "no-measurement" information at the decision makers. Note, however, that in this case the only pairs that make $J(\gamma)$ finite, are $u^1 = u^2 = u \in \mathbb{R}$, since

$$E[e^{x^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{+\frac{x^2}{2}} dx = \infty.$$

The set of permissible policies not being an open set, we cannot talk about stationarity in this case. *Theorem 10.3.1* (which does not involve stationarity) is applicable here. Note also that for every $u \in \mathbb{R}$, $u^1 = u^2 = u$ is a *pbp optimal* solution, but only one of these is team optimal.

Now, perhaps as a more practical case, consider the measurement scheme:

$$y^1 = x + w^1; \qquad y^2 = x + w^2$$

where w^1 and w^2 are independent random variables uniformly distributed on the interval $[-1, 1]$, which are also independent of x. Note that here the random state of nature, ξ, is chosen as $(x, w^1, w^2)'$. Clearly, $u^1 = u^2 = 0$ is team-optimal for this case also, but it is not obvious at the outset whether it is stationary or not. Toward this end, let us evaluate (10.11) for $i = 1$ and with $\gamma^2(y^2) = 0$:

$$(\partial/\partial u^1) E_{x,y^2|y^1}\{(u^1)^2 e^{\xi^2}\} = (\partial/\partial u^1)[(u^1)^2 E_{x|y^1}\{e^{\xi^2}\}] = 2u^1 E_{x|y^1}\{e^{\xi^2}\}$$

where the last step follows because the conditional probability density of x given y^1 is nonzero only in a finite interval (thus making the conditional expectation finite). By symmetry, it follows that both derivatives in (10.11) vanish at $u^1 = u^2 = 0$, and hence the team-optimal solution is stationary. It is not difficult to see that in fact this is the only pair of stationary policies. Note that all the hypotheses of *Theorem 10.3.2* are satisfied here, under condition (c.2). ◇

Quadratic-Gaussian teams

Given a probability space $(\Omega, \mathbf{F}, P_\Omega)$, and an associated vector-valued random variable ξ, let $\{J; \Gamma^i, i \in \mathcal{N}\}$ be a static stochastic team problem with the following specifications [349]:

(i) $\mathbb{U}^i = \mathbb{R}^{m_i}$, $i \in \mathcal{N}$; i.e., the action spaces are unconstrained Euclidean spaces.

(ii) The loss function is a quadratic function of $\mathbf{u}$ for every ξ (where we use the notation L instead of c):

$$L(\xi; \mathbf{u}) = \sum_{i,j \in \mathcal{N}} u^{i'} R_{ij}(\xi) u^j + 2 \sum_{i \in \mathcal{N}} u^{i'} r_i(\xi) + c(\xi) \tag{10.15}$$

where $R_{ij}(\xi)$ is a matrix-valued random variable (with $R_{ij} = R'_{ji}$), $r_i(\xi)$ is a vector-valued random variable, and $c(\xi)$ is a random variable, all generated by measurable mappings on the random state of nature, ξ.

(iii) $L(\xi; \mathbf{u})$ is strictly (and uniformly) convex in $\mathbf{u}$ a.s., i.e., there exists a positive scalar α such that, with $R(\xi)$ defined as a matrix comprised of N blocks, with the *ij*'th block given by $R_{ij}(\xi)$, the matrix $R(\xi) - \alpha I$ is positive definite a.s., where I is the appropriate dimensional identity matrix.

(iv) $R(\xi)$ is uniformly bounded above, i.e., there exists a positive scalar β such that the matrix $\beta I - R(\xi)$ is positive definite a.s.

(v) $Y^i = \mathbb{R}^{r_i}$, $i \in \mathcal{N}$, i.e., the measurement spaces are Euclidean spaces.

(vi) $y^i = \eta^i(\xi), i \in \mathcal{N}$, for some appropriate Borel measurable functions $\eta^i, i \in \mathcal{N}$.

(vii) Γ^i is the (Hilbert) space of all Borel measurable mappings of $\gamma^i : \mathbb{R}^{r_i} \to \mathbb{R}^{m_i}$, which have bounded second moments, i.e., $E_{y^i}\{\gamma^{i'}(y^i)\gamma^i(y^i)\} < \infty$.

(viii) $E_\xi[r'_i(\xi) r_i(\xi)] < \infty, i \in \mathcal{N}$; $E_\xi[c(\xi)] < \infty$.

Definition 10.3.2 *A static stochastic team is quadratic if it satisfies (i)–(viii) above. It is a standard quadratic team if furthermore the matrix R is constant for all ξ (i.e., it is deterministic). If, in addition, ξ is a Gaussian distributed random vector, and $r_i(\xi) = Q_i \xi, \eta^i(\xi) = H^i \xi$, $i \in \mathcal{N}$, for some deterministic matrices Q_i, H^i, $i \in \mathcal{N}$, the decision problem is a quadratic-Gaussian team (more widely known as a linear-quadratic-Gaussian (LQG) team under some further structure on Q_i and H^i).* ◇

One class of quadratic teams for which the team-optimal solution can be obtained in closed form are those where the random state of nature ξ is a Gaussian random vector. Let us decompose ξ into $N+1$ block vectors

$$\xi = (x', y^{1'}, y^{2'}, \ldots, y^{N'})' \tag{10.16}$$

of dimensions $r_0, r_1, r_2, \ldots, r_N$, respectively. Being a Gaussian random vector, ξ is completely described in terms of its mean value and covariance matrix, which we specify below:

$$E[\xi] =: \bar{\xi} = (\bar{x}', \overline{y}^{1'}, \ldots, \overline{y}^{N'}) \tag{10.17}$$

$$\text{cov}\,(\xi) =: \Sigma, \text{ with } [\Sigma]_{ij} =: \Sigma_{ij}, \quad i,j = 0,1,\ldots,N \tag{10.18}$$

$[\Sigma]_{ij}$ denotes the *ij*'th block of the matrix Σ of dimension $r_i \times r_j$, which stands for the cross-variance between the *i*'th and *j*'th block components of ξ. We further assume (in addition to the natural condition $\Sigma \geq 0$) that $\Sigma_{ii} > 0$ for $i \in \mathcal{N}$, which means that the measurement vectors y^i's have nonsingular distributions. To complete the description of the quadratic-Gaussian team, we finally take the linear terms $r_i(\xi)$ in the loss function (10.15) to be linear in x, which makes x the "payoff relevant" part of the state of nature:

$$r_i(\xi) = D_i x, \quad i \in \mathcal{N} \tag{10.19}$$

where D_i is an $(r_i \times r_0)$ dimensional constant matrix.

In the characterization of the team-optimal solution for the quadratic-Gaussian team we will need the following important result on the conditional distributions of Gaussian random vectors, generalizing our earlier results in *Chapter* 6.

Lemma 10.3.1 *Let z and y be jointly Gaussian distributed random vectors with mean values $\bar{z}$, $\bar{y}$, and covariance*

$$cov\,(z,y) = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zy} \\ \Sigma_{zy}' & \Sigma_{yy} \end{pmatrix} \geq 0, \quad \Sigma_{yy} > 0. \tag{10.20}$$

Then, the conditional distribution of z given y is Gaussian, with mean

$$E[z|y] = \bar{z} + \Sigma_{zy}\Sigma_{yy}^{-1}(y - \bar{y}) \tag{10.21}$$

and covariance

$$cov(z|y) = \Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{zy}' \tag{10.22}$$

$\diamond$

The complete solution to the quadratic-Gaussian team is given in the following.

Theorem 10.3.3 *[349] The quadratic-Gaussian team decision problem as formulated above admits a unique team-optimal solution, that is affine in the measurement of each agent:*

$$\gamma^{i*}(y^i) = \Pi^i(y^i - \overline{y}^i) + M^i\bar{x}, \quad i \in \mathcal{N}. \tag{10.23}$$

Here, Π^i is an $(m_i \times r_i)$ matrix $(i \in \mathcal{N})$, uniquely solving the set of linear matrix equations:

$$R_{ii}\Pi^i\Sigma_{ii} + \sum_{j\in\mathcal{N}, j\neq i} R_{ij}\Pi^j\Sigma_{ji} + D_i\Sigma_{0i} = 0, \tag{10.24}$$

and M^i is an $(m_i \times r_0)$ matrix for each $i \in \mathcal{N}$, obtained as the unique solution of

$$\sum_{j\in\mathcal{N}} R_{ij}M^j + D_i = 0, \quad i \in \mathcal{N}. \tag{10.25}$$

Remark 10.3. The proof of this result follows immediately from Theorem 10.3.1 and noting that Condition (c.2) holds. However, a *Projection Theorem* based concise proof can also be provided exploiting the quadratic nature of the problem

(see [349, p. 55], [261] and [139]), by defining the problem as an inner-product minimization and projection (onto the closed subspace of decentralized control policies viewed as a product of individual policies of each DM) problem the solution of which builds on an orthogonality condition.

An important application of the above result is the following static Linear Quadratic Gaussian Problem: Consider a two-controller system evolving in $\mathbb{R}^n$ with the following description: Let x_1 be Gaussian and $x_2 = Ax_1 + B^1u_1^1 + B^2u_1^2 + w_1$

$$y_1^1 = C^1x_1 + v_1^1,$$

$$y_1^2 = C^2x_1 + v_1^2,$$

with w, v^1, v^2 zero-mean, i.i.d. disturbances. For $\rho_1, \rho_2 > 0$, let the goal be the minimization of

$$J(\gamma^1, \gamma^2) = \mathcal{E}\left[||x_1||_2^2 + \rho_1||u_1^1||_2^2 + \rho_2||u_1^2||_2^2 + ||x_2||_2^2\right] \tag{10.26}$$

over the control policies of the form:

$$u_t^i = \mu_t^i(y_1^i), \quad i = 1, 2$$

For such a setting, optimal policies are linear.

10.4 Static Reduction of Dynamic Teams: Policy-Dependent and Policy-Indepent Reductions

Following Witsenhausen [331], we say that two information structures are equivalent if: (i) The policy spaces are equivalent/isomorphic in the sense that policies under one information structure are realizable under the other information structure, (ii) the costs achieved under equivalent policies are identical almost surely, and (iii) if there are constraints in the admissible policies, the isomorphism among the policy spaces preserves the constraint conditions.

A large class of sequential team problems admit an equivalent information structure which is static. This is called the *static reduction* of an information structure.

10.4.1 Static reduction I: Dynamic teams with quasi-classical information structures and their policy-dependent static reduction

An important information structure which is not nonclassical, is of the *quasi-classical* type, also known as *partially nested*; an *IS* is partially nested if an agent's information at a particular stage t can depend on the action of some other agent at some stage $t' \leq t$ only if she also has access to the information of that agent at stage t'. For such team problems with partially nested information, one talks about *precedence relationships* among agents: an agent DM i is *precedent* to another agent DM j (or DM i *communicates* to DM j), if the former agent's actions affect the information of the latter, in which case (to be partially nested) DM j has to have the information based on which the action-generating policy of DM i was constructed.

For partially nested (or quasi-classical) information structures, static reduction has been studied by Ho and Chu in the specific context of LQG systems [171] and for a class of non-linear systems satisfying restrictive invertibility properties [172].

Under quasi-classical information, LQG stochastic team problems are tractable by conversion into equivalent static team problems: Consider the following dynamic team with N agents, where each agent acts only once, with $\mathbf{A}k$, $k \in \mathcal{N}$, having the following measurement

$$y^k = C^k\xi + \sum_{i:i\to k} D_{ik}u^i, \tag{10.27}$$

where ξ is an exogenous random variable picked by nature, and $i \to k$ denotes the precedence relation that the action of $\mathbf{A}i$ affects the information of $\mathbf{A}k$ and u^i is the action of $\mathbf{A}i$.

If the information structure is quasi-classical, then

$$\mathcal{I}^k = \{y^k, \{\mathcal{I}^i, i \to k\}\}.$$

That is, $\mathbf{A}k$ has access to the information available to all the signaling agents. Such an IS is equivalent to the IS $\mathcal{I}^k = \{\tilde{y}^k\}$, where $\tilde{y}^k$ is a static measurement given by

$$\tilde{y}^k = \left\{ C^k\xi, \{C^i\xi, i \to k\} \right\}. \tag{10.28}$$

Such a conversion can be done provided that the policies adopted by the agents are deterministic, with the equivalence to be interpreted in the sense that any deterministic policy measurable under the original *IS* being measurable also under the new (static) *IS* and vice versa, since the actions are determined by the measurements. The restriction of using only deterministic policies is, however, without any loss of optimality: with policies of all other agents fixed (possibly randomized) no agent can benefit from randomized decisions in such team problems. We discussed this property of irrelevance of random information/actions in optimal stochastic control in *Chapter 5* in view of Blackwell's Irrelevant Information Theorem (see [346, Remark 2]).

This observation, made by Ho and Chu [171] leads to the following result.

Theorem 10.4.1 *Consider an LQG system with a partially nested information structure. For such a system, optimal solutions are affine (that is, linear plus a constant).*

The linearity condition can be relaxed via the following more general condition, which is also due to Ho and Chu [172].

Assumption 10.4.1 *Under a quasi-classical information structure, with*

$$\mathcal{I}^k = \{y^k, \{\mathcal{I}^i, i \to k\}\}.$$

if $y^k = g(\xi, u^{[1,k-1]})$, then the map $g(\cdot, u^{[1,k-1]}) : \xi \mapsto y^k$ is invertible.

Under this assumption, static reduction is possible.

Policy-dependence of the static reduction. In the above, under Assumption 10.4.1, while mapping the policies that are equivalent under the dynamic setup to those that are expressed in terms of exogenous variables in the static-reduced form, we note that the policies' dependence on the exogenous variables explicitly depend on the policies adopted by the preceding DMs. Accordingly, we refer to the static reduction of partially nested dynamic teams as *policy-dependent static reduction* (as opposed to the *policy-independent* reduction to be presented in the following). Some restrictions and limitations due to such policy-dependence will be studied later in the chapter.

If the control actions are also shared under the static measurement reductions, called *static-measurements with control-sharing reduction* [277], even though this reduced information structure is not static in a strict sense, where only the measurements are so, this reduction is policy-independent.This follows since $g(\cdot, u^{[1,k-1]})$ in Assumption 10.4.1 is invertible given previous actions $u^{[1,k-1]}$ regardless of the policies of the previous decision makers.

Remark 10.4. Another class of dynamic team problems that can be converted into solvable dynamic optimization problems are those where even though the information structure is nonclassical, there is no incentive for signaling because any signaling from say agent $\mathbf{A}i$ to agent $\mathbf{A}j$ conveys information to the latter which is "cost irrelevant", that is it does not lead to any improvement in performance [342] [349].

10.4.2 Static reduction II: Non-classical information structures and their policy-independent reduction

In this sub-section, we introduce another static reduction method, due to Witsenhausen [331], applicable also to non-classical information structures and call it a *policy-independent static reduction*, since this reduction does not depend on the policies adopted.

For some of the results of the chapter, we need to go beyond a static reduction, and we will need to make the measurements independent of each other as well as ω_0. This is not possible for every team which admits a static reduction, for example quasi-classical team problems with LQG models [171] do not admit such a further reduction, since the measurements are partially nested. Witsenhausen refers to such an information structure as *independent static* in [331, Section 4.2(e)].

Consider now dynamic team setting according to the intrinsic model where each DM t measures

$$y^t = g_t(\omega_0, \omega_t, y^1, \dots, y^{t-1}, u^1, \dots, u^{t-1}),$$

and the decisions are generated by $u^t = \gamma^t(y^t)$, with $1 \leq t \leq N$. Here $\omega_0, \omega_1, \cdots, \omega_N$ are primitive (exogenous) variables. We will indeed, for every $1 \leq n \leq N$, view the relation

$$P(dy^n | \omega_0, y^1, y^2, \cdots, y^{n-1}, u^1, u^2, \cdots, u^{n-1}),$$

as a (controlled) stochastic kernel (to be defined later), and through standard stochastic realization results (see [144, Lemma 1.2] or [56, Lemma 3.1]), we can represent this kernel in a functional form through

$$y^n = g_n(\omega_0, \omega_n, y^1, y^2, \cdots, y^{n-1}, u^1, u^2, \cdots, u^{n-1})$$

for some independent ω_n and measurable g_n.

This team admits an *independent-measurements* reduction provided that the following absolute continuity condition holds: For every $t \in \mathcal{N}$, there exists a function f_t such that for all Borel S:

$$\begin{aligned} &P(y^t \in S | \omega_0, u^1, u^2, \cdots, u^{t-1}, y^1, y^2, \cdots, y^{t-1}) \\ &\quad = \int_S f_t(y^t, \omega_0, u^1, u^2, \cdots, u^{t-1}, y^1, y^2, \cdots, y^{t-1}) Q_t(dy^t), \end{aligned} \tag{10.29}$$

We can then write (since the action of each DM is determined by the measurement variables under a policy)

$$\begin{aligned} &P(d\omega_0, d\mathbf{y}, d\mathbf{u}) \\ &= P(d\omega_0) \prod_{t=1}^{N} \left(f_t(y^t, \omega_0, u^1, u^2, \cdots, u^{t-1}, y^1, y^2, \cdots, y^{t-1}) Q_t(dy^t) 1_{\{\gamma^t(y^t) \in du\}} \right). \end{aligned}$$

The cost function $J(\underline{\gamma})$ can then be written as

$$J(\underline{\gamma}) = \int P(d\omega_0) \prod_{t=1}^{N} (f_t(y^t, \omega_0, u^1, u^2, \cdots, u^{t-1}, y^1, y^2, \cdots, y^{t-1}) Q_t(dy^t)) c(\omega_0, \mathbf{u}), \tag{10.30}$$

with $u^k = \gamma^k(y^k)$ for $1 \leq k \leq N$, and where now the measurement variables can be regarded as independent from each other, and also from ω_0, and by incorporating the $\{f_t\}$ terms into c, we can obtain an equivalent *static team* problem. Hence, the essential step is to appropriately adjust the probability space and the cost function.

The new cost function may now explicitly depend on the measurement values, such that

$$c_s(\omega_0, \mathbf{y}, \mathbf{u}) = c(\omega_0, \mathbf{u}) \prod_{t=1}^{N} f_t(y^t, \omega_0, u^1, u^2, \cdots, u^{t-1}, y^1, y^2, \cdots, y^{t-1}). \tag{10.31}$$

Here we can reformulate even a static team to one which is, clearly still static, but now with independent measurements which are also independent from the cost relevant exogenous variable ω_0.

Such a condition is in general not restrictive. Indeed, as Witsenhausen notes, a static reduction always holds when the measurement variables take values from countable set since a reference measure as in Q^i above can be always constructed on the measurement space $\mathbb{Y}^i$ (e.g., $Q^i(z) = \sum_{j \geq 1} 2^{-j} 1_{\{z = m_j\}}$ where $\mathbb{Y}^i = \{m_j, j \in \mathbb{N}\}$) so that the absolute continuity condition always holds. We refer the reader to [80] for relations with classical continuous-time stochastic control

where the relation with Girsanov's classical measure transformation [145] [33] is recognized. For discrete-time partially observed stochastic control, similar arguments had been presented in Borkar [58], [62] again in the context of measure transformation.

Remark 10.5. [Change of Measure Formula] Denote the joint probability measure on $(\omega_0, u^1, \ldots, u^N, y^1, \ldots, y^N)$ by P, and the probability measure of ω_0 by $\mathbb{P}^0$. If the preceding absolute continuity condition (10.29) holds, then (under any admissible policy profile $\gamma^1, \cdots, \gamma^N$) there exists a joint reference probability measure $\mathbb{Q}$ on $(\omega_0, u^1, \ldots, u^N, y^1, \ldots, y^N)$ such that the probability measure P is absolutely continuous with respect to $\mathbb{Q}$ ($P \ll \mathbb{Q}$), so that for every Borel set A in $(\Omega_0 \times \prod_{i=1}^N (\mathbb{U}^i \times \mathbb{Y}^i))$

$$P(A) = \int_A \frac{dP}{d\mathbb{Q}} \mathbb{Q}(d\omega_0, du^1, \ldots, du^N, dy^1, \ldots, dy^N), \tag{10.32}$$

where the reference probability measure

$$\mathbb{Q}(d\omega_0, du^1, \ldots, du^N, dy^1, \ldots, dy^N) := \mathbb{P}^0(d\omega_0) \prod_{i=1}^N Q^i(dy^i) 1_{\{\gamma^i(y^i) \in du^i\}}, \tag{10.33}$$

leads to a Radon-Nikodym derivative, which is *policy-independent*:

$$\frac{dP}{d\mathbb{Q}}(\omega_0, u^1, \ldots, u^1, y^1 \ldots, y^N) = \prod_{i=1}^N f^i(y^i, \omega_0, u^1, \ldots, u^{i-1}, y^1, \ldots, y^{i-1}). \tag{10.34}$$

Indeed, one may slightly relax the condition in (10.29) (which requires the absolute continuity to hold for all $\omega_0, u^1, \cdots, u^{t-1}, y^1, \cdots, y^{t-1}$), to an almost sure existence condition of a derivative under a reference measure, in the sense that (10.34) holds.

Witsenhausen's Counterexample and its static reduction

The celebrated Witsenhausen's counterexample [325] is a dynamic non-classical team problem

Suppose x and w_1 are two independent, zero-mean Gaussian random variables with variance σ^2 and 1 so that

$$y^0 = x, \qquad y^1 = u_0 + w_1$$

$$u_0 = \gamma_0(x), \qquad u_1 = \gamma_1(y).$$

with the performance criterion:

$$Q_W(x, u_0, u_1) = k(u_0 - x)^2 + (u_1 - u_0)^2\,, \tag{10.35}$$

This can also be viewed as a standard discrete-time two-stage stochastic optimal control problem, with state equations (see Figure 10.3)

$$x_1 = x_0 + v_0\,, \qquad x_2 = x_1 - v_1\,, \tag{10.36}$$

measurement equations

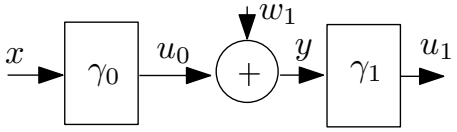

Fig. 10.2: Flow of information in Witsenhausen's counterexample.

$$y_0 = x_0\,, \qquad y_1 = x_1 + w_1\,, \tag{10.37}$$

and memoryless controls

$$v_0 = \gamma_0(y_0)\,, \qquad v_1 = \gamma_1(y_1)\,, \tag{10.38}$$

where μ_0 and μ_1 are the instantaneous measurement output control policies at stages 0 and 1, respectively. This becomes equivalent to the earlier formulation in view of the correspondences

$$u_0 = x_0 + v_0\,, \quad u_1 = v_1\,, \quad x = x_0\,, \quad w = w_1\,, \quad y = y_1\,,$$

if we pick the cost function as

$$\tilde{Q}(x_2, v_0) = (x_2)^2 + k(v_0)^2 \equiv Q_W(x_1 - v_0, x_1, x_1 - x_2)\,.$$

This problem is described by a linear system; all primitive variables are Gaussian and the performance criterion is quadratic, yet linear policies are not optimal. We note that this is a non-convex problem [351] and thus variational methods do not necessarily lead to optimality. In fact, we don't even have a good lower bound on the optimal cost for Witsenhausen's counterexample even though approximation results exist (see [274] for a detailed discussion).

The static reduction for Witsenhausen's counterexample proceeds as follows:

$$\begin{aligned}
&\int (k(u^1 - y^1)^2 + (u^1 - u^2)^2) Q(dy^1)\gamma^1(du^1|y^1)\gamma_1(du^2|y^2)P(dy^2|u^1) \\
&= \int (k(u^1 - y^1)^2 + (u^1 - u^2)^2) Q(dy^1)\gamma^1(du^1|y^1)\gamma_1(du^2|y^2)\eta(y^2 - u^1)dy^2 \\
&= \int \left((ku_0^2 + (u_0 - u_1)^2)\gamma^1(du^1|y^1)\gamma_1(du^2|y^2)\frac{\eta(y^2 - u^1)dy^2}{\eta(y^2)} \right) Q(dy^1)\eta(y^2)dy^2 \\
&= \int \left((ku_0^2 + (u_0 - u_1)^2)\gamma^1(du^1|y^1)\gamma_1(du^2|y^2)\frac{\eta(y^2 - u^1)dy^2}{\eta(y^2)} \right) Q(dy^1)Q(dy^2)
\end{aligned} \tag{10.39}$$

where Q denotes a Gaussian measure with zero mean and unit variance and η its density.

10.4.3 Equivalent static reductions preserve optimality but may not person-by-person-optimality or stationarity

We showed earlier that static reduction can be a very useful method for deriving and studying properties of optimal policies in stochastic teams. We note, however, that static reduction has its limitations when the solution concept is not global optimality but only person-by-person-optimality or stationarity. This, in particular, is a concern for policy-dependent reductions.

Theorem 10.4.2 *[277] Consider a stochastic dynamic team with a policy-independent static reduction.*

(i) A policy $\underline{\gamma}^*$ is pbp optimal (globally optimal) for a dynamic team if and only if $\underline{\gamma}^*$ is pbp optimal (globally optimal) for a policy-independent static reduction of the dynamic team;

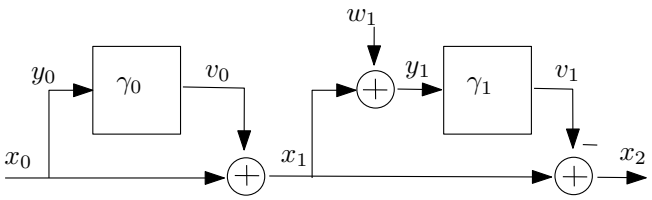

Fig. 10.3: Witsenhausen's counterexample in two-stage state-space linear stochastic control form.

(ii) Let a policy $\underline{\gamma}^*$ satisfy P-almost surely

$$\nabla_{u^i} E_{\mathbb{Q}}^{\gamma^{-i*}} \left[\frac{dP}{d\mathbb{Q}} \middle| y^i \right] \Big|_{u^i = \gamma^{i*}(y^i)} = 0 \qquad \forall i \in \mathcal{N}, \tag{10.40}$$

where $\frac{dP}{d\mathbb{Q}}$ is defined in (10.34)*. Then, $\underline{\gamma}^*$ is stationary for dynamic team if and only if $\underline{\gamma}^*$ is stationary for a policy-independent static reduction of dynamic team.*

Theorem 10.4.3 *[277] Consider a stochastic dynamic team with partially nested information structure. Let Assumption 10.4.1 hold. Then:*

(i) $\underline{\gamma}^{D,*}$ is a globally optimal policy for for the dynamic team if and only if its static equivalent $\underline{\gamma}^{S,*}$ is a globally optimal policy for its static reduction, under the policy-dependent static reduction.

(ii) If $\underline{\gamma}^{D,*}$ is a stationary (pbp optimal) policy for ($\mathcal{P}^D$), then its static equivalent $\underline{\gamma}^{S,*}$ is not necessarily a stationary (pbp optimal) policy for ($\mathcal{P}^S$) under the policy-dependent static reduction;

(iii)If $\underline{\gamma}^{S,*}$ is a stationary (pbp optimal) policy for a static reduced dynamic team, then $\underline{\gamma}^{D,*}$, satisfying the policy-dependent static reduction relation, is not necessarily pbp optimal for the dynamic team problem.

10.4.4 All stochastic dynamic teams are nearly static (with independent measurements) reducible

Now that we have seen the benefits of static reduction, a natural question arises as to whether we can perturb any stochastic dynamic team by adding some additive noise to the measurements to make it static-reducible with arbitrarily small error in the optimal cost: That is, are all dynamic team problems ϵ-away from being static reducible as far as optimal cost is concerned for any $\epsilon > 0$? This is indeed the case, see [173].

10.5 Expansion of information Structures: A recipe for identifying sufficient information

We start with a general result on *optimum-performance equivalence* of two stochastic dynamic teams with different information structures. This is in fact a result which has a very simple proof, but it is quite effective as we will see shortly.

Proposition 10.5.1 *Let D_1 and D_2 be two stochastic dynamic teams with the same loss function, and differing only in their information structures, $\underline{\eta}_1$ and $\underline{\eta}_2$, respectively, with corresponding composite strategy spaces $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$, such that $\mathbf{\Gamma}_2 \subseteq \mathbf{\Gamma}_1$. Let D_1 admit a team-optimal solution, denoted by $\underline{\gamma}_1^* \in \mathbf{\Gamma}_1$, with the further property that $\underline{\gamma}_1^* \in \mathbf{\Gamma}_2$. Then $\underline{\gamma}_1^*$ also solves D_2.*

A recipe for utilizing the result above would be [349]:

> Given a team problem, say D_2, with *IS* $\underline{\eta}_2$, which is presumably difficult to solve, obtain a *finer IS* $\underline{\eta}_1$, and solve the team problem under this expanded *IS* (assuming that this new team problem is easier to solve). Then, if the team-optimal solution here is adapted to the sigma-field generated by the original coarser *IS*, it solves also the original problem D_2.

10.6 Convexity of Decentralized Stochastic Control Problems

We have already seen the utility of convexity in team theoretic problems earlier in the chapter, e.g. in Theorem 10.3.1. We will study convexity further in this section.

10.6.1 Convexity of static team problems and an equivalent representation of cost functions

Definition 10.6.1 *A (static or dynamic) team problem is convex on* $\mathbf{\Gamma}$ *if* $J(\underline{\gamma}) < \infty$ *for all* $\underline{\gamma} \in \mathbf{\Gamma}$ *and for any* $\alpha \in (0,1), \underline{\gamma}_1, \underline{\gamma}_2 \in \mathbf{\Gamma}$*:*

$$J(\alpha \underline{\gamma}_1 + (1-\alpha)\underline{\gamma}_2) \leq \alpha J(\underline{\gamma}_1) + (1-\alpha) J(\underline{\gamma}_2)$$

We state the following immediate result without proof, more general refinements will be stated later in the chapter.

Theorem 10.6.1 *Consider a static team.* $J(\underline{\gamma})$ *is convex if* $c(\omega_0, \mathbf{u})$ *is convex in* $\mathbf{u}$ *for all* ω_0*, provided that* $J(\underline{\gamma}) < \infty$ *for all* $\underline{\gamma} \in \mathbf{\Gamma}$.

The condition in Theorem 10.6.1 is not tight, however, due to information structure and measurability aspects.

Example 10.6. Consider $\Omega = [0,1]$ and let P be the uniform distribution on Ω, with $N = 2$, $\mathbb{U}^1 = \mathbb{U}^2 = [1,2]$. Let:

$$c(\omega, u^1, u^2) = 1_{\{\omega \in [0, 0.9]\}} \left((u^1 - 2)^2 + (u^2 - 2)^2 \right) + 1_{\{\omega \in (0.9, 1]\}} \left(\sqrt{1 + u^1} + \sqrt{1 + u^2} \right)$$

Now, suppose further that $I^1 = I^2 = \eta^1(\omega) = \eta^2(\omega) = 1_{\{\omega \in [0, 0.1)\}}$. It follows that here the team problem is convex, even though $c(\omega, u^1, u^2)$ is not convex on $\{\omega : \omega \in (0.9, 1]\}$, which has a non-zero probability measure. To see this, note that one may view this optimization problem as $J(u_1^1, u_2^1; u_1^2, u_2^2)$ where $u_j^i = \gamma^i(\omega_j)$, with $\omega_1 \equiv \{\omega : \omega \in [0, 0.1)\}$ and $\omega_2 \equiv \{\omega : \omega \in [0.1, 1]\}$. It follows that

$$J(u_1^1, u_2^1; u_1^2, u_2^2) = \sum_{i=1,2} 0.1(u_1^i - 2)^2 + 0.8(u_2^i - 2)^2 + 0.1(\sqrt{u_2^i + 1})$$

The Hessian of J is a diagonal matrix with strictly positive entries, leading to the convexity of the problem.

In the following, we will make use of the fact that $u^k \leftrightarrow y^k \leftrightarrow \{\mathbf{y}^{-k}, \omega\}$ form a Markov chain almost surely. Before proceeding further, let us note that the *join* of two σ-fields over some set $\mathbb{X}$ is the coarsest σ-field containing both. The *meet* of two σ-fields is the finest σ-field which is a subset of both. Let $\mathcal{F}^i$ be the σ-field generated by η^i over Ω, and let $\mathcal{F}_c = \bigcap_k \mathcal{F}^k$ be the meet of these fields, this is termed as *common knowledge* by Aumann [22] for finite probabilities spaces. In addition, let $\mathcal{F}_j$ be the *join* of the σ-field, denoted with $\mathcal{F}_j = \bigcup_k \mathcal{F}^k$.

In the following, as earlier in the chapter, we assume that the measurement and the control action spaces are standard Borel.

An equivalent representation of the cost through iterated expectations. Let us express the expected cost under a given measurable team policy $\underline{\gamma}$ as follows. With the interpretation that $P(u^k \in \cdot | y^k) = 1_{\{u^k = \gamma^k(y^k) \in \cdot\}}$, we obtain from the law of the iterated expectations that

$$E[c(\omega_0, \mathbf{u})] = E\Big[E[c(\omega_0, \mathbf{u}) | \mathbf{y}] \Big] \tag{10.41}$$

Under any measurable policy, given $\mathbf{y}$, $\mathbf{u}$ is specified. Thus, with

$$\tilde{c}(y^1, \cdots, y^N, u^1, \cdots, u^N) := E[c(\omega_0, \mathbf{u}) | \mathbf{y}],$$

that is $\tilde{c}(y^1, \cdots, y^N, u^1, \cdots, u^N) = \left(\int P(d\omega_0 | \mathbf{y}) c(\omega_0, u^1, \cdots, u^N) \right)$, the cost function becomes $E[\tilde{c}(y^1, \cdots, y^N, u^1, \cdots, u^N)]$. We will use this representation in the following.

Theorem 10.6.2 *(i)If a team problem is convex, then*

$$E[c(\omega_0, \mathbf{u}) | \mathcal{F}_c]$$

is convex in u almost surely.
(ii) If

$$E[c(\omega_0, \mathbf{u})|\mathcal{F}_j]$$

is convex in u almost surely, then the team problem is convex on the set of team policies that satisfy $J(\underline{\gamma}) < \infty$.

Proof. (i) We will show the contra-positive. Let B be a Borel set such that $P(B) > 0$, $B \in \mathcal{F}_c$, and $E[c(\omega_0, \mathbf{u})|B]$ be non-convex so that there exist $\mathbf{u}$ and $\mathbf{u}'$ and $\lambda \in (0, 1)$ such that

$$E[c(\omega_0, \lambda\mathbf{u} + (1-\lambda)\mathbf{u}')|B] > \lambda E[c(\omega_0, \mathbf{u})|B] + (1-\lambda)E[c(\omega_0, \mathbf{u}')|B]$$

Now, let $\underline{\gamma}$ and γ be two team policies so that these only differ on B; and on B $\underline{\gamma} = \mathbf{u}$ and $\gamma = \mathbf{u}'$. Such measurable policies exist, for example by taking $\underline{\gamma}(\omega) = \{0, 0, \cdots, 0\}$ when $\omega \notin B$. These policies are both Borel measurable and are admissible given the information structure. Then

$$J(\lambda\underline{\gamma} + (1-\lambda)\gamma) > \lambda J(\underline{\gamma}) + (1-\lambda)J(\gamma)$$

and convexity fails.

(ii) We adopt the equivalent representation (10.41) in this part of the proof. Note that under any measurable policy, the random variable $\tilde{c}(y^1, \cdots, y^N, u^1, \cdots, u^N)$ is measurable on the σ-field generated by $\mathbf{y}$ and thus the join σ-field. The proof then follows from the following. Consider two policies $\underline{\gamma}$ and $\underline{\bar{\gamma}}$ with finite expected costs. It follows then that

$$\begin{aligned}
&J(\lambda\underline{\gamma} + (1-\lambda)\underline{\bar{\gamma}}) \\
&= \int P(d\mathbf{y})\tilde{c}(y^1, \cdots, y^N, \lambda\gamma^1(y^1) + (1-\lambda)\bar{\gamma}^1(y^1), \cdots, \lambda\gamma^N(y^N) + (1-\lambda)\bar{\gamma}^N(y^N)) \\
&\leq \int P(d\mathbf{y})\Big(\lambda\tilde{c}(y^1, \cdots, y^N, \gamma^1(y^1), \cdots, \gamma^N(y^N)) \\
&\qquad +(1-\lambda)\tilde{c}(y^1, \cdots, y^N, \bar{\gamma}^1(y^1), \cdots, \bar{\gamma}^N(y^N))\Big) \\
&= \lambda J(\underline{\gamma}) + (1-\lambda)J(\underline{\bar{\gamma}})
\end{aligned}$$

$\diamond$

It can be observed that Example 10.6 satisfies the conditions of Theorem 10.6.2. These conditions will also be used to study Witsenhausen's counterexample [325] later in the chapter.

A generalization of Radner and Krainak et. al.'s theorems

We provide a generalization of Radner's or Krainak et al.'s theorem by utilizing an information structure dependent nature of convexity. For example, Radner or Krainak et.al's theorems are not applicable to Example 10.6.

Theorem 10.6.3 *Let $\{J; \Gamma^i, i \in \mathcal{N}\}$ be a static stochastic team problem, the loss function $E[c(\omega_0, \mathbf{u})|\mathcal{F}_j]$ is convex and continuously differentiable in $\mathbf{u}$ almost surely. Let $\underline{\gamma}^* \in \mathbf{\Gamma}$ be a policy N-tuple with a finite cost $(J(\underline{\gamma}^*) < \infty)$, and suppose that for every $\underline{\gamma} \in \mathbf{\Gamma}$ such that $J(\underline{\gamma}) < \infty$, the following holds:*

$$\sum_{i \in \mathcal{N}} E\{\nabla_{u^i}\tilde{c}(\mathbf{y}, \underline{\gamma}^*(\mathbf{y}))[\gamma^i(y^i) - \gamma^{i*}(y^i)]\} \geq 0, \tag{10.42}$$

where $\tilde{c}(\mathbf{y}, \mathbf{u}) := E[c(\omega_0, \mathbf{u})|\mathcal{F}_j]$. Then, $\underline{\gamma}^*$ is a team-optimal policy, and it is unique if $\tilde{c}(\mathbf{y}, \mathbf{u})$ is strictly convex in $\mathbf{u}$ almost surely.

Proof. The proof follows by defining the new loss function:

$$\tilde{c}(\mathbf{y}, \mathbf{u}) = E[c(\omega_0, \mathbf{u})|\mathcal{F}_j],$$

almost surely. The result then follows as in Theorem 10.3.1. ◇

Theorem 10.6.4 *Let $\{J; \Gamma^i, i \in \mathcal{N}\}$ be a static stochastic team problem which satisfies all the hypotheses of Theorem 10.6.3, with the exception of inequality (10.42). Instead of (10.42), let either (c.5) or (c.6) be satisfied with c replaced with $\tilde{c}$. Then, if $\underline{\gamma}^* \in \mathbf{\Gamma}$ is a stationary policy it is also team optimal. Such a policy is unique if $E[c(\omega_0, \mathbf{u})|\mathcal{F}_j]$ is strictly convex in $\mathbf{u}$, a.s.*

Proof. The proof follows by defining the new loss function $\tilde{c}$ as in the proof of Theorem 10.6.3, and following Theorem 10.3.2. ◇

10.6.2 Convexity of Sequential Dynamic Teams

Convexity of the reduced model

The static reduction of a sequential dynamic team problem, if exists, is not unique. However, the following holds: Either all of the static reductions are convex or none is. This holds under a minor technicality for quasi-classical patterns. Here, first the information is to be expanded to allow for control sharing. Thus, we can state that a stochastic dynamic team problem with a static reduction is convex if and only if its static reduction is.

Non-convexity of the Witsenhausen counterexample and its variants. Consider the celebrated Witsenhausen's counterexample [325]: This is a dynamic non-classical team problem with y^1 and w^1 zero-mean independent Gaussian random variables with unit variance and $u^1 = \gamma^1(y^1)$, $u^2 = \gamma^2(u^1 + w^1)$ and the cost function $c(\omega, u^1, u^2) = k^2(y^1 - u^1)^2 + (u^1 - u^2)^2$ for some $k > 0$: The static reduction is given in (10.39).

Another interesting example is the point-to-point communication problem: Here, the setup is exactly as in the Witsenhausen's counterexample, but $c(\omega, u^1, u^2) = k^2(u^1)^2 + (y^1 - u^2)^2$. This problem is a peculiar one in that, even though the information structure is non-classical, and is non-convex; an optimal encoder and decoder is linear. A proof of this result builds on information theoretic ideas, such as the data-processing inequality (see *Chapters 3, 11* in [349] for a detailed account). In this case, the reduction (10.30) writes as:

$$\begin{aligned}
&\int (k(u^1)^2 + (y^1 - u^2)^2) Q(dy^1) \gamma^1(du^1|y^1) \gamma_1(du^2|y^2) P(dy^2|u^1) \\
&= \int \left((k(u^1)^2 + (y^1 - u^2)^2) Q(dy^1) \gamma^1(du^1|y^1) \gamma_1(du^2|y^2) \frac{\eta(y^2 - u^1) dy^2}{\eta(y^2)} \right) Q(dy^1) Q(dy^2)
\end{aligned} \tag{10.43}$$

Consider the static reduction of Witsenhausen's counterexample and the Gaussian signaling problem (10.39)-(10.43). For both (10.39) and (10.43), using the fact that e^{-x^2} is not a convex function, we recognize that this problem is not convex by Theorem 10.6.2(i) (with the common knowledge/information being the trivial σ-algebra consisting of the empty set and its complement).

We note that Witsenhausen states without proof in [325] (p. 134) that the counterexample is non-convex in γ^1 for every optimal γ^2 (selected as a best response to γ^1). The discussion above can be viewed as an explicit proof for this result. Note also that for both problems above, linear policies contain person by person optimal policies, but this does not imply global optimality. For the first problem, Witsenhausen had shown the suboptimality of linear policies. For the second problem (10.43), however, linear policies are indeed optimal.

Partially nested information structures: Convexity of the reduced model

As reviewed earlier, an important information structure which is not nonclassical, is of the *partially nested* type. For such team problems with partially nested information, a static reduction exists under certain invertibility conditions as discussed earlier. For such problems, the cost function is not altered by the static reduction. This leads to the following result.

Theorem 10.6.5 *Consider a partially nested stochastic dynamic team which admits a static reduction where the cost function* $c(\omega_0, \mathbf{u})$ *convex in* $\mathbf{u}$*. If the information structure is expanded to also include control sharing whenever measurements are shared under the partially nested information structure, then the team problem is convex.*

See [351]. We note that Ho and Chu [171] established this result that for the special setup involving the partially nested LQG teams. In this case, optimal policies are linear through an equivalence to static teams.

10.6.3 Symmetric Team Problems: Optimality of Symmetric Policies

If a team problem, static or dynamic, is convex and symmetric (i.e., exchangeable; meaning that any permutation of DM policies does not alter the induced expected cost), then an optimal team policy can be taken to be symmetric across agents without loss.

In the absence of convexity, one can only show exchangeability of an optimal team policy [278–280].

10.7 The Strategic Measures Approach to Decentralized Stochastic Control

For classical stochastic control problems, strategic measures were defined (see [284], [253], [117] and [124]) as the set of probability measures induced on the product (sequence) spaces of the states, measurements, and actions; that is, given an initial state distribution and a policy, one can uniquely define a probability measure on the product space of the states, measurements, and actions. Certain measurability, compactness, and convexity properties of strategic measures for classical stochastic control problems were studied in [49, 117, 124, 253].

In [351], strategic measures for decentralized stochastic control problems were introduced and many of their properties were established. For decentralized stochastic control problems, considering the set of strategic measures along with compactification and/or convexification of these sets of measures through introducing private and/or common randomness allow one to place operationally flexible topologies (such as those leading to a standard Borel space, e.g., weak convergence topology, among others) on the set of strategic measures, as we will study in the following.

10.7.1 Measurable policies as a subset of randomized policies and strategic measures

A common method in control theory is to view a measurable policy as a special case of *relaxed* policies where relaxation is often employed by randomization. Such an approach has been ubiquitously adopted in various fields often with different terminology (e.g., relaxed controls (Young topology) in optimal deterministic control [224] [340], distributional strategies in economics [238] [230], local hidden variables in quantum information theory etc.)

We recall here the following representation result [56]. Let $\mathbb{X}, \mathbb{M}$ be Borel spaces. Let the notation $\mathcal{P}(\mathbb{X})$ denote the set of probability measures on $\mathbb{X}$. Consider the set of probability measures

$$\Theta := \left\{ \zeta \in \mathcal{P}(\mathbb{X} \times \mathbb{M}) : \zeta(dx, dm) = P(dx)\, Q^f(dm|x), Q^f(\cdot|x) = 1_{\{f(x) \in \cdot\}}, f : \mathbb{X} \to \mathbb{M} \right\},$$

on $\mathbb{X} \times \mathbb{M}$ having fixed input marginal P on $\mathbb{X}$ and the stochastic kernel from $\mathbb{X}$ to $\mathbb{M}$ is realized by some measurable function $f : \mathbb{X} \to \mathbb{M}$. We equip this set with weak convergence topology. This set is the (Borel measurable) set of the extreme points of the set of probability measures on $\mathbb{X} \times \mathbb{M}$ with a fixed marginal P on $\mathbb{X}$. For compact $\mathbb{M}$, the Borel

measurability of Θ follows from [251] since the set of probability measures on $\mathbb{X} \times \mathbb{M}$ with a fixed marginal P on $\mathbb{X}$ is a convex and compact set in a complete separable metric space, and therefore, the set of its extreme points is Borel measurable. Moreover, the non-compact case holds by [56, Lemma 2.3]. Furthermore, given a fixed marginal P on $\mathbb{X}$, any stochastic kernel Q from $\mathbb{X}$ to $\mathbb{M}$ can be identified by a probability measure $\xi \in \mathcal{P}(\Theta)$ such that

$$Q(\cdot|x) = \int_{\Theta} \xi(dQ^f)\, Q^f(\cdot|x). \tag{10.44}$$

In particular, a stochastic kernel can thus be viewed as an integral representation over probability measures induced by deterministic policies.

For a team setup, for any DM k, let

$$\Theta^k := \Big\{ \zeta \in \mathcal{P}(\mathbb{Y}^k \times \mathbb{U}^k) : \zeta = P_k\, Q^{\gamma^k}, \\ Q^{\gamma^k}(\cdot|y^k) = 1_{\{\gamma^k(y^k)\in\cdot\}}, \gamma^k \in \Gamma^k, P_k(\cdot) = P(y^k \in \cdot) \Big\}.$$

For a static team, P_k would be fixed; that is, independent of the policies of the preceding DMs. Therefore, in static case, in view of (E.4), any element $\zeta \in \mathcal{P}(\mathbb{Y}^k \times \mathbb{U}^k)$ with fixed marginal P_k on $\mathbb{Y}^k$ can be expressed as the mixture of Θ^k

$$\zeta(A) = \int_{\Theta^k} \xi^k(dQ)\, Q(A), \quad A \in \mathcal{B}(\mathbb{Y}^k \times \mathbb{U}^k), \tag{10.45}$$

for some $\xi \in \mathcal{P}(\Theta^k)$. In the sequel, we generalize this idea to the set of strategic measures induced by measurable policies and define various relaxed policies that are obtained as a mixture of measurable policies. Indeed, instead of viewing N-tuple of policies as the joint strategy of DMs, we regard the induced probability distribution on the product space of state, measurements, and actions as the joint strategy and name it *strategic measure*.

10.7.2 Sets of strategic measures for static teams

Consider a static team problem defined under Witsenhausen's intrinsic model. In the following, $B = B^0 \times \prod_{k=1}^N (A^k \times B^k)$ are used to denote the cylindrical Borel sets in $\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)$.

Let $L_A(\mu)$ be the set of strategic measures induced by all admissible measurable policies with $(\omega_0, \mathbf{y}) \sim \mu$; that is, $P \in L_A(\mu) \subset \mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)\Big)$ if and only if

$$P(B) = \int_{B^0 \times \prod_{k=1}^N A^k} \mu(d\omega_0, d\mathbf{y}) \prod_{k=1}^N 1_{\{u^k = \gamma^k(y^k) \in B^k\}}, \tag{10.46}$$

for all cylindrical $B \in \mathcal{B}\Big(\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)\Big)$, where $\gamma^k \in \Gamma^k$ for $k = 1, \ldots, N$. Let $L_A(\mu, \underline{\gamma})$ be the strategic measure under a particular strategy $\underline{\gamma} \in \mathbf{\Gamma}$.

The first relaxation is obtained via individual randomization of policies. Namely, let $L_R(\mu)$ be the set of strategic measures induced by all individually randomized team policies where $\omega_0, \mathbf{y} \sim \mu$; that is,

$$L_R(\mu) := \Big\{ P \in \mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)\Big) : P(B) = \int_B \mu(d\omega_0, d\mathbf{y}) \prod_{k=1}^N \Pi^k(du^k|y^k) \Big\},$$

where Π^k takes place from the set of stochastic kernels from $\mathbb{Y}^k$ to $\mathbb{U}^k$ for each $k = 1, \ldots, N$.

Another relaxation, which is stronger than the former one, is obtained by taking the mixture of the elements of $L_A(\mu)$. To this end, define $\Upsilon = [0,1]^N$. We then let

$$L_C(\mu) := \Big\{ P \in \mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^{N} (\mathbb{Y}^k \times \mathbb{U}^k) \Big) : P(B) = \int \eta(dz) L_A(\mu, \underline{\gamma}(z))(B),\ \eta \in \mathcal{P}(\Upsilon) \Big\},$$

where $\underline{\gamma}(z)$ denotes a collection of team policies measurably parametrized by $z \in \Upsilon$ so that the map $L_A(\mu, \underline{\gamma}(\cdot)) : \Upsilon \to L_A(\mu)$ is Borel measurable as $L_A(\mu)$ is a Borel subset of $\mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^{N} (\mathbb{Y}^k \times \mathbb{U}^k) \Big)$ under weak convergence topology (as we will see in Theorem 10.10).

Let L_{CR} denote the set of strategic measures that are induced by some *fixed* but common independent randomness and arbitrary private independent randomness; that is,

$$L_{CR}(\mu) := \Big\{ P \in \mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^{N} (\mathbb{Y}^k \times \mathbb{U}^k) \Big) : P(B) = \int_{B \times \Upsilon} \eta(dz) \mu(d\omega_0, d\mathbf{y}) \prod_k \Pi^k(du^k | y^k, z) \Big\},$$

where Π^k takes place from the set of stochastic kernels from $\mathbb{Y}^k \times \Upsilon$ to $\mathbb{U}^k$ for each $k = 1, \dots, N$. Here, the common randomness η is fixed.

Let L_{CCR} denote the set of strategic measures that are induced by some arbitrary but common independent randomness and arbitrary private independent randomness, as in $L_C(\mu)$; that is,

$$L_{CCR}(\mu) := \Big\{ P \in \mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^{N} (\mathbb{Y}^k \times \mathbb{U}^k) \Big) : P(B) = \int_{B \times \Upsilon} \eta(dz) \mu(d\omega_0, d\mathbf{y}) \prod_k \Pi^k(du^k | y^k, z),\ \eta \in \mathcal{P}(\Upsilon) \Big\},$$

where Π^k takes place from the set of stochastic kernels from $\mathbb{Y}^k \times \Upsilon$ to $\mathbb{U}^k$ for each $k = 1, \dots, N$. Here, the common randomness η is arbitrary, unlike $L_{CR}(\mu)$. The following result, essentially from [351], states some structural results about above-defined sets of strategic measures. In particular, it establishes convexity related properties of these sets.

There also exist further convex relaxations: Quantum Relaxations, Non-Signaling Relaxations and Local-Markov Relaxations. We do not discuss these in these notes.

Theorem 10.7. *Consider a static team problem. Then, we have the following characterizations.*
(i) $L_R(\mu)$ *has the following representation:*

$$L_R(\mu) = \Big\{ P \in \mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^{N} (\mathbb{Y}^k \times \mathbb{U}^k) \Big) : P(B) = \int U(dz) L_A(\mu, \underline{\gamma}(z))(B), \ U \in \mathcal{P}(\Upsilon), U(dv_1, \cdots, dv_N) = \prod_s \eta_k(dv_k), \eta_k \in \mathcal{P}([0,1]) \Big\};$$

that is, $U \in \mathcal{P}(\Upsilon)$ *is constructed by the product of* N *independent random variables on* $[0,1]$.
(ii) $L_C(\mu) = L_{CCR}(\mu)$ *and this is a convex set. The set of extreme points of* $L_C(\mu)$ *is* $L_A(\mu)$. *Furthermore,* $L_R(\mu) \subset L_C(\mu)$.
(iii) We have the following equalities:

$$\inf_{\underline{\gamma} \in \mathbf{\Gamma}} J(\underline{\gamma}) = \inf_{P \in L_A(\mu)} \int P(ds) c(s) = \inf_{P \in L_R(\mu)} \int P(ds) c(s) = \inf_{P \in L_C(\mu)} \int P(ds) c(s).$$

In particular, deterministic policies are optimal among the randomized class. In other words, individual and common randomness does not improve the optimal team cost.

(iv) The sets $L_R(\mu)$ and $L_{CR}(\mu)$ are not convex. In particular, the presence of independent or (fixed) common randomness does not convexify the set of strategic measures.
(v) $L_R(\mu)$ and $L_C(\mu)$ are not necessarily weakly closed.

10.7.3 Sets of strategic measures for dynamic teams in the absence of static reduction

Note that if the dynamic team setup admits a static reduction (in particular independent static reduction), then one can define strategic measures by considering equivalent static problem and characterize the convexity properties of the set of strategic measures, as done in the previous section. In this section, we suppose that dynamic team does not admit a static reduction. Let μ be the distribution of ω_0. Recall that in dynamic setup, the distribution of measurements $\mathbf{y}$ is not fixed as opposed to the static case. In this case, we present the following characterization for strategic measures in dynamic sequential teams. Let, for all $n \in \mathcal{N}$,

$$h_n = \{\omega_0, y^1, u^1, \cdots, y^{n-1}, u^{n-1}, y^n, u^n\},$$

and $p_n(dy^n|h_{n-1}) := P(dy^n|h_{n-1})$ be the transition kernel characterizing the measurements of DM n according to the intrinsic model. We note that this may be obtained by the relation:

$$\begin{aligned}
&p_n(y^n \in \cdot \,|\, \omega_0, y^1, u^1, \cdots, y^{n-1}, u^{n-1}) \\
&:= P\Big(\eta^i(\omega, \mathbf{u}^{[1,i-1]}) \in \cdot \,\Big|\, \omega_0, y^1, u^1, \cdots, y^{n-1}, u^{n-1}\Big) \\
&= P\Big(g^n(\omega_0, \omega_n, u^1, \cdots, u^{n-1}) \in \cdot \,\Big|\, \omega_0, y^1, u^1, \cdots, y^{n-1}, u^{n-1}\Big).
\end{aligned} \tag{10.47}$$

Note that once a policy is fixed, $p_n(dy^n|h_{n-1})$ represents the conditional distribution of y^n given the past history h_{n-1}. Let $L_A(\mu)$ be the set of strategic measures induced by measurable policies and let $L_R(\mu)$ be the set of strategic measures induced by individually randomized policies for the dynamic team. We have the following characterizations of $L_A(\mu)$ and $L_R(\mu)$ that are quite useful when establishing the closedness of these sets.

Theorem 10.8 ([351, Theorem 2.2]). *Consider a dynamic team problem that does not admit a static reduction. Then, we have the following characterizations.*
(i) A probability measure $P \in \mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)\Big)$ is a strategic measure induced by a measurable policy (that is in $L_A(\mu)$) if and only if, for every $n = 1, \ldots, N$, we have

$$\int P(dh_{n-1}, dy^n)\, g(h_{n-1}, y^n) = \int P(dh_{n-1}) \left(\int_{\mathbb{Y}^n} g(h_{n-1}, z)\, p_n(dz|h_{n-1}) \right)$$

and

$$\int P(dh_n)\, g(h_{n-1}, y^n, u^n) = \int P(dh_{n-1}, dy^n) \left(\int_{\mathbb{U}^n} g(h_{n-1}, y^n, a)\, 1_{\{\gamma^n(y^n) \in da\}} \right),$$

for all continuous and bounded function g with appropriate arguments, where $P(d\omega_0) = \mu(dw_0)$ and $\gamma^n \in \Gamma^n$.
(ii) A probability measure $P \in \mathcal{P}\Big(\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)\Big)$ is a strategic measure induced by a individually randomized policy (that is in $L_R(\mu)$) if and only if, for every $n = 1, \ldots, N$, we have

$$\int P(dh_{n-1}, dy^n)\, g(h_{n-1}, y^n) = \int P(dh_{n-1}) \left(\int_{\mathbb{Y}^n} g(h_{n-1}, z)\, p_n(dz|h_{n-1}) \right) \tag{10.48}$$

and

$$\int P(dh_n)\, g(h_{n-1}, y^n, u^n) = \int P(dh_{n-1}, dy^n) \left(\int_{\mathbb{U}^n} g(h_{n-1}, y^n, a^n)\, \Pi^n(da^n|y^n) \right) \tag{10.49}$$

for all continuous and bounded function g with appropriate arguments, where $P(d\omega_0) = \mu(dw_0)$ and Π^n is a stochastic kernel from $\mathbb{Y}^n$ to $\mathbb{U}^n$.

Remark 10.9. A result similar to Theorem 10.7 can also be stated for the dynamic case, in particular with regard to $L_A(\mu)$ being the set of extreme points of the convex hull of $L_R(\mu)$. The reader is referred to [351, Theorem 2.3] which essentially establishes this; see also [123, Theorem 1.c] for related discussions.

A celebrated result in economics theory, known as Kuhn's theorem [199], notes that the convex hull of admissible (i.e. those in $L_A(\mu)$) strategic measures (hence $L_C(\mu)$) is equivalent to $L_R(\mu)$ when the information structure is classical. We can thus state that this does not apply in the absence of classical-ness, as $L_R(\mu)$ would not be convex (if the information structure is not classical, then convexity fails [351, p.12]), but the convex hull of admissible policies is, by definition, convex; but the convex hull of $L_R(\mu)$ is $L_C(\mu)$.

10.7.4 Measurability properties of sets of strategic measures

As noted earlier, the set $L_A(\mu)$ is a Borel subset of $\mathcal{P}\left(\Omega_0 \times \prod_k (\mathbb{Y}^k \times \mathbb{U}^k)\right)$ under weak convergence topology. The same is true for $L_R(\mu)$, which is stated in the following theorem. This result will be crucial in the analysis to follow.

Theorem 10.10 ([351, Theorem 2.10]). *Consider a sequential (static or dynamic) team.*

(i) The set of strategic measures $L_R(\mu)$ is Borel when viewed as a subset of the space of probability measures on $\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)$ under the topology of weak convergence.

(ii) The set of strategic measures $L_A(\mu)$ is Borel when viewed as a subset of the space of probability measures on $\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)$ under the topology of weak convergence.

For further properties of the sets of strategic measures, see [351].

10.8 Existence of Optimal Solutions

The following theorem states a general existence result for static teams and for dynamic teams admitting static reduction. Its proof depends on Weierstrass Extreme Value Theorem.

Theorem 10.11. *Consider a static team or the static reduction of a dynamic team with c denoting the cost function. Let c be lower semi-continuous in $\mathbf{u}$ for every fixed $\omega_0, \mathbf{y}$ and $L_R(\mu)$ or $L_C(\mu)$ be a compact set under weak convergence topology. Then, there exists an optimal team policy. This policy can be chosen deterministic and hence induces a strategic measure in $L_A(\mu)$.*

Remark 10.12. Since the cost function c_s in independent static reduction of a dynamic team also depends on the measurements $\mathbf{y}$, we include $\mathbf{y}$ as an argument to the cost function c in the previous theorem.

Theorem 10.13. *[346, Theorem 5.2] Consider a static or a dynamic team that admits an independent static reduction. Let c be lower semi-continuous in $\mathbf{u}$ for any $\omega_0, \mathbf{y}$. Suppose further that $\mathbb{U}^i$ is $\sigma-$compact (that is, $\mathbb{U}^i = \cup_n K_n$ for a countable collection of increasing compact sets K_n) and, without any loss, the control laws can be restricted to those with $E[\phi^i(u^i)] \leq M$ for some lower semi-continuous $\phi^i : \mathbb{U}^i \to \mathbb{R}_+$ which satisfies $\lim_{n\to\infty} \inf_{u^i \notin K_n} \phi^i(u^i) = \infty$. Then, an optimal team policy exists.*

Remark 10.14. Building on [351, Theorems 2.3 and 2.5] and [152, p. 1691] (due to Blackwell's theorem on irrelevant information [48, 51], [349, p. 457]), an optimal policy, when exists, can be assumed to be *deterministic*.

So far, we presented existence results for static or dynamic teams that admit independent static reduction. In the following, we present existence results for teams that do not admit independent static reduction.

Theorem 10.15. *[351, Theorem 2.9] Consider a sequential team with a classical information structure with the further property that $\sigma(\omega_0) \subset \sigma(y^1)$ (under every policy, y^1 contains ω_0). Suppose further that $\prod_{k=1}^{N} \mathbb{U}^k$ is compact. If c is lower semi-continuous and each of the kernels p_n (defined in (10.47)) is weakly continuous so that*

$$\int f(y^n)\, p_n(dy^n|\omega_0, y^1, \ldots, y^{n-1}, u^1, \cdots, u^{n-1}) \tag{10.50}$$

is continuous in $\omega_0, y^1, \cdots, y^{n-1}, u^1, \cdots, u^{n-1}$ for every continuous and bounded f, then there exists an optimal team policy which is deterministic.

A further existence result along similar lines, for a class of static teams, is presented next.

Theorem 10.16. *[346, Theorem 5.6] Consider a static team with a classical information structure (that is, with an expanding information structure so that $\sigma(y^n) \subset \sigma(y^{n+1}), n \geq 1$). Suppose further that $\prod_{k=1}^{N}(\mathbb{Y}^k \times \mathbb{U}^k)$ is compact. If*

$$\tilde{c}(y^1, \cdots, y^N, u^1, \cdots, u^N) := E[c(\omega_0, \mathbf{u})|\mathbf{y}, \mathbf{u}]$$

is lower semi-continuous in $\mathbf{u}$ for every $\mathbf{y}$, then there exists an optimal team policy which is deterministic.

Remark 10.17. The power of this last result may first seem limited. However, some reflection leads to the conclusion that, in the continuous-time theory of stochastic control, a related but not identical argument has remarkable consequences. If one makes the measurements independent via a change of measure argument, as in Girsanov's celebrated argument, so that the information structure is first made static, and then makes the information structure classical by considering the actions at time t measurable on the filtration generated by the past noise processes and actions up to time t; the proof of Theorem 10.16 can be slightly adapted to show that such a set of measurement-action measures (with fixed marginal on the measurements) that satisfy conditional independence $u_{[0,t]} \leftrightarrow y_{[0,t]} \leftrightarrow y_s - y_t$ is weakly closed (these are known as wide-sense admissible control policies). Furthermore, the value is continuous in this joint measure on $\{(u, y)_s, s \in [0, T]\}$ and this set of measures is tight. These lead to the compactness-continuity conditions and accordingly an existence result for optimal policies follows. Furthermore, by showing that the set of $\{(u, y)_s, s \geq 0\}$ measures which have quantized support in the measurement variable are dense, one can show also that piece-wise constant control policies are nearly optimal. This allows one to approximate a continuous-time process with a (sampled) discrete-time process and the machinery developed earlier in the lecture notes are applicable. This approach is the essence of Kushner's method [203], though stated somewhat differently.

10.8.1 Some Applications and Revisiting Existence Results for Classical (Single-DM) Stochastic Control

Witsenhausen's counterexample with Gaussian variables

Consider the celebrated Witsenhausen's counterexample [325] as depicted in Figures 10.3 and 10.2: This is a dynamic non-classical team problem with y^1 and w^1 zero-mean independent Gaussian random variables with unit variance and $u^1 = \gamma^1(y^1)$, $u^2 = \gamma^2(u^1 + w^1)$ and the cost function $c(\omega, u^1, u^2) = k(y^1 - u^1)^2 + (u^1 - u^2)^2$ for some $k > 0$. Witsenhausen's counterexample can be expressed, through a change of measure argument (also due to Witsenhausen) as in (10.39).

Since the optimal policy for $\gamma^2(y^2) = E[u^1|y^1]$ and $E[(E[u^1|y^1])^2] \leq E[(u^1)^2]$, it is evident with a two-stage analysis (see [152, p. 1701]) that without any loss we can restrict the policies to be so that $E[(u^i)^2] \leq M$ for some finite M, for $i = 1, 2$; this ensures a weak compactness condition on both $\hat{\gamma}^1$ and $\hat{\gamma}^2$. Since the reduced cost $\left((k(u^1 - y^1)^2 + (u^1 - u^2)^2) \frac{\eta(y^2 - u^1)}{\eta(y^2)} \right)$ is continuous in the actions, Theorem 10.13 applies.

Existence for partially observable Markov Decision Processes (POMDPs)

Consider a partially observable stochastic control problem (POMDP) with the following dynamics.

$$x_{t+1} = f(x_t, u_t, w_t), \qquad y_t = g(x_t, v_t).$$

Here, x_t is the $\mathbb{X}$-valued state, u_t is the $\mathbb{U}$-valued the control, y_t is the $\mathbb{Y}$-valued measurement process. In this section, we will assume that these spaces are finite dimensional real vector spaces. Furthermore, (w_t, v_t) are i.i.d noise processes and $\{w_t\}$ is independent of $\{v_t\}$. The controller only has causal access to $\{y_t\}$: A deterministic admissible control policy $\mathit{\Pi}$ is a sequence of functions $\{\gamma_t\}$ so that $u_t = \gamma(y_{[0,t]}; u_{[0,t-1]})$. The goal is to minimize

$$E^{\mathit{\Pi}}_{x_0}[\sum_{t=0}^{T-1} c(x_t, u_t)],$$

for some continuous and bounded $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$.

Such a problem can be viewed as a decentralized stochastic control problem with increasing information, that is, one with a *classical* information structure.

Any POMDP can be reduced to a (completely observable) MDP [352], [262], whose states are the posterior state distributions or *beliefs* of the observer. A standard approach for solving such problems then is to reduce the partially observable model to a fully observable model (also called the belief-MDP) by defining

$$\pi_t(A) := P(x_t \in A | y_{[0,t]}, u_{[0,t-1]}), A \in \mathcal{B}(\mathbb{X})$$

and observing that (π_t, u_t) is a controlled Markov chain where π_t is $\mathcal{P}(\mathbb{X})$-valued with $\mathcal{P}(\mathbb{X})$ being the space of probability measures on $\mathbb{X}$ under the weak convergence topology. Through such a reduction, existence results can be established by obtaining conditions which would ensure that the controlled Markovian kernel for the belief-MDP is weakly continuous, that is if $\int F(\pi_{t+1}) P(d\pi_{t+1} | \pi_t = \pi, u_t = u)$ is jointly continuous (weakly) in π and u for every continuous and bounded function F on $\mathcal{P}(\mathbb{X})$.

This was studied recently in [128, Theorem 3.7, Example 4.1] and [184] (see also [73] in a control-free context). In the context of the example presented, if $f(\cdot, \cdot, w)$ is continuous and g has the form: $y_t = g(x_t) + v_t$, with g continuous and w_t admitting a continuous density function η, an existence result can be established building on the measurable selection criteria under weak continuity [165, Theorem 3.3.5, Proposition D.5], provided that $\mathbb{U}$ is compact.

On the other hand, through Theorem 10.16, such an existence result can also be established by obtaining a static reduction under the aforementioned conditions. Indeed, through (10.31), with η denoting the density of v_n, we have $P(y_n \in B | x_n) = \int_B \eta(y - g(x_n))dy$. With η and g continuous and bounded, taking $y^n := y_n$, by writing $x_{n+1} = f(x_n, u_n, w_n) = f(f(x_{n-1}, u_{n-1}, w_{n-1}), u_n, w_n)$, and iterating inductively to obtain

$$x_{n+1} = h_n(x_0, \mathbf{u}_{[0,n-1]}, \mathbf{w}_{[0,n-1]}),$$

for some h_n which is continuous in $\mathbf{u}_{[0,n-1]}$ for every fixed $x_0, \mathbf{w}_{[0,n-1]}$, one obtains a reduced cost (10.31) that is a continuous function in the control actions. Theorem 10.16 then implies the existence of an optimal control policy. This reasoning is also applicable when the measurements are not additive in the noise but with $P(y_n \in B | x_n = x) = \int_B m(y, x)\eta(dy)$ for some m continuous in x and η a reference measure.

Revisiting fully observable Markov Decision Processes with the construction presented in the chapter

Consider a fully observed Markov decision process where the goal is to minimize

$$E^{\mathit{\Pi}}_{x_0}[\sum_{t=0}^{T-1} c(x_t, u_t)],$$

for some continuous and bounded $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$. Suppose that the controller has access to $x_{[0,t]}, u_{[0,t-1]}$ at time t. This system can always be viewed as a sequential team problem with a classical information structure. Under the assumption that the transition kernel according to the usual formulation, that is $P(dx_1|x_0 = x, u_0 = u)$ is weakly continuous (in the sense discussed in the previous application above), it follows that the transition kernel according to the formulation introduced in (10.47) is also weakly continuous by an application [287, Theorem 3.5]. It follows that when $\mathbb{U}$ is compact, and hence the existence of an optimal policy follows. A similar analysis is applicable when one considers the case where $P(dx_1|x_0 = x, u_0 = u)$ is strongly continuous in u for every fixed state x and the bounded cost function is continuous only in u (this is another typical setup where measurable selection conditions hold (see Assumptions 5.2.1 and 5.2.2)).

10.9 Approximation of Optimal Solutions via Finite Approximations

In this section, we consider the finite approximation of static team problems. Since results of this section can also be applied to static reduction of dynamic teams, we suppose that the cost function c also depends on the measurements $\mathbf{y}$ (which is not the case in the original problem formulation). Recall that, in the independent static reduction of a dynamic team, the reduced cost function c_s is a function of ω_0, $\mathbf{u}$, and $\mathbf{y}$. To obtain finite approximation result, the following assumptions are imposed on the components of the model.

Assumption 10.9.1 *(a) The cost function c is continuous in $(\mathbf{u}, \mathbf{y})$ for any fixed ω_0. In addition, it is bounded on any compact subset of $\Omega_0 \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)$.*

(b) For each k, $\mathbb{U}^k$ is a closed and convex subset of a completely metrizable locally convex vector space.

(c) For each k, $\mathbb{Y}^k$ is locally compact.

(d) For any subset G of $\prod_{k=1}^N \mathbb{U}^k$, the function $w_G(\omega_0, \mathbf{y}) := \sup_{\mathbf{u} \in G} c(\omega_0, \mathbf{y}, \mathbf{u})$ is integrable with respect to $\mu(d\omega_0, d\mathbf{y})$, for any compact subset G of $\prod_{k=1}^N \mathbb{U}^k$ of the form $G = \prod_{k=1}^N G^k$.

(e) For any $\underline{\gamma} \in \mathbf{\Gamma}$ with $J(\underline{\gamma}) < \infty$ and each k, there exists $u^{k,*} \in \mathbb{U}^k$ such that $J(\underline{\gamma}^{-k}, \gamma^k_{u^{k,*}}) < \infty$, where $\gamma^k_{u^{k,*}} \equiv u^{k,*}$.

In what follows, for any subset G of $\prod_{k=1}^N \mathbb{U}^k$, we let

$$\mathbf{\Gamma}_G := \left\{ \underline{\gamma} \in \mathbf{\Gamma} : \underline{\gamma} \left(\prod_{k=1}^N \mathbb{Y}^k \right) \subset G \right\}$$

and $\mathbf{\Gamma}_{c,G} := \mathbf{\Gamma}_c \cap \mathbf{\Gamma}_G$, where $\mathbf{\Gamma}_c$ denotes the set of continuous strategies. Using these definitions, let us define the following set of strategic measures for any subset G of $\prod_{k=1}^N \mathbb{U}^k$:

$$L_A^G(\mu) := \left\{ P \in L_A(\mu) : P(B) = \int_{B^0 \times \prod_{k=1}^N A^k} \mu(d\omega_0, d\mathbf{y}) \prod_{k=1}^N 1_{\{u^k = \gamma^k(y^k) \in B^k\}},\ \underline{\gamma} \in \mathbf{\Gamma}_G \right\}.$$

Let $L_A^{G,c}(\mu)$ denote the set of strategic measures in $L_A^G(\mu)$ induced by continuous policies.

The following result states that, there exists a near optimal strategic measure whose support on the product of action spaces $\prod_{k=1}^N \mathbb{U}^k$ is convex and compact (and thus bounded) subset G of it, and conditional distributions of actions given measurements are induced by continuous policies.

Proposition 10.18. *Suppose Assumption 10.9.1 holds. Then, for any $\varepsilon > 0$ there exists a compact subset G of $\prod_{k=1}^N \mathbb{U}^k$ of the form $G = \prod_{i=1}^N G^i$, where each G^i is convex and compact, such that*

$$\inf_{P \in L_A^{G,c}(\mu)} \int P(ds)\, c(s) < J^* + \varepsilon.$$

Given any strategic measure, using Assumption 10.9.1-(e) and the fact that every measure on a Borel space is tight [249, Theorem 3.2], one can construct a strategic measure in $L_A(\mu)$ whose support on the product of action spaces is convex and compact and whose cost is $\varepsilon/2$-close to the cost of the given strategic measure. For the new strategic measure, since it has a convex and compact support on the product of action spaces, using Lusin's theorem [111, Theorem 7.5.2], we can construct a strategic measure induced by continuous policies whose cost function is $\varepsilon/2$-close to the cost of bounded support strategic measure. We can complete the proof by combining these two results.

Since each $\mathbb{Y}^i$ is a locally compact separable metric space, there exists an increasing sequence of compact subsets $\{K_l^i\}$ such that $K_l^i \subset \operatorname{int} K_{l+1}^i$ and $\mathbb{Y}^i = \bigcup_{l=1}^{\infty} K_l^i$ [4, Lemma 2.76], where $\operatorname{int} D$ denotes the interior of the set D.

Let d_i denote the metric on $\mathbb{Y}^i$. For each $l \geq 1$, let $\mathbb{Y}_{l,n}^i \coloneqq \left\{y_{i,1}, \ldots, y_{i,i_{l,n}}\right\} \subset K_l^i$ be an $1/n$-net in K_l^i. Recall that if $\mathbb{Y}_{l,n}^i$ is an $1/n$-net in K_l^i, then for any $y \in K_l^i$ we have

$$\min_{z \in \mathbb{Y}_{l,n}^i} d_i(y,z) < \frac{1}{n}.$$

For each l and n, let $q_{l,n}^i : K_l^i \to \mathbb{Y}_{l,n}^i$ be a nearest neighborhood quantizer given by

$$q_{l,n}^i(y) = \arg \min_{z \in \mathbb{Y}_{l,n}^i} d_i(y,z),$$

where ties are broken so that $q_{l,n}^i$ is measurable. If $K_l^i = [-M, M] \subset \mathbb{Y}^i = \mathbb{R}$ for some $M \in \mathbb{R}_+$, the finite set $\mathbb{Y}_{l,n}^i$ can be chosen such that $q_{l,n}^i$ becomes a uniform quantizer. We let $Q_{l,n}^i : \mathbb{Y}^i \to \mathbb{Y}_{l,n}^i$ denote the extension of $q_{l,n}^i$ to $\mathbb{Y}^i$ given by

$$Q_{l,n}^i(y) \coloneqq \begin{cases} q_{l,n}^i(y), & \text{if } y \in K_l^i, \\ y_{i,0}, & \text{otherwise,} \end{cases}$$

where $y_{i,0} \in \mathbb{Y}^i$ is some auxiliary element.

Define $\Gamma_{l,n}^i = \Gamma^i \circ Q_{l,n}^i \subset \Gamma^i$; that is, $\Gamma_{l,n}^i$ is defined to be the set of all strategies $\tilde{\gamma}^i \in \Gamma^i$ of the form $\tilde{\gamma}^i = \gamma^i \circ Q_{l,n}^i$, where $\gamma^i \in \Gamma^i$. Define also $\mathbf{\Gamma}_{l,n} \coloneqq \prod_{i=1}^N \Gamma_{l,n}^i \subset \mathbf{\Gamma}$. Note that, for any $i = 1, \ldots, N$, $\Gamma_{l,n}^i$ is the set of policies for DM i which can only use the output levels of the quantizer $Q_{l,n}^i$. In other words, in addition to measurement channel $g^i(dy^i|\omega_0)$ between DM i and the Nature, there is also an analog-to-digital converter (quantizer) between them.

Using these definitions, let us define the following set of strategic measures for any l and n:

$$L_A^{l,n}(\mu) \coloneqq \left\{ P \in L_A(\mu) : P(B) = \int_{B^0 \times \prod_{k=1}^N A^k} \mu(d\omega_0, d\mathbf{y}) \prod_{k=1}^N 1_{\{u^k = \gamma^k(y^k) \in B^k\}},\ \underline{\gamma} \in \mathbf{\Gamma}_{l,n} \right\}.$$

The following theorem states that an optimal (or almost optimal) strategic measure can be approximated with arbitrarily small approximation error for the induced costs by strategic measures in $L_A^{l,n}(\mu)$ for sufficiently large l and n.

Theorem 10.19. *[274] For any $\varepsilon > 0$, there exist $(l, n(l))$, a compact subset G of $\prod_{k=1}^N \mathbb{U}^k$ of the form $G = \prod_{i=1}^N G^i$, where each G^i is convex and compact, and $P \in L_A^{l,n(l)}(\mu) \bigcap L_A^G(\mu)$ such that*

$$\int P(ds)\, c(s) < J^* + \varepsilon$$

For each (l, n), we define a team model with finite measurement spaces. We prove that, for sufficiently large l and n, optimal strategic measure of the team model corresponding to (l, n) will provide a strategic measure to the original team model which is nearly optimal.

To this end, fix any (l,n). For the pair (l,n), the corresponding finite measurement team model has the following measurement spaces: $\mathbb{Z}^i_{l,n} \coloneqq \{y_{i,0}, y_{i,1}, \ldots, y_{i,i_{l,n}}\}$ (i.e., the output levels of $Q^i_{l,n}$), $i \in \mathcal{N}$. The stochastic kernels $g^i_{l,n}(\,\cdot\,|\omega_0)$ from Ω_0 to $\mathbb{Z}^i_{l,n}$ denotes the measurement constraints and given by:

$$g^i_{l,n}(\,\cdot\,|\omega_0) \coloneqq \sum_{j=0}^{i_{l,n}} g(S^{l,n}_{i,j}|\omega_0)\,\delta_{y_{i,j}}(\,\cdot\,),$$

where $S^{l,n}_{i,j} \coloneqq \big\{y \in \mathbb{Y}^i : Q^i_{l,n}(y) = y_{i,j}\big\}$. Indeed, $g^i_{l,n}(\,\cdot\,|\omega_0)$ is the push-forward of the measure $g^i(\,\cdot\,|\omega_0)$ with respect to the quantizer $Q^i_{l,n}$.

Let $\Phi^i_{n,l} \coloneqq \big\{\phi^i : \mathbb{Z}^i_{l,n} \to \mathbb{U}^i, \phi^i \text{ measurable}\big\}$ denote the set of measurable policies for DM i and let $\mathbf{\Phi}_{l,n} \coloneqq \prod_{i=1}^N \Phi^i_{l,n}$. The cost of this team model is $J_{l,n} : \mathbf{\Phi}_{l,n} \to \mathbb{R}_+$ and defined as

$$J_{l,n}(\underline{\phi}) \coloneqq \int_{\Omega_0 \times \prod_{i=1}^N \mathbb{Z}^i_{l,n}} c(\omega_0, \mathbf{y}, \mathbf{u})\, P_{l,n}(d\omega_0, d\mathbf{y}),$$

where $\underline{\phi} = (\phi^1, \ldots, \phi^N)$, $\mathbf{u} = \underline{\phi}(\mathbf{y})$, and

$$P_{l,n}(d\omega_0, d\mathbf{y}) = P(d\omega_0) \prod_{i=1}^N g^i_{l,n}(dy^i|\omega_0) =: \mu_{l,n}(d\omega_0, d\mathbf{y}).$$

For any compact subset G of $\prod_{k=1}^N \mathbb{U}^k$, we also define $\mathbf{\Phi}^G_{l,n} \coloneqq \{\underline{\phi} \in \mathbf{\Phi}_{l,n} : \underline{\phi}(\prod_{i=1}^N \mathbb{Z}^i_{l,n}) \subset G\}$.

In order to obtain the approximation result, we need to impose the following additional assumption.

Assumption 10.9.2 *For any compact subset G of $\prod_{k=1}^N \mathbb{U}^k$ of the form $G = \prod_{i=1}^N G^i$, we assume that the function w_G is uniformly integrable with respect to the measures $\{\mu_{l,n}\}$; that is,*

$$\lim_{R\to\infty} \sup_{l,n} \int_{\{w_G > R\}} w_G(\omega_0, \mathbf{y})\, d\mu_{l,n} = 0.$$

Note that Assumption 10.9.1-(d),(e) hold if the cost function is bounded. Indeed, conditions in Assumption 10.9.1 are quite mild and hold for the celebrated counterexample of Witsenhausen.

Theorem 10.20. *[274] Suppose Assumptions 10.9.1 and 10.9.2 hold. Then, for any $\varepsilon > 0$, there exists a pair $(l, n(l))$ and a compact subset $G = \prod_{i=1}^N G^i$ of $\prod_{k=1}^N \mathbb{U}^k$ such that an optimal (or almost optimal) strategic measure $P^{l,n(l)}$ in the set $T^G_A(\mu_{l,n(l)})$ for the $(l, n(l))$ team is ε-optimal for the original team problem when $P^{l,n(l)}$ is extended to $\Omega \times \prod_{k=1}^N (\mathbb{Y}^k \times \mathbb{U}^k)$ via quantizers $Q^i_{l,n(l)}$; that is,*

$$P^{l,n(l)}_{\text{ex}}(\cdot) = \int_{\cdot} \mu(d\omega_0, d\mathbf{y}) \prod_{k=1}^N 1_{\{u^k = \gamma^k \circ Q^k_{l,n(l)}(y^k) \in \cdot\}}$$

where

$$P^{l,n(l)}(B) = \int_{\cdot} \mu_{l,n(l)}(d\omega_0, d\mathbf{y}) \prod_{k=1}^N 1_{\{u^k = \gamma^k(y^k) \in \cdot\}}$$

10.10 Dynamic Programming and Centralized MDP Reduction Approaches to Team Decision Problems

10.10.1 Dynamic programming approach based on Common Information and a Controlled Markov State

In a team problem, if all the random information at any given decision maker is common knowledge between all decision makers, then the system is essentially centralized. If only some of the system variables are common knowledge, the remaining unknowns may or may not lead to a computationally tractable program generating an optimal solution. A possible approach toward establishing a tractable program is through the construction of a controlled Markov chain where the controlled Markov state may now live in a larger state space (for example a space of probability measures) and the actions are elements in possibly function spaces. This controlled Markov construction may lead to a computation of optimal policies.

Such a *dynamic programming approach* has been adopted extensively in the literature (see for example, [19], [338], [87], [3], [342], and generalized in [241,242]) through the use of a team-policy which uses common information to generate partial functions for each DM to generate their actions using local information. Thus, in the dynamic programming approach, a separation of team decision policies in the form of a two-tier architecture, *a higher-level controller* and a *lower-level controller*, can be established with the use of common knowledge.

In the following, we present the ingredients of such an approach, as generalized in [242] and termed *the common information approach*:

1. Elimination of irrelevant information at the DMs: In this step, irrelevant local information at the DMs, say DM k, is identified as follows. By letting the policy at other DMs to be arbitrary, the policy of DM k can be optimized as a best-response function, and irrelevant data at DM k can be removed.

2. Construction of a coordinated system: This step identifies the common information and local/private information at the DMs, after Step 1 above has been carried out. A *fictitious coordinator* (*higher-level controller*) uses the common information to generate team policies, which in turn dictates the (*lower-level*) DMs what to do with their local information.

3. Formulation of the cost function as a Partially Observed Markov Decision Process (POMDP), in view of the coordinator's optimal control problem: A fundamental result in stochastic control is that the problem of optimal control of a partially observed Markov chain (with additive per-stage costs) can be solved by turning the problem into a fully observed one on a larger state space where the state is replaced by the "belief" on the state.

4. Solution of the POMDP leads to the structural results for the coordinator to generate optimal team policies, which in turn dictates the DMs what actions to take given their local information realizations.

5. Establishment of the equivalence between the solution obtained and the original problem, and translation of the optimal policies. Any coordination strategy can be realized in the original system. Note that, even though there is no real coordinator, such a coordination can be realized implicitly, due to the presence of common information.

We will provide a further explicit setting with such a recipe at work, in the context of the *k-stage periodic belief sharing pattern* in the next section. In particular, Lemma 10.10.1 and Lemma 10.10.2 will highlight this approach. When a given information structure does not allow for the construction of a controlled Markov chain even in a larger, but fixed for all time stages, state space, one question that can be raised is what information requirements would lead to such a structure. We will also investigate this problem in the context of the *one-stage belief sharing pattern* in the next section.

k-Stage Periodic Information or Belief Sharing Pattern

In this section, we will use the term *belief* for a probability measure-valued random variable. This terminology has been used particularly in the artificial intelligence and computer science communities, which we adopt here. We will, however, make precise what we mean by such a belief process in the following.

As mentioned earlier in *Chapter 6*, a fundamental result in stochastic control is that the problem of optimal control of a partially observed Markov chain can be solved by turning the problem into a fully observed one on a larger state space

where the state is replaced by the belief on the state. Such an approach is very effective in the centralized setting; in a decentralized setting, however, the notion of a state requires further specification. In the following, we illustrate this approach under the k-step periodic belief sharing information pattern.

Consider a joint process $\{x_t, y_t, t \in \mathbb{Z}_+\}$, where we assume for simplicity that the spaces where x_t, y_t take values from are finite dimensional real-valued or countable. They are generated by

$$x_{t+1} = f(x_t, u_t^1, \ldots, u_t^L, w_t),$$

$$y_t^i = g(x_t, v_t^i),$$

where x_t is the state, $u_t^i \in \mathbb{U}^i$ is the control action, $(w_t, v_t^i, 1 \leq i \leq L)$ are second order, zero-mean, mutually independent, i.i.d. noise processes. We also assume that the state noise, w_t, either has a probability mass function, or a probability measure with a density function.

Suppose that there is a common information vector $\mathcal{I}_t^c$ at some time t, which is available to all the decision makers. At times $ks-1$, with $k > 0$ fixed, and $s \in \mathbb{Z}_+$, the decision makers share all their information: $\mathcal{I}_{ks-1}^c = \{\mathbf{y}_{[0,ks-1]}, \mathbf{u}_{[0,ks-1]}\}$ and for $\mathcal{I}_0^c = \{P(x_0)\}$, that is at time 0 the DMs have the same *a priori* belief on the initial state. Hence, at time t, DM i has access to $\{y_{[ks,t]}^i, \mathcal{I}_{ks-1}^c\}$.

Until the next *common* observation instant $t = k(s+1)-1$ we can regard the individual decision functions specific to DM i as $\{u_t^i = \gamma_s^i(y_{[ks,t]}^i, \mathcal{I}_{ks-1}^c)\}$; we let γ_s denote the ensemble of such decision functions and let $\underline{\gamma}$ denote the team policy.

It then suffices to generate γ_s for all $s \geq 0$, as the decision outputs conditioned on $y_{[ks,t]}^i$, under $\gamma_s^i(y_{[ks,t]}^i, \mathcal{I}_{ks-1}^c)$, can be generated. In such a case, we can define $\gamma_s(., \mathcal{I}_{ks-1}^c)$ to be the joint team decision rule mapping $\mathcal{I}_{ks-1}^c$ into a space of action vectors: $\{\gamma_s^i(y_{[ks,t]}^i, \mathcal{I}_{ks-1}^c), i \in \mathcal{L} = \{1, 2 \ldots, L\}, t \in \{ks, ks+1, \ldots, k(s+1)-1\}\}$.

Let $[0, T-1]$ be the decision horizon, where T is divisible by k. Let the objective of the decision makers be the joint minimization of

$$E_{x_0}^{\gamma^1, \gamma^2, \ldots, \gamma^L}[\sum_{t=0}^{T-1} c(x_t, u_t^1, u_t^2, \ldots, u_t^L)],$$

over all policies $\gamma^1, \gamma^2, \ldots, \gamma^L$, with the initial condition x_0 specified. The cost function

$$J_{x_0}(\underline{\gamma}) = E_{x_0}^{\underline{\gamma}} \sum_{t=0}^{T-1} c(x_t, \mathbf{u}_t)$$

can be expressed as:

$$J_{x_0}(\underline{\gamma}) = E_{x_0}^{\underline{\gamma}}[\sum_{s=0}^{\frac{T}{k}-1} \bar{c}(\gamma_s(., \mathcal{I}_{ks-1}^c), \bar{x}_s)]$$

with

$$\bar{c}(\gamma_s(., \mathcal{I}_{ks-1}^c), \bar{x}_s) = E_{\bar{x}_s}^{\underline{\gamma}}[\sum_{t=ks}^{k(s+1)-1} c(x_t, \mathbf{u}_t)]$$

Lemma 10.10.1 *[342] Consider the decentralized system setup above. Let $\mathcal{I}_t^c$ be a common information vector supplied to the DMs regularly every k time stages, so that the DMs have common memory with a control policy generated as described above. Then, $\{\bar{x}_s := x_{ks}, \gamma_s(\cdot, \mathcal{I}_{ks-1}^c), s \geq 0\}$ forms a controlled Markov chain.*

In view of the above, we have the following separation result.

Lemma 10.10.2 *[342] Let $\mathcal{I}_t^c$ be a common information vector supplied to the DMs regularly every k time steps. There is no loss in performance if $\mathcal{I}_{ks-1}^c$ is replaced by $P(\bar{x}_s | \mathcal{I}_{ks-1}^c)$.*

An essential issue for a tractable solution is to ensure a common information vector which will act as a sufficient statistic for future control policies. This can be done via sharing information at every stage, or some structure possibly requiring larger but finite delay.

The above motivates us to introduce the following pattern.

Definition 10.10.1 **k-stage periodic belief sharing pattern** *[342] An information pattern in which the decision makers share their posterior beliefs to reach a joint belief about the system state is called a* belief sharing information pattern. *If the belief sharing occurs periodically every k-stages ($k > 1$), the DMs also share the control actions they applied in the last $k-1$ stages, together with intermediate belief information. In this case, the information pattern is called the k-*stage periodic belief sharing information pattern.

Remark 10.21. For $k > 1$, it should be noted that, the exchange of the control actions is essential. ◇

The above generalize to models with standard Borel spaces [**?**], where weak Feller regularity are also obtained for the reduced model. Accordingly, the numerical and learning theoretic results are applicable.

10.10.2 A Universal Dynamic Program

[346] considered the following topology on control policies, while developing a universal dynamic programming algorithm applicable to any sequential decentralized stochastic control problem, generalizing Witsenhausen's program [327] which is tailored primarily for countable probability spaces.

Define

(i) State: $x_t = \{\omega_0, u^1, \cdots, u^{t-1}, y^1, \cdots, y^t\}$, $1 \leq t \leq N$.

(i') Extended State: $\pi_t \in \mathcal{P}(\Omega_0 \times \prod_{i=1}^{t} \mathbb{Y}^i \times \prod_{i=1}^{t-1} \mathbb{U}^i)$ where, for Borel $B \in \Omega_0 \times \prod_{i=1}^{t} \mathbb{Y}^i \times \prod_{i=1}^{t-1} \mathbb{U}^i$,

$$\pi_t(B) := E_{\pi_t}[1_{\{(\omega_0, y^1, \cdots, y^t; u^1, \cdots, u^{t-1}) \in B\}}].$$

Thus, $\pi_t \in \mathcal{P}(\Omega_0 \times \prod_{i=1}^{t} \mathbb{Y}^i \times \prod_{i=1}^{t-1} \mathbb{U}^i)$ where the space of probability measures is endowed with the weak convergence topology.

(ii) Control Action: Given π_t, $\hat{\gamma}^t$ is a probability measure in $\mathcal{P}(\Omega_0 \times \prod_{k=1}^{t} \mathbb{Y}^k \times \prod_{k=1}^{t} \mathbb{U}^k)$ that satisfies the conditional independence relation:

$$u^t \leftrightarrow y^t \leftrightarrow x_t = (\omega_0, y^1, \cdots, y^t; u^1, \cdots, u^{t-1})$$

(that is, for every Borel $B \in \mathbb{U}^i$, almost surely under $\hat{\gamma}^t$, the following holds:

$$P(u^t \in B | y^t, (\omega_0, y^1, \cdots, y^t; u^1, \cdots, u^{t-1})) = P(u^t \in B | y^t)$$

with the restriction

$$x_t \sim \pi_t.$$

Denote with $\Gamma^t(\pi_t)$ the set of all such probability measures. Any $\hat{\gamma}^t \in \Gamma^t(\pi_t)$ defines, for almost every realization y^t, a conditional probability measure on $\mathbb{U}^t$. When the notation does not lead to confusion, we will denote the action at time t by $\gamma^t(du^t|y^t)$, which is understood to be consistent with $\hat{\gamma}^t$.

(ii')Alternative Control Action for Static Teams with Independent Measurements: Given π_t, $\hat{\gamma}^t$ is a probability measure on $\mathbb{Y}^t \times \mathbb{U}^t$ with a fixed marginal $P(dy^t)$ on $\mathbb{Y}^t$, that is $\pi_t^{\mathbb{Y}^t}(dy^t) = P(dy^t)$. Denote with $\Gamma^t(\pi_t^{\mathbb{Y}^t})$ the set of all such probability measures. As above, when the notation does not lead to confusion, we will denote the action at time t by $\gamma^t(du^t|y^t)$, which is understood to be consistent with $\hat{\gamma}^t$. In particular, (y^t, u^t) is independent of (y^k, u^k) for $k \neq t$.

With the control actions defined as in the above [346] developed a universal dynamic program for any sequential decentralized stochastic control and established, as a corollary of the program, further existence results, one which is essentially

identical to that presented in 10.13, but slightly more restrictive in that the cost function is assumed to be continuous in all of its arguments.

Theorem 10.22. *[346]*

(i) Under the kernel (10.47) and controlled Markov construction presented, the optimal team problem admits a well-defined backwards-induction (dynamic programming) recursion.

(ii) In particular, if the problem is independent static-reducible, actions are compact-valued and the cost function is continuous, an optimal policy exists and the value function is continuous in the prior (that is, in the distribution of primitive noise variables) under weak convergence.

Remark 10.23. The above construction is related to an interpretation put forward by Witsenhausen in his standard form [327] where all the uncertainly is embedded into the initial state and the controlled system evolves deterministically. Witsenhausen had considered only countable probability spaces for an optimality analysis.

Remark 10.24. The fully observed MDP setup can be viewed as a special case of the above. In this context, by Blackwell's theorem (Theorem 5.1.1), we know that we can reduce the search space to policies that are Markov. In this case, the optimality analysis via Bellman's principle (Theorem 5.1.3) can be recovered via the Universal Dynamic Program.

10.11 Bibliographic Notes

We primarily followed [346], [351] and *Chapters 2, 3, 4* and *12* of [349] for this topic. For a more complete coverage, the reader may follow [349].

In the economics and game theory literature, information structures are also studied extensively. Stochastic team problems are termed as *identical interest games*. In this literature, $L_C(\mu)$ appears in the analysis of Aumann's correlated equilibrium [23]. Common and independent randomness discussions appear in the analysis of comparison of information structures [216]. For further discussions, including a multi-stage generalization known as communication equilibria, see [135]. For a detailed treatment, we refer the reader to [230, p. 131].

10.12 Exercises

Exercise 10.12.1 *Consider the following static team decision problem with dynamics:*

$$x_1 = ax_0 + b_1u_0^1 + b_2u_0^2 + w_0,$$

$$y_0^1 = x_0 + v_0^1,$$

$$y_0^2 = x_0 + v_0^2,$$

Here v_0^1, v_0^2, w_0 are independent, Gaussian, zero-mean with unit variance.

Let $\gamma^i : \mathbb{R} \to \mathbb{R}$ be policies of the controllers: $u_0^1 = \gamma_0^1(y_0^1), u_0^2 = \gamma_0^2(y_0^2)$.

Find

$$\min_{\gamma^1,\gamma^2} E_{\nu_0}^{\gamma^1,\gamma^2}[x_1^2 + \rho_1(u_0^1)^2 + \rho_2(u_0^2)^2],$$

where ν_0 is a zero-mean Gaussian distribution and $\rho_1, \rho_2 > 0$.

a) Find an optimal team policy $\underline{\gamma} = \{\gamma^1, \gamma^2\}$.

b) When $b_1 = b_2$ and $\rho_1 = \rho_2$, can you conclude that an optimal solution will be identical for both Decision Makers? See [278–280] for further structural results on convex and exchangeable teams.

Exercise 10.12.2 *Consider the following team decision problem with dynamics:*

$$x_{t+1} = ax_t + b_1 u_t^1 + b_2 u_t^2 + w_t,$$

$$y_t^1 = x_t + v_t^1,$$
$$y_t^2 = x_t + v_t^2,$$

Here x_0, v_t^1, v_t^2, w_t are mutually and temporally independent zero-mean Gaussian random variables.

Let $\{\gamma^i\}$ be the policies of the controllers so that $u_t^i = \gamma_t^i(y_0^i, y_1^i, \cdots, y_t^i)$ for $i = 1, 2$.

Consider:

$$\min_{\gamma^1,\gamma^2} E_{x_0}^{\gamma^1,\gamma^2}\left[\left(\sum_{t=0}^{T-1} x_t^2 + \rho_1 (u_t^1)^2 + \rho_2 (u_t^2)^2]\right) + x_T^2\right],$$

where $\rho_1, \rho_2 > 0$.

Explain if the following are correct or not:

a) For $T = 1$, the problem is a static team problem.

b) For $T = 1$, optimal policies are linear.

c) For $T = 1$, linear policies may be person-by-person-optimal. That is, if γ^1 is assumed to be linear, then γ^2 is linear; and if γ^2 is assumed to be linear then γ^1 is linear.

d) For $T = 2$, optimal policies are linear.

e) For $T = 2$, linear policies may be person-by-person-optimal.

Exercise 10.12.3 *Consider a common probability space (with a finite sample space Ω) on which the information available to two decision makers DM^1 and DM^2 are defined, such that I_1 is available at DM^1 and I_2 is available at DM^2.*

R. J. Aumann [22] defines that an information E is common knowledge between two decision makers DM^1 and DM^2, if whenever E happens, DM^1 knows E, DM^2 knows E, DM^1 knows that DM^2 knows E, DM^2 knows that DM^1 knows E, and so on.

Let Ω be finite. Suppose that one claims that an event E is common knowledge if and only if $E \in \sigma(I_1) \cap \sigma(I_2)$, where $\sigma(I_1)$ denotes the $\sigma-$field over Ω generated by information I_1 and likewise for $\sigma(I_2)$.

Is this argument correct? Provide an answer with precise arguments. You may wish to consult [22], [244], [71] and Chapter *12 of [349].*

Exercise 10.12.4 *Let X be a binary random variable. Suppose two decision makers DM 1 and DM 2 have access to some local random variables Y^1 and Y^2, respectively, defined on a common probability space and correlated with X, and exchange their conditional expectations over time. Suppose further that:*

- *the information σ-fields at each decision maker is increasing: $\mathcal{F}_t^i \subset \mathcal{F}_{t+1}^i$, $i = 1, 2$, $t \in \mathbb{Z}_+$.*
- *for all $n \in \mathbb{N}$, there exists $m > n$ such that $\mathcal{F}_m^i$ contains information on $E[X|\mathcal{F}_n^j]$, $i, j = 1, 2$. That is, the decision makers exchange their estimates (but not their raw data -Y^i is private to DM i, $i = 1, 2$-) infinitely often.*

State and rigorously justify your answers for the following:

a) [10 Points] Is there a limit for $\lim_{n\to\infty} E[X|\mathcal{F}_n^j]$, $j = 1, 2$? Either argue that the limit exists, or provide a counterexample.

b) [10 Points] For the cases where the limit exists, is it the case that

$$\lim_{n\to\infty} E[X|\mathcal{F}_n^1] = \lim_{n\to\infty} E[X|\mathcal{F}_n^2]$$

Either prove the result, or provide a counterexample.

Hint: See [65] (see also [140] and [307])

Exercise 10.12.5 *Consider a linear Gaussian system with mutually independent and i.i.d. noises:*

$$\begin{aligned} x_{t+1} &= Ax_t + \sum_{j=1}^{L} B^j u_t^j + w_t\,, \\ y_t^i &= C^i x_t + v_t^i\,, \quad 1 \le i \le L, \end{aligned} \tag{10.51}$$

with the one-step delayed observation sharing pattern.

Construct a controlled Markov chain for the team decision problem: First show that one could have

$$\{y_t^1, y_t^2, \ldots, y_t^L, P(dx_t | y_{[0,t-1]}^1, y_{[0,t-1]}^2, \ldots, y_{[0,t-1]}^L)\}$$

as the state of the controlled Markov chain.

Consider the following problem:

$$E_{\nu_0}^{\gamma}[\sum_{t=0}^{T-1} c(x_t, u_t^1, \cdots, u_t^L)]$$

For this problem, if at time $t \ge 0$ each of the decision makers (say DM i) has access to $P(dx_t | y_{[0,t-1]}^1, y_{[0,t-1]}^2, \ldots, y_{[0,t-1]}^L)$ and their local observation $y_{[0,t]}^i$, show that they can obtain a solution where the optimal decision rules only uses $\{P(dx_t | y_{[0,t-1]}^1, y_{[0,t-1]}^2, \ldots, y_{[0,t-1]}^L), y_t^i\}$:

What if, they do not have access to $P(dx_t | y_{[0,t-1]}^1, y_{[0,t-1]}^2, \ldots, y_{[0,t-1]}^L)$, and only have access to $y_{[0,t]}^i$? What would be a sufficient statistic for each decision maker for each time stage?

Exercise 10.12.6 *Two decision makers, Alice and Bob, wish to control a system:*

$$\begin{aligned} x_{t+1} &= ax_t + u_t^a + u_t^b + w_t, \\ y_t^a &= x_t + v_t^a, \\ y_t^b &= x_t + v_t^b, \end{aligned}$$

where u_t^a, y_t^a are the control actions and the observations of Alice, u_t^b, y_t^b are those for Bob and v_t^a, v_t^b, w_t are independent zero-mean Gaussian random variables with finite variance. Suppose the goal is to minimize for some $T \in \mathbb{Z}_+$:

$$E_{x_0}^{\Pi^a, \Pi^b}\left[\sum_{t=0}^{T-1} x_t^2 + r_a (u_t^a)^2 + r_b (u_t^b)^2\right],$$

for $r_a, r_b > 0$, where Π^a, Π^b denote the policies adopted by Alice and Bob. Let the local information available to Alice be $I_t^a = \{y_s^a, u_s^a, s \le t-1\} \cup \{y_t^a\}$, and $I_t^b = \{y_s^b, u_s^b, s \le t-1\} \cup \{y_t^b\}$ is the information available at Bob at time t.

Consider an $n-$step delayed information pattern: In an $n-$step delayed information sharing pattern, the information at Alice at time t is

$$I_t^a \cup I_{t-n}^b,$$

and the information available at Bob is

$$I_t^b \cup I_{t-n}^a.$$

State if the following are true or false:

a) If Alice and Bob share all the information they have (with $n = 0$), it must be that, the optimal controls are linear.

b) Typically, for such problems, for example, Bob can try to send information to Alice to improve her estimation on the state, through his actions. When is it the case that Alice cannot benefit from the information from Bob, that is for what values of n, there is no need for Bob to signal *information this way?*

c) If Alice and Bob share all information they have with a delay of 2*, then their optimal control policies can be written as*

$$u_t^a = f_a(E[x_t|I_{t-2}^a, I_{t-2}^b], y_{t-1}^a, y_t^a),$$

$$u_t^b = f_b(E[x_t|I_{t-2}^a, I_{t-2}^b], y_{t-1}^b, y_t^b),$$

for some functions f_a, f_b. Here, $E[.|.]$ denotes the expectation.

d) If Alice and Bob share all information they have with a delay of 0*, then their optimal control policies can be written as*

$$u_t^a = f_a(E[x_t|I_t^a, I_t^b]),$$

$$u_t^b = f_b(E[x_t|I_t^a, I_t^b]),$$

for some functions f_a, f_b. Here, $E[.|.]$ denotes the expectation.

11

Controlled Stochastic Differential Equations

This chapter introduces the basics of stochastic differential equations and then studies controlled such equations. A complete treatment is beyond the scope of these notes, however, the essential tools and ideas will be presented so that a student who is comfortable with the discrete-time discussion thus far in the notes can realize that with a little additional effort the continuous-time case can also be followed with ease. The reader is referred to e.g. [15, 175, 181, 204, 248] for more comprehensive treatments on various aspects ranging from mathematical foundations, stability, optimal control, filtering, and numerical methods.

Our approach here will primarily be to map the material presented so far in the notes to the continuous-time case, with the understanding that the discrete-time theory is well understood. With X_t an $\mathbb{R}$-valued random variable for each $t \in \mathbb{R}_+$, consider a stochastic process $X_t, t \in \mathbb{R}_+$. Given a sufficiently regular function f suppose that we can define

$$\lim_{h\to 0} \frac{E[f(X_h)|X_0 = x] - f(x)}{h} =: \mathcal{A}f(x), \quad x \in \mathbb{R}$$

for some map $\mathcal{A}$ (to be studied further). This means that $E[f(X_h)|X_0 = x] = f(x) + \mathcal{A}f(x)h + o(h)$, where $\frac{o(h)}{h} \to 0$ as $h \to 0$. With $\mu_t(B) = E[1_{\{X_t \in B\}}]$, for all Borel B, the above implies under mild conditions on $\mathcal{A}$ that

$$\int \mu_t(dx) f(x) = \int_0^t \left(\int \mathcal{A}f(z)\mu_s(dz) \right) ds + \int f(x)\mu_0(dx)$$

We will observe that the above can be viewed as a limit (as $h \to 0$) of interpolations of the sampled (and thus discrete with $k \in \mathbb{Z}_+$) stochastic process

$$X_{(k+1)h} = X_{kh} + hb(X_{kh}) + \sigma(X_{kh})\sqrt{h}Z \tag{11.1}$$

where $Z \sim \mathcal{N}(0, 1)$ and $\mathcal{A}f(x) = b(x)\frac{d}{dx}f(x) + \frac{1}{2}(\sigma^2(x))\frac{\partial^2 f}{\partial x^2}(x)$. In the limit as $h \to 0$, we arrive in some particular sense (that of weak convergence of path valued random processes under the topology of uniform convergence over compact sets), at the limit equation

$$dX_t = b(X_t) + \sigma(X_t)dB_t,$$

which is called a stochastic differential equation. Here, B_t is the Brownian motion.

The discussion (11.1) also leads to the following chain rule: Let $f(x, t)$ be differentiable so that the operations to follow are well-defined (e.g., twice continuously differentiable in x and continuously differentiable in t): Then, if we attempt to write

$$f(x_{t+h}, t + h) \approx f(x, t) + \frac{\partial f}{\partial t}h + \frac{\partial f}{\partial x}dx = f(x, t) + \frac{\partial f}{\partial t}h + \frac{\partial f}{\partial x}(b(x)h + \sigma(x)\sqrt{h}Z)$$

what we observe is that in the last term $\frac{\partial f}{\partial x}dx = \frac{\partial f}{\partial x}(b(x)h + \sigma(x)\sqrt{h}Z)$, when normalization by h is made, the expression $\sqrt{h}/h$ *does not* decay to zero and and the second derivative term appearing in the Taylor's expansion, which would be $\frac{1}{2}\frac{\partial^2 f}{\partial x^2}(dx)(dx)$, is *non-negligible*. Accordingly, a more appropriate expression is:

$$f(x_{t+h}, t+h) \approx f(x,t) + \frac{\partial f}{\partial t}h + \frac{\partial f}{\partial x}dx + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(dx)(dx)$$

leading to

$$f(x_{t+h}, t+h) \approx f(x,t) + \frac{\partial f}{\partial t}h + \frac{\partial f}{\partial x}(hb(x_t) + \sigma(x_t)\sqrt{h}Z) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\sigma^2(x_t)hZ^2$$

This essentially leads to Itô's formula to be studied. A number of technical questions will arise with respect to the notion of convergence as $h \downarrow 0$ and the non-differentiability of the Brownian process B_t. This model will be generalized, and there will also be control entering the flow, e.g. via $b(x,u)$ with u denoting the control term and possibly in $\sigma(\cdot)$ as well.

We will restrict the model to certain systems, e.g. those driven by the Brownian process, though one can in principle study more general models (the term multiplying $\sigma(X_{kh})$ does not need to be a Gaussian measure and there exist many other processes that can be considered.

We should note that the construction of a stochastic process on a continuous time interval, such as $[0,T]$ requires more caution when compared with a discrete-time stochastic process, as we will observe. In this chapter, we will primarily be concerned with controlled Markov processes X_t, each taking values in $\mathbb{R}^n$ for $t \in [0,T]$ or $t \in [0,\infty)$ and where the integration term involves the Brownian process or semimartingale processes [175].

11.1 Continuous-time Markov processes

11.1.1 Two ways to construct a continuous-time Markov process

As discussed in *Chapter 1* and Section 1.4, one way to define a stochastic process is to view it as a vector valued random variable. This requires us to place a proper topology on the set of sample paths, to be discussed further below.

Another definition would involve defining the process on finitely many time instances: Let $\{X_t(\omega), t \in [0,T]\}$ be stochastic process so that for each t, $X_t(\omega)$ is an $\mathbb{R}^n$-valued random variable measurable on some probability space $(\Omega, \mathcal{F}, P)$. We can define the σ-algebra generated by *cylinder* sets (as in *Chapter 1*) of the form:

$$\{\omega \in \Omega : X_{t_1}(\omega) \in A_1, X_{t_2}(\omega) \in A_2, \cdots, X_{t_N}(\omega) \in A_N, A_k \in \mathcal{B}(\mathbb{R}^n), N \in \mathbb{N}\}$$

By defining a stochastic process in this fashion and assigning probabilities to such finite dimensional events, Theorem 1.2.3 implies that there exists a unique stochastic process on the σ-algebra generated by the sets of this form. However, unlike a discrete-time stochastic process, in general, not all properties of the stochastic process are captured by finite dimensional distributions of it and the σ-field generated by such sets is not a sufficiently rich set of sets. For example the set of sample paths that satisfy $\sup_{t\in[0,1]} |X_t(\omega)| \leq 10$ may not be a well-defined event (that is, a set) in this σ-algebra. Likewise, the extension theorem considered in Theorem 1.2.2 requires a probability measure already defined on the cylinder sets; it may not be possible to define such a probability measure by only considering finite dimensional distributions [332].

If a stochastic process has continuous sample paths, then by specifying the process on rational time instances will uniquely define the process. Thus, if the process is known to admit certain regularity properties, the technical issues with regard to defining a process on finitely many sample points will disappear.

11.1.2 The Brownian motion

Definition 11.1.1 *A stochastic process B_t is called a Wiener process or Brownian motion if (i) the finite dimensional distributions of B_t are such that for any $n \in \mathbb{N}$ and any sequence $0 < t_1 < t_2 < \cdots, t_n$, the collection of random variables $B_{t_1}, B_{t_2} - B_{t_1}, \cdots, B_{t_n} - B_{t_{n-1}}$ are independent Gaussian zero mean random variables with $B_k - B_s \sim \mathcal{N}(0, k-s)$, and (ii) B_t has continuous sample paths.*

Such a process exists and can be constructed as a limit of random walks as briefly suggested in Remark 11.1 below. Going back to the construction we discussed in the previous section, we can define the Brownian motion as a $C([0,\infty))$-valued

(that is, a continuous path valued) random variable: The topology on $C([0,\infty))$ is the topology of uniform convergence on compact sets (this is a stronger convergence than the topology of point-wise convergence but weaker than the topology of uniform convergence over $\mathbb{R}$). This is in agreement with the finite dimensional characterization through which we could define the Brownian motion.

Remark 11.1. [Why Brownian Motion?] The Gaussian property of the continuous limit process is universal in the sense that, any continuous time process with sufficiently regular independent increments must be the Brownian process (via a result known as Donsker's theorem). In particular, even though typically in the construction of the Brownian motion (or its existence), one considers Gaussian i.i.d. random increments and takes its limit; this is not necessary for the Gaussian properties of the limit: Let $\{Z_1, Z_2, \cdots, \}$ be an i.i.d. random sequence with mean 0 and variance 1. For each $n \in \mathbb{N}$ define the random variable (with variance t for each $n \in \mathbb{N}$):

$$W_n(t) = \frac{1}{\sqrt{n}} \sum_{1 \le k \le \lfloor nt \rfloor} Z_k, \quad t \in [0,1]. \tag{11.2}$$

This is a random function. By the central limit theorem, $W_n(t) \to \mathcal{N}(0,t)$ (in distribution, that is weakly) for each $t \in \mathbb{R}_+$ and the same holds for any finite collection of time values. With this insight, one can also show that the path-valued random variable converges weakly (where one needs to define an appropriate metric on the path-valued realization space, as the elements of the sequence may not be continuous) to the standard Brownian motion. In this context, an appropriate topology is the *Skorokhod topology* defined on the space of functions which are right continuous with left limits: Such a topology defines a separable metric space [42].

For many interesting properties of the Brownian motion, the reader is referred to [252].

Remark 11.2 (Going beyond the Brownian motion). While the discussion above justifies the typical usage of the Brownian motion for many stochastic integration models studied later in the chapter, one can consider more general processes (known as semimartingales) for the analysis in the following sections to be applicable [175]. Some applications may force one to even be more general and consider driving signals that are to be studied under the theory of rough-paths [136, 221], which seeks to give a sample path sense meaning to stochastic integration, to be discussed further, as well as several robustness properties to approximate models and continuity of solutions in the driving noise.

On White Noise

In many physical systems, one encounters models of the form

$$\frac{dx}{dt} = (f(x_t) + u_t) + n_t,$$

where n_t is some noise process. In engineering, one would like to model the noise process to be *white*, in the sense that n_t, n_s are independent for $t \neq s$, and n_t is zero-mean. We call such a process *white*, because the Fourier transform (and thus the frequency spectrum) of the correlation function defined as $R(\tau) = E[n_t n_{t+\tau}]$ of such a process is a constant: If the process is a discrete-time process with finite support, then this interpretation would be directly applicable since the Fourier transform of a discrete-time impulse would be constant for all frequency values. For a continuous-time process, however, if $R(\tau) = E[n_t n_{t+\tau}] = 0$ for all $\tau \neq 0$, then a mathematical complication arises: If $R(0) < \infty$, then this signal has zero-energy and its Fourier transform would be identically 0. If $R(0) = \infty$, then such an R would have significant irregularities; such a process would have its correlation function as $E[n_s n_t] = \delta(t-s)$; where the Dirac delta function δ is a *distribution* acting on a proper set of test functions (such as the Schwartz signals $\mathcal{S}$ [174]). Such a process is not a well-defined Gaussian process since it does not have a well-defined correlation function as δ itself is not a function. But, one can view this process as a distribution, or always cautiously always work under an integral; this way one can make an operational use for such a definition.

With such a cautious interpretation, as we see in Exercise 11.10.4, the Fourier transform of a Brownian motion, over a bounded support, is i.i.d. across its discrete spectrum coefficients, and Gaussian. This justifies the term *white noise*. Thus, it is evident that n_t would not be a well-defined process and instead of n_t, we will only work with its integral B_t. Thus,

while working with B_t, instead of derivatives, we will study integral equations. On the other hand, it will be evident that we cannot take the ordinary Lebesgue or Riemann integrations for B_t (unless the function we work with is too regular) since B_t is too irregular. Instead, a method to obtain integrations will be introduced: the Itô integral. The properties of integration, differentiation, chain rule etc. for such integrations is called *stochastic calculus*. Later in the chapter, we will add control to the dynamics.

11.2 Stochastic Integration, the Itô Integral and Stochastic Differential Equations

11.2.1 Some subtleties on stochastic integration

We define a differential equation or a stochastic integral as an appropriate limit to make sense of the expression:

$$\int_0^T f(s,\omega)dB_s(\omega)$$

In the following, let $t_k^{(n)} := kT2^{-n}, k = \{0, 1, \cdots, 2^n - 1\}$. Thus, we have $B_{t_k^n} := B_{kT2^{-n}}$.

We first note that one cannot define the above in the Riemann-Stieltjes sense (i.e., by partitioning the domain and taking limits as the partition gets finer) for an arbitrary measurable f[1]. To gain further insight as to why this leads to an issue, we discuss the following. Using the *independent-increments* property (that is (i) in Definition 11.1.1) of the Brownian motion (e.g. via the construction of (11.2)), the following can be shown:

Lemma 11.2.1 *In L_2 (that is, mean-square) and hence in probability*

$$\lim_{n\to\infty}\sum_k (B_{t_{k+1}^n} - B_{t_k^n})^2 = T.$$

Observe the following [312].

Theorem 11.2.1 *Define the total variation of the Brownian process in the interval $[a, b]$ as:*

$$TV(B,a,b) = \sup_{a\le t_1\le t_2\le\cdots\le t_k\le b, k\in\mathbb{N}} \sum_k |B_{t_{k+1}} - B_{t_k}|$$

Almost surely, $TV(B, a, b) = \infty$.

Proof. By Lemma 11.2.1, and Theorem B.3.2, it follows that there exists some subsequence n_m so that $\sum_k (B_{t_{k+1}^{n_m}} - B_{t_k^{n_m}})^2 \to b - a$ almost surely (see Theorem B.3.2). Now, if $TV(B, a, b) < \infty$, this would imply that

$$\sum_k (B_{t_{k+1}^{n_m}} - B_{t_k^{n_m}})^2 \le \sup_k |B_{t_{k+1}^n} - B_{t_k^n}| \sum_k |B_{t_{k+1}^n} - B_{t_k^n}| \to 0,$$

as $n \to \infty$, since by continuity of sample paths (which would then be uniformly continuous due to compactness of the support, for any sample path with probability one) $\sup_k |B_{t_{k+1}^n} - B_{t_k^n}| \to 0$. This would lead to a contradiction. ⋄

To appreciate some subtleties on stochastic integration, let us consider simple functions of the form:

$$f(t,\omega) = \sum_{k=0}^{2^n-1} e_k(\omega) 1_{\{t\in[kT2^{-n},(k+1)T2^{-n}]\}}$$

where $n \in \mathbb{N}$. Let us define

[1] However, this would be applicable if one has further regularities on the integrand f [339].

$$\int_0^T f(t,\omega)dB_t(\omega) = \sum_k e_k(\omega)[B_{t_{k+1}}(\omega) - B_{t_k}(\omega)]$$

where $t_k = t_k^{(n)} = kT2^{-n}, k = \{0, 1, \cdots, 2^n - 1\}$. In the following, we use the notation: $B_{t_k^n} := B_{kT2^{-n}}$.

Now, note that if one has:

$$f_1(t,\omega) = \sum_k B_{\{kT2^{-n}\}} 1_{\{t\in[kT2^{-n},(k+1)T2^{-n}]\}}$$

it can be shown that

$$E[\int_0^T f_1(t,\omega)dB_t(\omega)] = 0,$$

but instead with

$$f_2(t,\omega) = \sum_k B_{\{(k+1)T2^{-n}\}} 1_{\{t\in[kT2^{-n},(k+1)T2^{-n}]\}}$$

it can be shown that

$$E[\int_0^T f_2(t,\omega)dB_t(\omega)] = T.$$

Thus, even though both f_1 and f_2 look to be reasonable approximations for some function $f(t,\omega)$, such as $B_t(\omega)$, the integrals have drastically different meanings.

In particular the variations in the B_t process is too large to define an integration (in the usual sense of Riemann-Stieltjes), as we discussed above: It does make a difference on whether one defines $\int_0^T f(t,\omega)dB_t(\omega)$ as an appropriate limit of a sequence of expressions

$$\sum_k f(t_j^*,\omega)(B_{t_{k+1}^n}(\omega) - B_{t_k^n}(\omega))$$

for different $f(t_j^*,\omega)$ with $(t_j^* \in [t_j, t_{j+1}]$. If we take $t_j^* = t_j$ (the left end point), this is known as the *Itô Integral*. If we take $t_j^* = \frac{1}{2}(t_j + t_{j+1})$, this is known as the *Stratonovich* integral (and is denoted with $\int f \circ dB_t$, to distinguish it from the Itô integral).

11.2.2 The Itô Integral

Itô's integral will be well-defined[2], if we restrict the integrand $f(t,\omega)$ to be such that $f(t,\omega)$ is measurable on the σ-field generated by $\{B_s, s \le t\}$. We define $\mathcal{F}_t$ to be the σ-algebra generated by $B_s, s \le t$. In other words, $\mathcal{F}_t$ is the smallest σ-algebra containing sets of the form:

$$\{\omega : B_{t_1}(\omega) \in A_1, \cdots, B_{t_k}(\omega) \in A_k\}, \quad t_k \le t,$$

for Borel $A_1, \cdots, A_k$. We also assume that all sets of measure zero are included in $\mathcal{F}_t$ (this operation is known as the completion of a σ-field).

Definition 11.2.1 *Let $\mathcal{N}_t$, $t \ge 0$, be an increasing family of σ-algebras of subsets of Ω. A process $g(t,\omega)$ is called $\mathcal{N}_t$ adapted if for each t, $g(t,\cdot)$ is $\mathcal{N}_t$-measurable.*

Definition 11.2.2 *Let $\mathcal{V}(S,T)$ be the class of functions:*

$$f(t,\omega) : [0,\infty) \times \Omega \to \mathbb{R}$$

[2]In the theory of integration, as a student has seen over many courses, one chooses a definition of integration and identifies conditions under which integration is possible. Of course, one expects that different integration concepts should be compatible whenever they are simultaneously applicable. We have seen the Riemann integration and the Lebesgue integration, and how they both are defined as limits of particular constructions. We will see in the following that the Itô integration has a similar flavour but with a very different construction. This approach carries over to other types of integrations, such as the rough integral [136, 221].

such that (i) $f(t,\omega)$ is $\mathcal{B}([0,\infty)) \times \mathcal{F}$-measurable, (ii) $f(t,\omega)$ is $\mathcal{F}_t$-adapted and (iii) $E[\int_S^T f^2(t,\omega)dt] < \infty$.

We will often take $S = 0$ in the following. For functions in $\mathcal{V}$, the Itô integral is defined as follows: A function ϕ is called elementary if it has the form:

$$\phi(t,\omega) = \sum_k e_k(\omega) 1_{\{t \in [t_k, t_{k+1})\}}$$

with e_k being $\mathcal{F}_{t_k}$-measurable. For elementary functions, we *define* the Itô integral as:

$$\int_0^T \phi(t,\omega) dB_t(\omega) = \sum_k e_k(\omega) \Big(B_{t_{k+1}}(\omega) - B_{t_k}(\omega) \Big) \tag{11.3}$$

With this definition, it follows that for a bounded and elementary ϕ,

$$E\left[\left(\int_0^T \phi(t,\omega) dB_t(\omega) \right)^2 \right] = E[\int_0^T \phi^2(t,\omega) dt]. \tag{11.4}$$

This property is known as the Itô isometry. The proof follows from expanding the summation in (11.3) and using the properties of the Brownian motion. Now, the remaining steps to define the Itô integral are as follows:

- **Step 1**: Let $g \in \mathcal{V}$ and $g(\cdot,\omega)$ be continuous for each ω. Then, there exist elementary functions $\phi_n \in \mathcal{V}$ such that

  $$E[\int_0^T (g - \phi_n)^2 dt] \to 0,$$

  as $n \to \infty$. The proof here follows from the dominated convergence theorem.

- **Step 2**: Let $h \in \mathcal{V}$ be bounded. Then there exist $g_n \in \mathcal{V}$ such that $g_n(\cdot,\omega)$ is continuous for all ω and n and

  $$E[\int_0^T (h - g_n)^2 dt] \to 0.$$

  One can follow Lusin's theorem (Theorem D.5.1) to establish this result.

- **Step 3**: Let $f \in \mathcal{V}$. Then, there exists a sequence $h_n \in \mathcal{V}$ such that h_n is bounded for each n and

  $$E[\int_0^T (f - h_n)^2 dt] \to 0.$$

Here, we use truncation and then the dominated convergence theorem.

Definition 11.2.3 (The Itô Integral) *Let $f \in \mathcal{V}(S,T)$ and ϕ_n be an approximating sequence of elementary functions as above given in Steps 1-2-3. The Itô integral of f is defined by*

$$\int f(t,\omega) dB_t(\omega) = \lim_{n\to\infty} \int_0^T \phi_n(t,\omega) dB_t(\omega),$$

where the convergence to the limit is in $L_2(P)$ in the sense; that is,

$$\lim_{n\to\infty} E\left[\left(\int_0^T \phi_n(t,\omega) dB_t(\omega) - \int f(t,\omega) dB_t(\omega) \right)^2 \right] = 0$$

The existence of a limit is established through the construction of a Cauchy sequence and the completeness of $L_2(P)$, the space of measurable functions with a finite second moment under P, with the corresponding norm. A computationally useful result is the following (generalizing (11.4)).

Corollary 11.2.1 *For all* $f \in \mathcal{V}(0,T)$

$$E\Big[\Big(\int_0^T f(t,\omega)dB_t(\omega)\Big)^2\Big] = E\Big[\int_0^T f^2(t,\omega)dt\Big].$$

And thus, if $f, f_n \in \mathcal{V}(0,T)$ *and*

$$E[\int_0^T (f_n - f)^2 dt] \to 0,$$

then in $L_2(P)$

$$\int_0^T f_n(t,\omega)dB_t(\omega) \to \int_0^T f(t,\omega)dB_t(\omega)$$

Example 11.3. Let us show that

$$\int_0^t B_s dB_s = \frac{1}{2}B_t^2 - \frac{1}{2}t. \tag{11.5}$$

With $B_j^n := B_{t_j^n}$, define the elementary function: $\phi_n(\omega) = \sum B_j^n(\omega)1_{\{t\in[t_j^n, t_{j+1}^n)\}}$, it follows that $E[\int_0^t(\phi_n - B_s)^2 ds] \to 0$. Therefore, the limit of the integrals of $\phi_n(\omega)$, that is the $L_2(P)$ limit of $\sum_j B_j^n(B_{j+1}^n - B_j^n)$, will be the integral. Observe now that

$$-(B_{j+1}^n - B_j^n)^2 = 2B_j^n(B_{j+1}^n - B_j^n) + B_j^{n2} - (B_{j+1}^n)^2$$

and thus summing over j, we obtain

$$\sum_j -(B_{j+1}^n - B_j^n)^2 = \sum_j 2B_j^n(B_{j+1}^n - B_j^n) + B_j^{n2} - (B_{j+1}^n)^2,$$

leading to

$$B_t^2 - \sum_j (B_{j+1}^n - B_j^n)^2 = \sum_j 2B_j^n(B_{j+1}^n - B_j^n) + B_0^2,$$

with $B_0 = 0$. Now, taking the intervals $[j, j+1]$ arbitrarily small, we see that the first term converges to $B_t^2 - t$ (see Lemma 11.2.1) and the term on the right hand side converges to $2\int_0^t B_s dB_s$, leading to the desired result. We will derive the same result using Itô's formula shortly. The message of this example is to highlight the computational method: Find a sequence of elementary function which converges in $L_2(P)$ to f, and then compute the integrals, and take the limit as the intervals shrink.

Remark 11.4. An important extension of the Itô integral is to a setup where f_t is $\mathcal{H}_t$-measurable, where $\mathcal{H}_t \subset \mathcal{F}_t = \sigma(B_s, s \le t)$. In applications, this is important to let us apply the integration to settings where the process that is integrated is measurable only on a subset of the filtration generated by the Brownian process. This allows one to define multi-dimensional Itô integrals as well. This is particularly useful for controlled stochastic differential equations, where the control policies are measurable with respect to a filtration that does not contain that generated by the Brownian motion, but the controller policy cannot depend on the future realizations of the Brownian motion either.

Remark 11.5 (Itô vs. Stratonovich Integrations). A curious reader may question the selection of choosing the Itô integral over any other, and in particular the Stratonovich integral: Different applications are more suitable for either interpretation. In stochastic control, measurability aspects (of admissible controls) are the most crucial ones. If one appropriately defines the functional or stochastic dependence between a function to be integrated or a noise process, the application of either will come naturally: If the functions are to not look at the future, then Itô's formula is appropriate. However, for many applications involving white noise-like disturbances, physical processes, or stochastic stability, ergodicity [17, 194] and smoothness properties of densities of solutions [155] to stochastic differential equations where one would build on connections with geometric control theory [297] with piece-wise constant control action sequences replacing the driving noise process, Stratonovich integral has been shown to be more relevant. Additionally, the Stratonovich integration has desirable robustness properties with regard to the approximation of the Brownian noise, as will be discussed in Section

11.8. A conclusion is that the application itself should determine the right notion of the stochastic integral to be used, in view of the assumptions imposed by the application.

11.2.3 The Itô Formula

Now that we have defined integration, we will study a generalization of the chain rule in classical calculus: Itô's formula allows us to take integrations of functions of processes.

Definition 11.2.4 *We say $v[0,T) \in \mathcal{W}_{\mathcal{H}}$ if*

$$v(t,\omega) : [0,T) \times \Omega \to \mathbb{R}$$

is such that (i) $v(t,\omega)$ is $\mathcal{B}([0,\infty)) \times \mathcal{F}$-measurable, (ii) $v(t,\omega)$ is $\mathcal{H}_t$-adapted where $\mathcal{H}_t$ is as in Remark 11.4 and (iii) $P(\int_0^T f^2(t,\omega)dt < \infty) = 1$.

Definition 11.2.5 (Itô Process) *Let B_t be a one-dimensional Brownian motion on $(\Omega, \mathcal{F}, P)$. A (one-dimensional) Itô process is a stochastic process X_t on $(\Omega, \mathcal{F}, P)$ of the form*

$$X_t = X_0 + \int_0^t b(s,\omega)ds + \int_0^t v(s,\omega)dB_s \tag{11.6}$$

where $v \in \mathcal{W}_{\mathcal{H}}$ so that v is $\mathcal{H}_t$-adapted and $P(\int_0^t v^2(t,\omega)dt < \infty) = 1$ for all $t \geq 0$. Likewise, b is also $\mathcal{H}_t$-adapted and $P(\int_0^t b^2(t,\omega)dt < \infty) = 1$ for all $t \geq 0$.

Instead of the integral form in (11.6), we may use the differential form notation:

$$dX_t = bdt + vdB_t,$$

with the understanding that this means the integral form.

Theorem 11.2.2 *[Itô Formula] Let X_t be an Itô process given by*

$$dX_t = bdt + vdB_t.$$

Let $g(t,x) \in C^{1,2}([0,\infty) \times \mathbb{R})$ (that is, g is continuously differentiable, C^1, in t, and twice continuously differentiable, C^2, in x). Then,

$$Y_t = g(t, X_t),$$

is again an Itô process and

$$dY_t = \frac{\partial g}{\partial t}(t,X_t)dt + \frac{\partial g}{\partial x}(t,X_t)dX_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t,X_t)(dX_t)^2$$

where

$$(dX_t)^2 = (dX_t)(dX_t)$$

with $dtdt = dtdB_t = dB_t dt = 0$ and $dB_t dB_t = dt$

Remark 11.6. Let us note that if instead of dB_t, we only had a differentiable function m_t so that $dX_t = bdt + vdm_t$, the regular chain rule would lead to:

$$dY_t = \frac{\partial g}{\partial t}(t,X_t)dt + \frac{\partial g}{\partial x}(t,X_t)(udt + vdm_t).$$

Note then that Itô's Formula is a generalization of the ordinary *chain rule* for derivatives. The difference is the presence of the quadratic term that appears in the formula; see also the discussion at the beginning of the chapter.

Example 11.7. Let us compute

$$\int_0^t B_s dB_s.$$

View the above as an application of Itô formula with $X_t = B_t$ so that $dX_t = dB_t$, and $Y_t = g(t, X_t) = \frac{1}{2}X_t^2$. Then, by Itô's formula,

$$dY_t = d(g(t, X_t)) = X_t dX_t + \frac{1}{2}dt = B_t dB_t + \frac{1}{2}dt$$

and thus

$$\int dY_s = \frac{1}{2}B_t^2 = \int B_s dB_s + \frac{1}{2}t.$$

Noting that $\int dY_s = Y_t - Y_0$, this result is in agreement with (11.5).

Itô's Formula can be extended to higher dimensions by considering each coordinate separately.

11.2.4 Stochastic Differential Equations

Consider now an equation of the form:

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t \tag{11.7}$$

with the interpretation that this means

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dB_s$$

Three natural questions are as follows: (i) Does there exist a solution to this differential equation? (ii) Is the solution unique? (iii) How can one compute the solution?

Theorem 11.2.3 (Existence and Uniqueness Theorem) *Let $T > 0$ and $b : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$ be measurable functions satisfying:*

$$|b(t, x)| + \|\sigma(t, x)\| \leq C(1 + |x|), \quad t \in [0, T], x \in \mathbb{R}^n$$

for some $C \in \mathbb{R}$ with $\|\sigma\|^2 := \mathrm{Trace}(\sigma\sigma^T)$, and

$$|b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq D(|x - y|), \quad t \in [0, T], x, y \in \mathbb{R}^n$$

for some constant D. Let $X_0 = Z$ be a random variable which is independent of $B_s, s \geq 0$ with $E[|Z|^2] < \infty$. Then, the stochastic differential equation (11.7) has a unique solution $X_t(\omega)$ that is continuous in t with the property that X_t is adapted to the filtration generated by $\{Z, B_s, s \leq t\}$ and $E[\int_0^T |X_t|^2] < \infty$.

Proof Sketch. The proof of existence follows from a similar construction for the existence of solutions to ordinary differential equations: One defines a sequence of iterations:

$$Y_t^{k+1} = X_0 + \int_0^t b(s, Y_t^k)ds + \int_0^t \sigma(s, Y_t^k)dBs$$

with $Y_t^0 := X_0$ for all $t \in [0, T]$. Then, the goal is to obtain a bound on the L_2-errors so that

$$\lim_{m,n\to\infty} E[|Y_t^m - Y_t^n|^2] \to 0,$$

so that Y_t^n is a Cauchy sequence under the $L_2(P)$ norm; this is where the Lipschitz bounds in the hypothesis are utilized. Call the limit X. The next step is to ensure that X indeed satisfies the equation and that there can only be one solution.

Finally, one proves that X_t can be taken to be continuous. ◇

Let us appreciate some of the conditions stated above in the context of deterministic models.

Remark 11.8. Consider the following *deterministic* differential equations:

– $$\frac{dx}{dt} = 4\frac{x}{t}$$
with $x(1) = 1$ does not admit a unique solution on the interval $[-1, 1]$.

– The differential equation $$\frac{dx}{dt} = x^2$$
with $x(0) = 1$ admits the solution $x_t = \frac{1}{1-t}$ and as $t \uparrow 1$, the solution *blows up* in finite time so that there is no solution for $t \geq 1$.

The solution discussed above is what is called a *strong solution*. Such a solution is such that X_t is unique for a given sample path. Furthermore, the solution is measurable on the filtration generated by the Brownian motion and the initial variable (which can be seen by the construction of the integral, where each pre-limit approximation is measurable on the filtration, and since L_2-limit implies a pointwise almost sure limit along a subsequence, the limit is also measurable, assuming completeness of the filtration). Such a solution concept has an important engineering/control appeal in that the solution is completely specified once the realizations of the Brownian motion (together with the initial state) are specified.

Weak solutions. In many applications, however, the conditions of Theorem 11.2.3 do not hold. In this case, one cannot always find a strong solution. However, in this case, one may be able to find a solution which satisfies the probabilistic flow in the system so that the evolution of the probabilities are well-defined: Note, however that, this solution may no longer be adapted to the filtration generated by the actual Brownian motion and the initial state; but may be adapted to *some other* Brownian process defined on *some* probability space. Such a solution is called a *weak solution* or a *martingale solution*. While such a definition has a physical interpretation limitation in the sense that the input-output relation does not correspond to one where the noise is an input and the solution is the output, this concept is instrumental in studying controlled stochastic differential equations as we will discuss later in the chapter and is appropriate if one is concerned with expected behaviour of the solutions. This concept is also related to the solution to the Fokker-Planck equation that we will discuss further in the chapter in Section 11.2.6. For weak solutions, it suffices to have b to be bounded and only σ to satisfy the Lipschitz continuity and the growth conditions provided that $\sigma(\cdot)\sigma^T(\cdot)$ has its eigenvalues uniformly bounded from below (at least locally). We will discuss this further in the context of Girsanov's measure transformation.

11.2.5 Some Properties of SDEs

Definition 11.2.6 *A diffusion (also called Itô diffusion) is a stochastic process $X_t(\omega)$ satisfying a stochastic differential equation of the form:*

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \quad t \geq s, X_s = x$$

where B_t is m-dimensional Brownian motion and b, σ satisfy the conditions of Theorem 11.2.3 so that

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq D|x - y|.$$

Note that here b, σ only depend on x and not on t. Thus, the process here is time-homogenous.

Theorem 11.2.4 *Let X_t be a diffusion and f be bounded and (Borel) measurable. Then, for $t, h \geq 0$:*

$$E_x[f(X_{t+h})|\mathcal{F}_t](\omega) = E_{X_t(\omega)}[f(X_h)]$$

Theorem 11.2.5 (Strong Markov Property) *Let f be bounded and Borel, and τ be a stopping time with respect to $\mathcal{F}_t = \sigma(\{B_s, s \leq t\})$. Then, for $h \geq 0$, conditioned on the event that $\tau < \infty$:*

$$E_x[f(X_{\tau+h})|\mathcal{F}_\tau](\omega) = E_{X_\tau(\omega)}[f(X_h)]$$

Definition 11.2.7 *Let X_t be a time-homogenous Itô diffusion in $\mathbb{R}^n$. The infinitesimal generator $\mathcal{A}$ of X_t is defined by:*

$$\mathcal{A}f(x) = \lim_{t \to 0} \frac{E_x[f(X_t)] - f(x)}{t}, \quad x \in \mathbb{R}^n,$$

whenever f is so that the limit is defined.

Lemma 11.2.2 *Let $Y_t = Y_t^x$ be an Itô process in $\mathbb{R}^n$ of the form:*

$$Y_t^x(\omega) = x + \int_0^t u(s,\omega) + \int_0^t v(s,\omega) dB_s(\omega).$$

Let $f \in C_c^2(\mathbb{R}^n)$, that is f is twice continuously differentiable and has compact support, and τ be a stopping time with respect to $\mathcal{F}_t$ with $E_x[\tau] < \infty$. Assume that u, v are bounded. Then,

$$E[f(Y_\tau)] = f(x) + E_x\left[\int_0^\tau \left(\sum_i u^i(s,\omega)\frac{\partial f}{\partial x^i}(Y_s) + \frac{1}{2}\sum_{i,j}(vv^T)_{ij}(s,\omega)\frac{\partial^2 f}{\partial x^i \partial x^j}(Y_s)\right) ds\right].$$

This lemma, combined with Definition 11.2.7 gives us the following result:

Theorem 11.2.6 *Let $dX_t = b(X_t)dt + \sigma(X_t)dB_t$. If $f \in C_c^2(\mathbb{R}^n)$, then,*

$$\mathcal{A}f(X_s) = \left(\sum_i b^i(x)\frac{\partial f}{\partial x^i}(X_s) + \frac{1}{2}\sum_{i,j}(\sigma\sigma^T)_{ij}(s,\omega)\frac{\partial^2 f}{\partial x^i \partial x^j}(X_s)\right)$$

A very useful result follows.

Theorem 11.2.7 (Dynkin's Formula) *Let $f \in C_c^2(\mathbb{R}^n)$ and τ be a stopping time with $E_x[\tau] < \infty$. Then,*

$$E_x[f(X_\tau)] = f(x) + E_x[\int_0^\tau \mathcal{A}f(X_s)ds]$$

Remark 11.9. The conditions for Dynkin's Formula can be generalized. As in Theorem 4.1.5, if the stopping time τ is bounded by a fixed constant, the conditions on f can be relaxed. Furthermore if τ is the exit time from a bounded set, then it suffices that the function is C^2 (and does not necessarily have compact support).

Remark 11.10. [Martingale characterization of weak solutions] Consider a stochastic differential equation:$dX_t = b(X_t)dt + \sigma(X_t)dB_t$. As we studied earlier, a probability measure P on the sample path space (or its stochastic realization X_t) is said to be a weak solution if under P

$$f(X_t) - \int_0^t \mathcal{A}f(X_s)ds \tag{11.8}$$

is a martingale with respect to $\mathcal{M}_t = \sigma(X_s, s \leq t)$, for any C^2 function f with bounded first and second order partial derivatives. As noted earlier, every strong solution is a weak solution, but not every weak solution is a strong solution; every such P admits a stochastic realization [181] but the stochastic realization may not be defined on the original probability space as a measurable function of the original Brownian motion. For example, if X_t can be defined to be randomized, where the randomization variables are independent noise processes, one could embed the noise terms into a larger filtration; this will lead to a weak solution but not a strong solution since there is additional information required (that is not contained in the original Brownian process).

11.2.6 Fokker-Planck equation

The discussion on the infinitesimal generator function (and Dynkin's formula) suggests that one can compute the evolution of the probability measure $\mu_t(\cdot) = P(X_t \in \cdot)$, by considering for a sufficiently rich class of functions $f \in \mathcal{D}$

$$E[f(X_t)] = \int \mu_t(dx) f(x).$$

Note that continuous and bounded functions are measure determining (as discussed in the proof of Theorem 10.8, see [42, p. 13] or [120, Theorem 3.4.5]) and since smooth signals are dense among such functions, we can take f to be smooth. Suppose that we assume that μ_t admits a density function and this is denoted by the same letter. Furthermore, let $p(x,t) := \mu_t(x)$. By taking $\mathcal{D}$ to be the space of smooth signals with compact support, which is a dense subset of the space of square integrable functions on $\mathbb{R}$, using the expectation of the infinitesimal generator function equation (11.8), writing

$$\frac{d}{dt}\int \mu_t(dx) f(x) = \frac{d}{dt}E[f(X_t)] = \frac{d}{dt}E[\int_0^t \mathcal{A}f(X_s)ds] = \int \mu_t(dx)\left(\frac{df}{dx}b(x) + \frac{1}{2}\frac{\partial^2 f(x)}{\partial x^2}\sigma^2(x)\right)$$

and applying integration by parts (twice for the term on the right), we obtain that for a process of the form

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t \tag{11.9}$$

the following holds:

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}(b(x)p(x,t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2(x)p(x,t)) \tag{11.10}$$

This is the celebrated *Fokker-Planck equation*. Notably, if there exists a stationary measure p, the time-independence on the right hand side will lead to an ODE for this stationary measure.

The Fokker-Planck equation is a partial differential equation whose existence for a solution requires certain technical conditions. As we discussed earlier, this is related to having a weak solution to a stochastic differential equation and in fact they typically imply one another. Of course, the Fokker-Planck equation may admit a density as a solution, but it may also admit a solution in a further weaker sense in that the evolution of the solution measure $P(X_t \in \cdot)$ may not admit a probability density function.

11.2.7 Rough Integration [Optional]

We end this section with a brief reflection on the limitations of the integrations noted above. From the way we have constructed the Itô integral is that the integral is constructed as an L_2 limit of approximations. In particular, (i) the integral is not defined in a sample path sense (and typically only would allow for convergence in probability and thus almost sure convergence along a subsequence, though this does occur also in a sample path almost sure sense under additional conditions on the regularity of the integrand; see e.g. [339] or [312, p. 91]), and (ii) it is not continuous with respect to the driving noise. A pathwise theory of solutions to differential equations would thus be a natural goal to arrive at; and this is attained by what is known as *rough integration*. The fundamental insight of rough paths theory is that the issue of defining solutions to differential equations driven by an irregular signal $X = (X^1, ..., X^d)$ can be reduced to defining iterated integrals of the form $\int_s^t (X^i(r) - X^i(s))dX^j(r)$. Rough paths theory allows for Hölder continuous driving signals - or in the case of stochastic differential equations, stochastic processes that are almost surely Hölder. Recall the definition of Hölder continuity: Define for $\alpha \in (0,1]$ the space $C^\alpha([0,T],\mathbb{R}^d)$ of α-Hölder functions $f : [0,T] \to \mathbb{R}^d$ equipped with the norm $\|f\|_\alpha := \sup_{t \neq s} \frac{|f(t)-f(s)|}{|t-s|^\alpha}$. Now, if X is a signal that is α-Hölder continuous with $\alpha \in (1/3, 1/2]$ and F is a smooth function, then for a partition of $[0,t]$, $\mathcal{P} = \{0 = t_0 < ... < t_n = t\}$ we have that in the integral

$$\int_0^t F(X(r))dX(r) = \sum_{k=0}^n \int_{t_k}^{t_{k+1}} F(X(r))dX(r)$$

$$= \sum_{k=0}^{n} \int_{t_k}^{t_{k+1}} F(X(t_k)) + F'(X(t_k))(X(r) - X(t_k)) + O(|r - t_k|^{2\alpha}) dX(r)$$
$$= \sum_{k=0}^{n} \Bigg(F(X(t_k))(X(t_{k+1}) - X(t_k)) + F'(X(t_k)) \int_{t_k}^{t_{k+1}} (X(r) - X(t_k)) dX(r)$$
$$+ O(|t_{k+1} - t_k|^{3\alpha}) \Bigg), \tag{11.11}$$

as $3\alpha > 1$, the remainder term should go to 0. This reduces the problem of defining the integral $\int_0^t F(X(r))dX(r)$ to just defining $\int_{t_k}^{t_{k+1}} (X(r) - X(t_k))dX(r)$. We take the right hand side as a *definition* of the left hand side, so long as we define the iterated integral first. However, if X is irregular then the iterated integral does not exist as a Riemann-Stieltjes integral and therefore must be defined. This leads to a new construction, called the rough integral. A rough path above a signal X is a pair $\mathbf{X}_{s,t} = (X_{s,t}, \mathbb{X}_{s,t})$ where $X_{s,t}$ is the increment of X and $\mathbb{X}_{s,t}$ is a *definition* or *postulation* of the iterated integral $\int_s^t (X(r) - X(s))dX(r)$ [137]. Then, one constructs the definition in a sense that it is compatible with the usual integration notions. For example, if $X \in C^1$ and $\int_s^t (X(r) - X(s))dX(r)$ is the Riemann-Stieltjes integral, $\mathbf{X}_{s,t} := (X(t) - X(s), \int_s^t (X(r) - X(s))dX(r))$ is consistent with such a postulation. The main utility of rough integration is that, by imposing conditions on the rough integration, solutions to equations of the form [137, Theorem 4.10]

$$dY^1 = b_1(Y^1)dt + \sigma(Y^1)d\mathbf{X}^1,$$

is continuous in the driving noise. It should be noted that the conditions impose that the rough integral definition itself is continuous in the driving noise.

11.3 Controlled Stochastic Differential Equations and the Hamilton-Jacobi-Bellman Equation

11.3.1 Revisiting the deterministic optimal control problem in continuous-time

Consider

$$\frac{dx}{dt} = f(x, u), \qquad x(0) = x$$

and suppose that the goal is to minimize

$$J(\gamma) = \int_{t_0}^{T} c(s, x(s), u(s))ds + P(x(T))$$

over all feedback control policies γ, where we take c to be continuous and bounded. Via Bellman's principle, as in Theorem 5.1.3, we define value functions:

$$V(t, x) = \inf_{\gamma} \int_{t}^{T} c(s, x(s), u(s))ds + P(x(T))$$

with the terminal condition

$$V(T, x(T)) = P(x(T))$$

Remark 11.11. For the existence of an optimal policy, very mild conditions can be arrived at via the theory of Young measures: See Section 11.5.1 for a detailed analysis leading to general existence conditions.

In the following, we first present an informal and non-rigorous derivation for an optimality equation, but the optimality analysis will be rigorously justified in Theorem 11.3.1. Applying Bellman's principle from the theory studied earlier, for a policy to be optimal it looks reasonable to arrive at the following (which will be justified shortly):

$$V(t,x) = \inf_{\gamma} \int_t^{t+h} c(s,x(s),u(s))ds + V(t+h,x(t+h)) \tag{11.12}$$

or

$$0 = \left(\inf_{\gamma} \int_t^{t+h} c(s,x(s),u(s))ds + V(t+h,x(t+h)) \right) - V(t,x) \tag{11.13}$$

for all $h \geq 0$. Consider then:

$$0 = \lim_{h \to 0} \frac{\left(\inf_{\gamma} \int_t^{t+h} c(s,x(s),u(s))ds + V(t+h,x(t+h)) \right) - V(t,x)}{h}$$

Now, for h small, we have that $x(t+h) = x(t) + f(x(t),u(t))h + o(h)$, where $o(h)/h \to 0$ as $h \to 0$. If we assume that V is continuously differentiable in its entries, we then have

$$V(t+h,x(t+h)) = V(t,x(t)) + V_t(x,t)h + (V_x(t,x) \cdot f(x,u))h + o(h),$$

leading to

$$0 = \lim_{h \to 0} \inf_{\gamma} \frac{\left\{ \int_t^{t+h} c(s,x(s),u(s))ds + V_t(t,x)h + (V_x(t,x) \cdot f(x,u))h + o(h) \right\}}{h}$$

Assuming that $o(h)/h \to 0$ uniformly for all control policies and that $c(s,x(s),u(s))$ is continuous in s (which is clearly not justified for an arbitrary policy!), we arrive at

$$0 = \inf_{\gamma} \left\{ c(t,x(t),u(t)) + V_t(t,x) + (V_x(t,x) \cdot f(x,u)) \right\}$$

In a more standard form, this leads to, provided the minimum exists,

$$-V_t(t,x) = \min_{u(t)} \left(c(t,x,u(t)) + (V_x(t,x) \cdot f(x,u(t))) \right) \tag{11.14}$$

This is the celebrated HJB (Hamilton-Jacobi-Bellman) equation. This defines a partial differential equation with boundary condition $V(T,x) = P(x)$ or $V(T,x(T)) = P(x(T))$.

The above analysis has several gaps: we imposed the value functions to be so that local linearized approximations could be made and some uniformity assumptions were not even justified. Nonetheless, as is often the case in applied mathematics, heuristic reasoning may lead to important equations whose validity however then needs to be rigorously justified. In particular, the above leads to an important equation which is a surprisingly strong result, as established in the following verification theorem:

Theorem 11.3.1 *[Optimality of HJB Solutions] Let $V(t,x)$ be C^1 (i.e., continuously differentiable) in both t and x, and solve the HJB equation (11.14). Suppose further that the policy γ satisfies (11.14) with $u(t) = \gamma(t)$. Then, γ is optimal.*

Proof. Let $V(t,x)$ be C^1 in both entries. Consider any admissible policy γ, which (under any history dependent measurable policy) can be viewed to be a function of time without any loss since the problem is deterministic. Let the HJB equation hold:

$$\frac{\partial V}{\partial t}(t,x) + \min_{u} \left(\frac{\partial V}{\partial x}(t,x) \cdot f(x,u) + c(t,x,u) \right) = 0, \quad V(T,x) = P(x_T),$$

and thus, for any policy with $u_t = \gamma(t)$:

$$\frac{\partial V}{\partial t}(t,x) + \left(\frac{\partial V}{\partial x}(t,x) \cdot f(x,\gamma(t)) + c(t,x,\gamma(t)) \right) \geq 0, \quad V(T,x) = P(x_T),$$

and thus, for any policy γ (which is open-loop without any loss in optimality for deterministic systems)

$$\frac{\partial V}{\partial t}(t,x) + \left(\frac{\partial V}{\partial x}(t,x) \cdot f(x,\gamma(t)) + c(t,x,\gamma(t))\right) \geq 0.$$

Now, consider $V(t,x_t^\gamma)$ where γ denotes the explicit dependence on the policy. We have that

$$\frac{dV(t,x_t^\gamma)}{dt} = \frac{\partial V}{\partial t}(t,x_t^\gamma) + \frac{\partial V}{\partial x}(t,x) \cdot f(x,u)|_{x=x_t^\gamma, u=\gamma(t)}$$

By the above, we have then

$$-(\frac{\partial V(t,x_t^\gamma)}{\partial t} + \frac{\partial V}{\partial x}(t,x) \cdot f(x,u)|_{x=x_t^\gamma, u=\gamma(t)}) \leq c(x,\gamma(t))$$

and

$$-\frac{dV(t,x_t^\gamma)}{dt} \leq c(x,\gamma(t)) \tag{11.15}$$

Taking the integral and noting that this holds for any γ, we arrive at

$$V(0,x_0) \leq \int_0^T c(x_t,\gamma(t))dt + V(T,x_T^\gamma) = \int_0^T c(x_t,\gamma(t))dt + P(x_T^\gamma)$$

Note that the initial value is independent of control and hence $V(0,x_0)$ is a lower bound for any control. Equality holds if the HJB is satisfied by some admissible control policy, which would then be optimal. ◇

Remark 11.12. For some generalizations on HJB and optimal control:

(i) One can relax the regularity conditions on V so that V may not be differentiable everywhere (leading to solution concepts such as viscosity solutions). An intuitive way to appreciate viscosity solutions is to consider the verification theorem above and replace V at a neighborhood of a point x where V is not differentiable with a continuously differentiable function ϕ which satisfies two properties: With $V(t,x) = \phi(t,x)$, if $V(s,y) - \phi(s,y)$ has its local minimum at (t,x), then a version of (11.14) holds with

$$-\phi_t(t,x) \leq \min_{u(t)} \left(c(t,x,u(t)) + (\phi_x(t,x) \cdot f(x,u(t)))\right),$$

which then leads to $\phi(t,x) - \phi(s,y) \geq V(t,x) - V(s,y)$ for (s,y) in a neighborhood of (t,x). Noting $(\phi_x(t,x) \cdot f(x,u(t))$ as the partial derivative of $\phi(t,\cdot)$, and moving it to the left hand side, by taking $s = t+h, y = x_{t+h}$, we arrive at (11.15) and this leads to the lower bound property of V. For the other direction, if $V(s,y) - \phi(s,y)$ has its local maximum at (t,x), we have that $\phi(t,x) - \phi(t,y) \leq V(t,x) - V(t,y)$ in a neighborhood of (t,x) and in the equation

$$-\phi_t(t,x) \geq \min_{u(t)} \left(c(t,x,u(t)) + (\phi_x(t,x) \cdot f(x,u(t)))\right),$$

fix the control policy attaining the minimum, and then arrive that

$$V(0,x_0) \geq \int_0^T c(x_t,\gamma(t))dt + P(x_T^\gamma)$$

Note the parallels in the argumentation, in terms of lower bounds and the attainability, with that in Theorem 5.5.3 in the order of the inequalities, as well as with the ACOI in Theorem 7.1.3.

(ii) Instead of sufficiency, one can arrive at necessary conditions via what is known as the *maximum principle* via variational local optimality conditions. We refer the reader to [218] for a rather comprehensive and accessible discussion and [131] for the stochastic setup.

11.3.2 The stochastic case and classes of admissible policies

Suppose now that we have a controlled system:

$$dX_t = b(t, X_t, u_t)dt + \sigma(t, X_t)dB_t,$$

where $u_t \in \mathbb{U}$ is the control action variable. We assume that u_t is measurable at least on $\mathcal{F}_t$ (but we can restrict this further so that it is measurable on a strictly smaller sigma field). Thus, the differential equation is well defined as an Itô integral. We will assume that a solution exists.

Often, one has a time-homogenous diffusion model, which is more suitable for infinite horizon analysis.

$$dX_t = b(X_t, U_t)dt + \sigma(X_t)dB_t \tag{11.16}$$

Control policies and existence of solutions

As we discussed extensively throughout the notes, the selection of the control actions need to be measurable with respect to some information at the controller and this dependency leads to fundamental differences on the behaviour of solutions and optimization methods.

(i) If for every t, u_t is measurable on the filtration generated by X_t, then the policy is called admissible.

(ii) If the control at time t is only a function of X_t and t, then the policy is called a Markov policy. If it only depends on X_t, then it is stationary. Randomization also is possible, but this requires a more careful treatment when compared with the discrete-time counterpart [15].

(iii) One often considers *adapted open-loop* policies; these are policies which are measurable with respect to $\sigma(X_0, (B_s, s \leq t))$ at time t.

(iv) One further relaxes these with *non-anticipative policies*; these are policies which satisfy the condition that $X_0, (U_s, B_s)$ is independent of $B_t - B_s$ for every $t > 0$.

The above entail subtle distinctions on whether they lead to strong solutions: To ensure existence and uniqueness of strong solutions, consider the following assumptions on the drift b and the diffusion matrix σ .

(A1) *Lipschitz continuity:* The functions σ, b are Lipschitz continuous in x (uniformly with respect to the other variables for b). In other words, for some constant $C > 0$

$$|b(x, \zeta) - b(y, \zeta)|^2 + \|\sigma(x) - \sigma(y)\|^2 \leq C|x - y|^2$$

for all x, y and $\zeta \in \mathbb{U}$, where $\|\cdot\|$is an appropriate metric on matrices such as $\|\sigma(x)\| = \sqrt{\text{Trace}\sigma(x)\sigma^T(x)}$. Furthermore, b is jointly continuous in (x, ζ).

(A2) *Affine growth condition:* b and σ satisfy a global growth condition of the form

$$\sup_{\zeta \in \mathbb{U}} |\langle b(x, \zeta), x\rangle| + \|\sigma(x)\|^2 \leq C_0(1 + |x|^2) \qquad \forall\, x$$

for some constant $C_0 > 0$.

Recall that in Theorem 11.2.3 we had noted that for a control-free stochastic differential equation, if b and σ satisfy Lipschitz regularity and growth properties, then a strong solution exists. However, if we only restrict that the control policy is measurable, with no additional assumptions, it is not guaranteed that a strong solution for X_t would exist. In view of this note, we state the following on the existence of strong or weak solutions under various classes of control policies:

(i) If b and σ satisfy similar regularity conditions, as in Theorem 11.2.3 with uniformity over control actions, and the control policies are non-anticipative, then the existence of strong solutions under the hypotheses above follows

from a similar argument [15, Theorem 2.2.4] (by viewing the control as an exogenous process). We also note that the above global Lipschitz constant C can be relaxed to a local one, that is one may have for all $R > 0$, $|b(x,\zeta) - b(y,\zeta)|^2 + \|\sigma(x) - \sigma(y)\|^2 \leq C_R|x-y|^2$ for all $\|x\|, \|y\| \leq R$; here one considers solutions up to an exit time, establishes a uniqueness result and takes take the exit set boundary to infinity.

(ii) If admissible (or feedback policies; that is, policies which are $\sigma(X_{[0,t]})$-measurable) are considered, one can apply measure transformation (via Girsanov's method), establish a strong solution for the control-free term $dX_t = \sigma(X_t)dB_t$, and then construct a solution under the original space; which then leads to a weak solution for the original space. Accordingly, under the hypothesis (A2) together with a Lipschitz condition as in (A1) but only for σ, for any admissible control (11.16) has a unique weak solution [15, Theorem 2.2.11].

(iii)Furthermore, under hypotheses (A1)–(A2) and under any stationary Markov strategy there is a unique strong solution which is a strong Feller process [15, Theorem 2.2.12].

(iv)Finally, if σ were allowed to also depend on control in general, existence in this setup is a non-trivial problem, since measure change arguments cannot be directly applied when the control is present in the diffusion term. However, if non-anticipative policies are considered, then under strict Lipschitz and growth conditions, a strong solution exists. We refer the reader to [15, Section 2] as well as [33, 95, 96] for a detailed analysis and literature review. Notably, if we replace $\sigma(x)$ by $\sigma(x,\zeta)$, if $\sigma(\cdot, v(\cdot))$ is Lipschitz continuous for stationary v then there is a unique strong solution. But in general stationary policies are just measurable functions, and existence of suitable strong solutions is more delicate (see, [15, Remarks 2.3.2]).

In the following, we will consider admissible policies and the solution concept will be taken to be of weak solutions so that the expected cost criteria are well-defined and the limitations on having strong solutions are not presented apriori. Nonetheless, we will observe that under verification theorems optimal policies (even among those which are admissible) will be Markov and the issues on existence of strong solutions will not be as critical, see the discussion above.

The reader is encouraged to revisit the verification theorems for discrete-time problems: These are Theorem 5.1.3 for finite horizon problems, Theorem 5.5.3 for discounted cost problems, and Theorem 7.1.1 for average cost problems. One can see that the essential difference is to express the expectations through Dynkin's formula and the differential generators.

Finite Horizon Cost Criterion

Suppose that given a control policy, the goal is to minimize

$$E[\int_0^T c(s, X_s, u_s)ds + c_T(X_T)],$$

where c_T is some terminal cost function.

As in *Chapter 5*, if a policy is optimal, we will arrive at the following equation for every possible state:

$$V(r, X_r) = \min_{\gamma} E[\int_r^t c(s, X_s, u_s)ds + V(t, X_t)|X_r].$$

In the following, following the same flow of ideas as in the deterministic case in Section 11.3.1, we first provide a rather informal derivation of the optimality equation, but the formal verification result will be precise. We assume that $V(s,x)$ is C^2 in x and C^1 in s. Then,

$$0 = \min_u \left(E[\int_r^t c(s, X_s, u_s)ds + V(t, X_t)|X_r] - V(r, X_r) \right)$$

In particular,

$$0 = \lim_{h \to 0} \min_u \left(\frac{E[\int_r^{r+h} c(s, X_s, u_s)ds + V(r+h, X_{r+h})|X_r] - V(r, X_r)}{h} \right)$$

Now, if V is so that it is in the domain of the generator for every control policy, with

$$\mathcal{L}_t^u V(t,x) = \sum_i b^i(t,x,u)\frac{\partial V}{\partial x^i}(t,x) + \frac{1}{2}\sum_{i,j}\sigma^i(t,x)\sigma^j(t,x)\frac{\partial^2 V}{\partial x^i \partial x^j}(t,x)$$

applying the mean-value theorem, assuming it would hold for now, we arrive at

$$\min_u \left(c(s,x,u) + \mathcal{L}_s^u V(s,x) + \frac{\partial V}{\partial s}(s,x) \right) = 0 \tag{11.17}$$

Thus, if a policy is optimal, it needs to satisfy the above property provided that V satisfies the necessary regularity conditions under the considered set of policies to validate the operations above and indeed the above would also be sufficient for optimality by the analysis to follow. However, as in the deterministic case, the analysis above is informal and we have not presented precise conditions under which the above would hold. As we have seen before in the earlier chapters, verification theorems show that a policy that satisfies the verification is optimal over all admissible policies:

Theorem 11.3.2 (Verification Theorem) *Consider: $dX_t = b(t,X_t,u_t)dt + \sigma(t,X_t,u_t)dB_t$. Suppose that V is C^2 in x and C^1 in t is so that:*

$$\frac{\partial V}{\partial t}(t,x) + \min_{u\in\mathbb{U}} \left(\mathcal{L}_t^u V(t,x) + c(t,x,u) \right) = 0, \quad V(T,x) = c_T(x), \tag{11.18}$$

Then, an admissible control policy which achieves the minimum for every (t,x) is optimal.

Proof.

The equation $\frac{\partial V}{\partial t}(t,x) + \min_u\{\mathcal{L}^u V(t,x) + c(t,x,u)\} = 0$ implies that for any admissible control realization:

$$c(t,x,u) \geq -\frac{\partial}{\partial t}V(t,x) - \mathcal{L}^u V(t,x), \tag{11.19}$$

and as in the deterministic case (in Theorem 11.3.1), for any admissible control policy γ

$$E^\gamma[\int_0^T \frac{-\partial}{\partial s}V(s,X_s^u) - \mathcal{L}_s^u V(s,X_s^u)] \leq E^\gamma[\int_0^T c(s,X_s^u,u_s)].$$

Using Itô's rule,

$$E[V(0,X_0)] = E[\int_0^T (\frac{-\partial V}{\partial s}(s,X_s^\gamma) - \mathcal{L}_s^\gamma V(s,X_s^\gamma))ds] + E[V(T,X_T^\gamma)],$$

and thus, we obtain that for any admissible control

$$E[V(0,X_0)] \leq E[\int_0^T c(s,X_s^\gamma,u_s)ds + c_T(X_T^\gamma)].$$

On the other hand, a policy γ^* which satisfies the equality in (11.18), leads to an equality in the above and optimality. $\diamond$

Example 11.13. [Optimal portfolio selection] We consider a continuous-time version of a problem considered in Exercise 2.9.2: A common example in finance applications is the portfolio selection problem where a controller (investor) would like to optimally allocate his wealth between a stochastic stock market and a market with a guaranteed income (see [312]): Consider a stock with an average return $\mu > 0$ and volatility $\sigma > 0$ and a bank account with interest rate $r > 0$. These are modeled by:

$$\begin{aligned} dS_t &= \mu S_t dt + \sigma S_t dB_t, \quad S_0 = 1 \\ dR_t &= rR_t dt, \quad R_0 = 1 \end{aligned} \tag{11.20}$$

Suppose that the investor can only use his own money to invest and let $u_t \in [0,1]$ denote the proportion of the money that he invests in the stock. This implies that at any given time, his wealth dynamics is given by:

$$dX_t = \mu u X_t dt + \sigma u X_t dW_t + r(1-u_t)X_t dt,$$

or $dX_t = \Big(\mu u + r(1-u)\Big)X_t dt + \sigma u X_t dB_t$. Suppose that the goal is to maximize $E[\log(X_T)]$ for a fixed time T (or minimize $-E[\log(X_T)]$). In this case, the Bellman equation writes as:

$$0 = \frac{\partial}{\partial t}V(t,x) + \min_u \left((\mu u + r(1-u))x \frac{\partial}{\partial x}V(t,x) + \sigma^2 u^2 x^2 \frac{1}{2}\frac{\partial^2}{\partial x^2}V(t,x) \right),$$

with $V(T,x) = -\log(x)$. With a guess of the value function of the form $V(t,x) = -\log(x) + b_t$, one obtains an ordinary differential equation for b_t with terminal condition $b_T = 0$. It follows that, if $\frac{\mu - r}{\sigma^2} \in [0,1]$ the optimal control is $u_t(x) = \frac{\mu - r}{\sigma^2}$, leading to $V(t,x) = -\log(x_t) - C(T-t))$, for some constant C.

Example 11.14. [The Linear Quadratic Regulator and Continuous-Time Riccati Equation] Consider a continuous-time counterpart of the LQG problem studied in Section 5.3: Let

$$dx_t = Ax_t dt + Bu_t dt + D dB_t,$$

where x_t, u_t are all $\mathbb{R}$-valued, with the goal of minimizing:

$$J(x,\gamma) = E_x^\gamma[(\int_0^N Qx_t^2 + Ru_t^2) + Q_N x_N^2],$$

where $R > 0, Q_N > 0, Q \geq 0$. By the HJB equation, we have

$$0 = \frac{\partial}{\partial t}V(t,x) + \min_{u\in\mathbb{R}} \left(\frac{\partial}{\partial x}V(t,x)(Ax+Bu) + \frac{1}{2}\frac{\partial^2}{\partial x^2}V(t,x)D^2 + Qx^2 + Ru^2 \right)$$

Taking $V(t,x) = P_t x^2 + K_t$ with $K_N = 0, P_N = Q_N$, we then have

$$-P_t' x^2 - K_t' = \min_{u\in\mathbb{R}} \left(2P_t x(Ax+Bu) + P_t D^2 + Qx^2 + Ru^2 \right) = \min_{u\in\mathbb{R}} \left(2AP_t x^2 + 2P_t Bxu + Ru^2 + P_t D^2 + Qx^2 \right)$$

By completion of squares: $2P_t Bxu + Ru^2 = (\sqrt{R}u + \frac{1}{\sqrt{R}}P_t Bx)^2 - \frac{1}{R}P_t^2 B^2 x^2$, we have that the optimal control is

$$u_t = -\frac{1}{R}P_t B x_t$$

where

$$-P_t' x^2 - K_t' = (2AP_t + Q - \frac{1}{R}P_t^2 B^2)x^2 + P_t D^2.$$

Thus, we arrive at (the continuous-time Riccati equation)

$$-P_t' = 2AP_t + Q - \frac{1}{R}P_t^2 B^2, \quad P_N = Q_N$$

and

$$K_t' = -P_t D^2, \quad K_N = 0.$$

11.3.3 Discounted Infinite Horizon Cost Criterion

Suppose that given a control policy, the goal is to minimize

$$E^{\gamma}[\int_0 e^{-\lambda s} c(X_s, u_s) ds].$$

In this section, we will consider a time-homogenous setup

$$dX_t = b(X_t, u_t)dt + \sigma(X_t, u_t)dB_t, \tag{11.21}$$

and let us define

$$\mathcal{L}^u g(x) = \sum_i b^i(x,u)\frac{\partial g}{\partial x^i}(x) + \frac{1}{2}\sum_{i,j}\sigma^i(x,u)\sigma^j(x,u)\frac{\partial^2 g}{\partial x^i \partial x^j}(x) \tag{11.22}$$

In this case, we have the following result:

Theorem 11.3.3 (Verification Theorem) *Suppose that V is C^2 and $\lim_{t\to\infty} e^{-\lambda t}E_x^{\gamma}[V(X_t^u)] = 0$ under any admissible policy γ and x. Let*

$$\min_{u\in\mathbb{U}}\Big(\mathcal{L}^u V(x) - \lambda V(x) + c(x,u)\Big) = 0. \tag{11.23}$$

Then, an admissible control policy which achieves the minimum above for every x is optimal.

Proof. Under any admissible policy γ and its control realization u_t at time t, we have that $\mathcal{L}^{\gamma}V(x_t) - \lambda V(x_t) + c(x_t, u_t) \geq 0$. This then leads to

$$e^{-\lambda t}c(x_t, u_t) \geq e^{-\lambda t}(-\mathcal{L}^{\gamma}V(x_t) + \lambda V(x_t))$$

Using Itô's rule for $V(X_t)e^{-\lambda t}$ one obtains:

$$E^{\gamma}[V(X_0)] - e^{-\lambda t}E[V(X_t^{\gamma})] = E^{\gamma}[\int_0^t e^{-\lambda s}(-\mathcal{L}^{\gamma}V(X_s) + \lambda V(X_s))ds] \leq E^{\gamma}[\int_0^t e^{-\lambda s}c(x_s, u_s)ds]$$

Taking $t \to \infty$, we show that $V(0)$ is a lower bound. Proceeding as before in the proof of Theorem 11.3.2, under equality, we have that the lower bound is attained. ◇

The above is a sufficiency analysis. One could say more with regard to necessity as well [15, Theorem 3.5.6] via an analysis based on the theory of partial differential equations: In addition to mild growth conditions [15, Section 2.2], if (i) $c : \mathbb{R}^d \times \mathbb{U} \to \mathbb{R}_+$ is continuous and locally Lipschitz in x uniform in u, and (ii) if $\sigma(x)$, not depending on u, is locally uniformly elliptic, i.e., $\sigma(x)\sigma^T(x)$ has its eigenvalues locally bounded from below, then (11.23) admits a unique solution, which is bounded, and which serves as the solution to the optimal cost. Thus, one has a complete characterization of optimality under additional regularity conditions. These also carry over to the average-cost setup presented in the following.

11.3.4 Average-Cost Infinite Horizon Cost Criterion

Suppose that given a control policy, the goal is to minimize

$$\limsup_{T\to\infty}\frac{1}{T}E^{\gamma}[\int_0^T c(X_s, u_s)ds].$$

Once again here we consider a time-homogenous setup with the controlled equation (11.21) and generator (11.22).

Theorem 11.3.4 (Verification Theorem) *Suppose that V is C^2 and $\eta \in \mathbb{R}$ so that*

$$\min_{u\in\mathbb{U}}\Big(\mathcal{L}^u V(x) - \eta + c(x,u)\Big) = 0, \tag{11.24}$$

and

$$\limsup_{T\to\infty} \frac{E_x^\gamma[V(X_0) - V(X_T^\gamma)]}{T} = 0, \tag{11.25}$$

under every admissible policy γ and initial state x. Then, an admissible control policy which achieves the minimum for every x is optimal.

Proof. Once again for any admissible control policy γ, $\mathcal{L}^\gamma V(x_t^\gamma) - \eta + c(x_t^\gamma, u_t) \geq 0$, leading to $c(x_t^\gamma, u_t) \geq \eta - \mathcal{L}^\gamma V(x_t^\gamma)$. Then, one obtains

$$\int_0^T c(x_t^\gamma, u_t)dt \geq \eta T - \int_0^T \mathcal{L}^\gamma V(x_t^\gamma dt.$$

Dividing by T and taking the limit superior, via (11.25), we have the lower bound. The achievability holds under equality (11.26). ◇

The convex analytic method

The analysis we made in *Chapter 5* and *7* applies to the diffusion setting as well. In particular, a discounted HJB equation plays the role of the discounted cost optimality equation. For the average cost problems, one can apply either a *vanishing discount* approach or an *convex-analytic* approach. We refer the reader to [15], [61, 66] and [295].

11.3.5 Control up to an Exit Time

In some applications, one studies cost criteria of the type:

$$E^\gamma[\int_0^\tau c(X_s, u_s)ds + h(X_\tau)],$$

where $\tau = \inf\{t \geq 0 : X_t \notin S\}$ for some $S \in \mathbb{R}^n$ which is a set with a smooth boundary $\partial(S)$ and h is a terminal cost function.

We again consider the time-homogenous setup with controlled equation (11.21) and generator (11.22).

Theorem 11.3.5 (Verification Theorem) *Suppose that V is C^2 and $\eta \in \mathbb{R}$ so that*

$$\min_{u\in\mathbb{U}} \Big(\mathcal{L}^u V(x) + c(x,u)\Big) = 0, \quad x \in S; \qquad V(x) = h(x) \quad \text{on} \quad \partial(S). \tag{11.26}$$

Then, an admissible control policy which achieves the minimum for every x is optimal.

Remark 11.15. All the equations stated in the verification theorems noted above demonstrate the strict connections with the theory of partial differential equations. Indeed, existence and optimality results can be utilized to obtain direct and very strong results, see [15, Chapter 3]. The regularity conditions on the value function can also be relaxed.

11.4 Partially Observed Case, Girsanov's Theorem and Separated Policies

Consider a partially observed setup with

$$Y_t = \int h(X_s)ds + B_t \tag{11.27}$$

for some independent Brownian process B_t.

Such a setup leads to a number of technical difficulties. The analysis (especially for the case with measurements that are not linear and Gaussian) can be quite subtle due to the fact that the control policy (only restricted to be measurable in general) may lead to issues on the existence of strong solutions for a given stochastic differential equation since the control policy may couple the state dynamics with the past in an arbitrarily complicated, though measurable, way and hence violating the existence conditions for strong solutions to stochastic differential equations. Even for linear models, the analysis requires some careful reflection: Lindquist [219] provides a detailed account on this aspect and provides a general separation theorem provided that the control laws are among those which lead to the existence of a solution to the controlled stochastic differential system, generalizing e.g. the analysis in [206] (where control laws of the Lipschitz type in the conditional estimate are considered) and [334] (where Lipschitz control policies in $y_{[0,t]}$ are considered) to ensure the existence of strong solutions.

To avoid such technical issues on strong solutions, relaxed solution concepts were introduced and studied in the literature based on measure transformation due to Girsanov [33, 95, 96] (see Exercise 11.10.2 for a heuristic derivation). Now, consider the measurement model given in (11.27).

For a moment, suppose that $h \equiv 0$, that is Y_t is just an independent process. Suppose also that there is no control in the diffusion process $dX_t = b(X_t)dt + \sigma(X_t)dB'_t$. In this case, it is evident that the measurement process and the noise process are independent and let us call this probability measure on the processes as $\mathbf{Q} := P_X \times Q_Y$. In this case, consider for any measurable bounded f on the paths of $X_{[0,t]}$ and $Y_{[0,t]}$ as:

$$E_{\mathbf{Q}}[f(X_{[0,t]}, Y_{[0,t]})|Y_{[0,t]}] = \int f(x_{[0,t]}, Y_{[0,t]})P_X(dx_{[0,t]}),$$

since the measurement processes Y_t gives no information on the state process X_t. Thus, the computation is quite simple in this case.

Now, consider our original process given by (11.27) where h is non-zero. Let $\mathbf{P}$ be the joint probability measure on the state and the measurement processes. Since $\mathbf{P} \ll \mathbf{Q}$, we have that

$$\int f(x_{[0,t]}, y_{[0,t]})\mathbf{P}(dx_{[0,t]}, dy_{[0,t]}) = \int G(x_{[0,t]}, y_{[0,t]})f(x_{[0,t]}, y_{[0,t]})\mathbf{Q}(dx_{[0,t]}, dy_{[0,t]})$$

for some $\mathbf{Q}$-integrable function G, which is the Radon-Nikodym derivative of $\mathbf{P}$ with respect to $\mathbf{Q}$. It turns out that under mild conditions, we have that

$$G_t := G(x_{[0,t]}, y_{[0,t]}) = \frac{d\mathbf{P}}{d\mathbf{Q}} = e^{\int_0^t h(x_s)dy_s - \frac{1}{2}\int_0^t |h(x_s)|^2 ds}$$

with $\int_0^t h(x_s)dy_s$ being a stochastic integral, this time with respect to the random measure/process Y_s. This relation allows us to view the partially observed problem as one with independent measurements, with the dependence pushed to the Radon-Nikodym derivative, not unlike what was done in Section 10.4.2 (see also Exercise 11.10.2).

Therefore,

$$E_{\mathbf{P}}[f(X_{[0,t]}, Y_{[0,t]})|y_{[0,t]}] = \frac{\int_{x_{[0,t]}, y_{[0,t]}} f(x_{[0,t]}, y_{[0,t]})G(x_{[0,t]}, y_{[0,t]})\mathbf{Q}(dx_{[0,t]}, y_{[0,t]})}{\int_{x_{[0,t]}} G(x_{[0,t]}, y_{[0,t]})\mathbf{Q}(dx_{[0,t]}, y_{[0,t]})} \tag{11.28}$$

This equation is known as the (Kushner-)Kallianpur-Striebel formula. If we focus on the numerator and focus on X_t only, this is known as the unnormalized filter [204].

11.4.1 Non-linear filtering in continuous time and Zakai's equation

We will now study the evolution of the numerator above where we restrict f to be a function of the current state only. In particular, we will study the evolution of $\int \mu_t^{y_{[0,t]}}(dx)f(x) = E_{\mathbf{P}}[f(X_t), y_{[0,t]}]$, where the notation $[\cdot, y_{[0,t]}]$ means that we

restrict the measurements $y_{[0,t]}$ to be fixed (but we are not computing the conditional measure), as measurements are also collected. Compare this with the Fokker-Planck equation (11.10) in which case measurements do not exist.

Now, under $\mathbf{Q}$, we have that the X and the Y process are independent. Note now that we can write

$$E_{\mathbf{P}}[f(X_t), y_{[0,t]}] = E_{\mathbf{Q}}[f(X_t)G_t],$$

where G_t is as defined earlier in this section. This relation will make the analysis below relatively immediate, following the analysis of the Fokker-Planck equation (11.10). We will follow a similar reasoning, except now we will also consider the realizations of the measurements by considering G_t as a variable which adds a time dependence. We have then:

$$\frac{d}{dt}\int \mu_t^{y_{[0,t]}}(dx)f(x) = \frac{d}{dt}E_{\mathbf{P}}[f(X_t), y_{[0,t]}] = \frac{d}{dt}E_{\mathbf{Q}}[f(X_t)G_t]$$

and thus,

$$\frac{d}{dt}\int \mu_t^{y_{[0,t]}}(dx)f(x) = \int \mu_t^{\mathbf{Q}}(dx)\left(\frac{df}{dx}b(x)G_t + \frac{1}{2}\frac{\partial^2 f(x)}{\partial x^2}\sigma^2(x)G_t + \frac{dG_t}{dt}f\right)$$

As before, applying integration by parts (twice for the term in the middle), and then computing via Itô's formula the derivative of $\frac{dG_t}{dt} = G_t(hdY)$ (this can be obtained by writing $G_t = e^{Z_t}$ where Z_t solves $dZ_t = -\frac{1}{2}|h(x_t)|^2dt + h(x_t)dB_t$), and then writing

$$\mu_t^{y_{[0,t]}}(dx) = \mu_t^{\mathbf{Q}}(dx)G_t$$

we obtain that for a process of the form

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t \tag{11.29}$$

with measurements given in 11.27), the non-normalized filter density (provided a smooth one exists) $p^y(x,t)$ evolves as:

$$dp^y(x,t) = \left(-\frac{\partial}{\partial x}(b(x)p^y(x,t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2(x)p^y(x,t))\right)dt + p^y(x,t)hdy \tag{11.30}$$

If one normalizes the above, this time by conditioning on y (that is, by dividing the above with the expectation over the random measurements by integrating over all state values under the measure $\mathbf{P}$), one arrives at another important equation, known as the *Kushner-Stratonovich* equation. In particular, for a given function f

$$E[f(X_t)|Y_{[0,t]} = y_{[0,t]}] = \frac{\int p^y(x,t)f(x)dx}{\int p^y(x,t)dx}$$

11.5 Existence of Optimal Policies under Full, Partial and Decentralized Information

11.5.1 A related existence discussion in deterministic continuous-time

We first revisit a version of the deterministic optimal control problem considered in Section 11.3.1. It is instructive to discuss here various control topologies that are already well-known in classical control theory (when there is a single controller who has access to the state variable).

In the following, we build on [271]. In deterministic nonlinear, geometric, and continuous-time control, properties on stabilizability, controllability, and reachability are drastically impacted by the restrictions on the classes of allowed controls (e.g., continuous, Lipschitz, finitely differentiable, or smooth control functions in the state or time when control is open-loop [72, 180, 267, 291]) and naturally the control topology induced is dictated by the class of admissible controls. For optimal control, to allow for continuity/compactness arguments, apriori imposing compactness over spaces of measurable functions would be an artificial restriction, and the use of powerful theorems such as the Arzela-Ascoli theorem which necessarily entail (usually very restrictive and suboptimal) conditions on continuity properties of the considered policies.

In deterministic optimal control theory, relaxed controls [322, 340] allow for the mathematical analysis on continuity-compactness to be applied with no artificial restrictions on the classes of control policies considered. A particularly consequential approach is via the study of topologies on *Young measures* defined by randomized/relaxed controls [228,340], [75, Section 2.1], [322, p. 254], [224] where one views the topology on control policies to be identified with the weak convergence topology of a measure defined on a product space with a fixed marginal at an input/state space (typically the Lebesgue measure in optimal deterministic control).

Let us consider an open-loop controller, where the control is only a function of the time variable. We let $\nu(dt, du)$ be a measure on $[0, T] \times \mathbb{U}$ where the first marginal $\lambda(dt)$ is the normalized Lebesge measure on time interval $[0, T]$ and let $\nu(du|t) = 1_{\{\gamma(t) \in du\}}$ be the conditional measure induced by deterministic open loop control. So, any deterministic open-loop control is embedded via:

$$\nu(dt, du) = \lambda(dt)\, 1_{\{\gamma(t) \in du\}}.$$

If allow for randomized policies, we obtain the set $\mathcal{P}_\lambda([0, T] \times \mathbb{U})$ of all probability measures with fixed marginal on $[0, T]$. This set is weakly closed, whose extreme points are those induced by deterministic policies. Thus, any deterministic optimal control problem, which can be written in an integral form and have lower semi-continuous cost functions in actions, will have an optimal solution, which will then be deterministic as these form the extreme points of randomized controls. It can also in fact be shown that such policies are dense in the space of randomized policies, in addition to these policies forming the extreme points in the set of randomized policies (see e.g., [32, Proposition 2.2] [210], [238, 19, Theorem 3], but also many texts in optimal stochastic control where denseness of deterministic controls have been established inside the set of relaxed controls [54]). We refer the reader to [224] for further discussion.

The following example builds on these, with somewhat different arguments. Let $\mathbb{X} = \mathbb{R}, \mathbb{U} = [0, 1]$, and let $f : \mathbb{X} \times \mathbb{U} \to [0, 1]$ and $c : \mathbb{X} \times \mathbb{U} \to [0, 1]$ be measurable functions continuous in the control action variable. Consider the following optimal control problem:

$$\inf_{\substack{\gamma:\mathbb{X}\to\mathbb{U} \\ u_t=\gamma(x_t)}} \int_0^1 c(x_t, u_t)\, \lambda(dt) \tag{11.31}$$

subject to

$$\frac{dx}{dt} = f(x_t, u_t) \tag{11.32}$$

The natural space to consider is the set of all control functions which depends on the current state, where the only restriction is measurability. However, allowing for measurability only does not facilitate continuity/compactness arguments since, as noted above, imposing compactness on a space of functions is an unnecessarily restrictive condition. Accordingly, one often cites appropriate but tedious measurable selection theorems building on optimality equations through dynamic programming.

On the other hand, every deterministic function of state can be expressed as a deterministic function of time, and so, be considered open-loop. Accordingly, we consider open loop controls and those which are relaxed. Let $\mathcal{P}_\lambda([0, T] \times \mathbb{U})$ be the set of relaxed open loop policies (known as Young measures). Now consider the space $C([0, 1]; \mathbb{X}) \times \mathcal{P}_\lambda([0, T] \times \mathbb{U})$, where $C([0, 1]; \mathbb{X})$ is the space of continuous functions from $[0, 1]$ to $\mathbb{X}$. We endow this space with the product topology with the first component being under the supremum norm and the second under Prohorov metric (or any weak convergence inducing metric). Note now that the cost (11.35) is continuous on $C([0, 1]; \mathbb{X}) \times \mathcal{P}_\lambda([0, T] \times \mathbb{U})$. Note that since f is uniformly bounded, we have that the set $\mathcal{S}$ of all admissible sample paths of the state $x : [0, 1] \to \mathbb{X}$ is equicontinuous, and so, by the Arzela-Ascoli theorem, $\mathcal{S}$ is relatively compact in $C([0, 1]; \mathbb{X})$. Accordingly, our space of interest $\mathcal{S} \times \mathcal{P}_\lambda([0, T] \times \mathbb{U})$ is a relatively compact subset of $C([0, 1]; \mathbb{X}) \times \mathcal{P}_\lambda([0, T] \times \mathbb{U})$.

Define now

$$H = \left\{ (x, m) \in C([0, 1]; \mathbb{X}) \times \mathcal{P}_\lambda([0, T] \times \mathbb{U}) : x_t - \int_0^t f(x_s, u)\, m_s(du)\, \lambda(ds) = 0 \right\}, \tag{11.33}$$

where $m_s(du) = m(du|s)$. This set is closed under the topology defined on $C([0, 1]; \mathbb{X}) \times \mathcal{P}_\lambda([0, T] \times \mathbb{U})$ and is a subset of $C([0, 1]; \mathbb{X}) \times \mathcal{P}_\lambda([0, T] \times \mathbb{U})$. Hence, H is compact. Now, the problem then is to find an optimal $(x, m) \in H$ which

minimizes (11.35), reformulated as:

$$\inf_{(x,m)\in H} \int_0^1 c(x_t,u)\, m(dt,du)$$

This is continuous in (x,m) by an application of the generalized weak convergence theorem under continuous convergence [287, Theorem 3.5] or [212, Theorem 3.5]. Therefore, there exists an optimal solution to the problem.

11.5.2 Existence of Optimal Policies for Fully Observed Stochastic Models

We start the discussion with the fully observed model. We build on [203], [205]) and [258]

Consider a continuous-time process X_t taking values in a Euclidean space $\mathbb{R}^N$, controlled by a control process $\{U_t\}$ taking values in a compact metric space $\mathbb{U}$. In the context of a diffusion process,

$$dX_t = b(X_t,U_t)dt + \sigma(X_t)dW_t, \tag{11.34}$$

driven by standard Brownian motion $\{W_t\}$, under the control policy U and the initial condition $x \in \mathbb{R}^d$. We also allow the control policy to be randomized, that is $\mathcal{P}(\mathbb{U})$-valued, where $\mathcal{P}(\mathbb{U})$ denotes the space of probability measures on $\mathbb{U}$ under the weak convergence topology. An *admissible control* is a $\mathcal{P}(\mathbb{U})$ valued non-anticipative process $\{U_t\}$.

To ensure existence and uniqueness of weak solutions of (11.34), we impose the following assumptions on the drift b and the diffusion matrix σ.

(A1) The function b is jointly continuous in (x,u) and σ is locally Lipschitz continuous, i.e., for some constant $C_R > 0$ depending on $R > 0$, we have

$$\|\sigma(x_1) - \sigma(x_2)\|^2 \le C_R\, |x_1 - x_2|^2$$

for all $x_1, x_2 \in B_R$, where $\|\sigma\| := \sqrt{(\sigma\sigma^T)}$. Also, we assume that b, σ are uniformly bounded, i.e.,

$$\sup_{u\in\mathbb{U}} |b(x,u)| + \|\sigma(x)\| \le C \qquad \forall\, x \in \mathbb{R}^N,$$

for some constant $C > 0$.

(A2) For some $\hat{C}_1 > 0$, it holds that

$$\sum_{i,j=1}^{d} a^{ij}(x) z_i z_j \ge \hat{C}_1 |z|^2 \qquad \forall\, x \in \mathbb{R}^N,$$

and for all $z = (z_1,\ldots,z_d) \in \mathbb{R}^N$, where $a := \frac{1}{2}\sigma\sigma$.

In view of (A2), one sees that σ^{-1} exists and it is bounded. For similar existence/approximation results in the fully observable setup, the authors in [203, 205, 208], assumed that b, σ are bounded and uniformly Lipschitz.

In the following, first we will trace some of the ideas presented by Kushner (see e.g. [208]), though with some presentational differences and then present an alternative approach. Suppose that one wishes to minimize the cost

$$J(U) := E_x^U\left[\int_0^T c(X_s,U_s)ds + c_T(X_T)\right], \tag{11.35}$$

over all admissible control policies. Here, c, c_T are continuous and bounded functions. We define a relaxed wide-sense admissible control policy in the following. We first place the Young topology on the control action space, by viewing the progressively measurable random control process $m(dt,du)(\omega)$ to be a random probability measure on $[0,T]\times\mathbb{U}$ with its fixed marginal on $[0,T]$ to be the Lebesgue measure; here $\mathcal{P}_\lambda([0,T]\times\mathbb{U})$ (the space of such probability measures) is endowed with the weak convergence topology. We require that $m_{[0,t]}$ be independent of $B_s - B_t, s > t$ for every $t \in [0,T]$. We let $m \in \mathcal{P}(\mathcal{P}_\lambda([0,T]\times\mathbb{U}))$. We then consider the $C([0,T];\mathbb{R}^N)$-valued (under sup-norm) X_t process (solution to the diffusion equation (11.34)) induced by $m(dt,du)(\omega)$ and then consider the space of probability measures on these random variables.

From [15, Theorem 2.2.11], it is easy to see that under any choice of control process $m(dt, du)(\omega)$, (11.34) admits a unique weak solution .

Toward an existence and approximation analysis, we adopt the following two approaches.

Weak convergence approach without measure transformation

In one approach, presented extensively by several seminal studies by Kushner and collaborators [203, 205, 208] as well as others such as Borkar [55], one considers the following. Given the above, we consider

$$\begin{aligned} H = \Bigg\{ & (\eta, m) \in \mathcal{P}(C([0,T];\mathbb{R}^N)) \times \mathcal{P}(\mathcal{P}_\lambda([0,T]\times\mathbb{U})) : \\ & E_x\Big[f(X_t) - f(X_0) - \int_0^t \mathcal{A}^u f(X_s) m_s(du)\lambda(ds)\Big] = 0, \\ & m_{[0,t]} \text{ is independent of } W_s - W_t,\ s > t,\ s,t \in [0,T] \Bigg\}, \end{aligned} \tag{11.36}$$

for all twice continuously differentiable function f with compact support and where $m_s(du) = m(du|s)$ and

$$\mathcal{A}^u f(x) \coloneqq \big(a(x)\nabla^2 f(x)\big) + b(x,u)\cdot\nabla f(x)\,.$$

In the following theorem we show that the space H is closed, we follow Kushner's *weak convergence* approach (see e.g., [203, 205, 208]) but under weaker conditions. For detailed proof see [258, Theorem 2.1] .

Theorem 11.5.1 *[258] Suppose that Assumptions (A1)–(A2) hold. Then, if $m^n \to m$, with $(\eta_n, m_n) \in H$, the measure on the state process $\eta^n \to \eta$ which is the measure on the state process under m (that is, $(\eta, m) \in H$). Thus, H is closed.*

Now, the problem is to find an optimal $(\eta, m) \in H$ which minimizes (11.35). From this, one can, as in the deterministic case summarized in [271, Section 7.1], arrive at general conditions for the existence of an optimal solution. One can also establish conditions (see e.g., [203, Theorem 4.4]) for compactness for this set under the weak topology. We can write the cost as

$$J(m) = E\bigg[\Big(\int_0^T \int_{\mathbb{U}} c(X_s, u) m_s(\omega)(du)\Big) + c_T(X_T)\bigg]\,. \tag{11.37}$$

Theorem 11.5.2 *[258] Suppose that Assumptions (A1)–(A2) hold. Then,*

$$J : \mathcal{P}(\mathcal{P}_\lambda([0,T]\times\mathbb{U})) \to \mathbb{R}$$

is a continuous map.

Next, using the above continuity result we want to prove the near optimality of piece-wise constant policies .

Theorem 11.5.3 *[258] Suppose that Assumptions (A1)–(A2) hold. Then, for every $\epsilon > 0$, there exists a piece-wise constant control policy in Γ_{RC} (thus also non-anticipative) which is ϵ-optimal.*

Proof. From [15, Theorem 2.3.1], we know that the set of non-anticipative measures with quantized support (in both time and control) is dense in $\mathcal{P}_\lambda([0,T]\times\mathbb{U})$. Thus, by the continuity of the cost as a function of policies (as we have established in Theorem 11.5.2), we obtain our result .

An approach with measure transformation

In an alternative approach, we consider the state process to be exogenous and the control only impacting the cost function. For the analysis of this subsection, we are assuming that the running cost $c(x,u)$ is bounded measurable and continuous in its second argument (i.e., only in u) and c_T is bounded measurable. differently from the previous subsection, we define relaxed wide-sense admissible control policy in the following. As in the above, we first place the Young topology on the control action space, by viewing the control to be a probability measure on $C([0,T];\mathbb{R}^N) \times \mathcal{P}_\lambda([0,T]\times\mathbb{U})$ with its fixed marginal on $C([0,T];\mathbb{R}^N)$ to be the Wiener measure, moreover we require that, under the measure-transformed model, $m_{[0,s]}$ be independent of $W_t - W_s$ for any $t > s$.

Let $\mathit{\Gamma}_{\mathsf{WRC}}$ denote the space of all wide-sense admissible control policies. A typical element of $\mathit{\Gamma}_{\mathsf{WRC}}$ is denoted by m (without loss of generality). To study this, we adopt Girsanov's measure transformation. Define

$$dX'_t = \sigma(X'_t)dW_t,$$

which, in fact, is policy independent (under new probability measure P_0), where

$$\frac{dP}{dP_0} =: Z_T = \exp\Big[\int_0^T \sigma^{-1}(X_s)b(X_s,U_s)dW_s - \frac{1}{2}\int_0^T |\sigma^{-1}(X_s)b(X_s,U_s)|^2 ds\Big].$$

provided that b is integrable (uniform over control policies) and $\sigma^{-1}(x)$ exists and is bounded (which is a consequence of (A2)). In this case, the marginal on the state process is fixed, but the cost function is now represented as:

$$J(m) := E^U_{P_0}\Big[\frac{dP}{dP_0}\Big(\int_0^T c(X_s,U_s)ds + c_T(X_T)\Big)\Big]$$

Here, one wishes to minimize (where the measure on the path process is fixed)

$$\inf_{m\in\mathit{\Gamma}_{\mathsf{WRC}}} J(m)\,. \tag{11.38}$$

In the following, we adopt the latter approach, as it will be much simpler to be generalized to information structures beyond the fully observed model, including decentralized information structures. This analysis is based on the supporting result in Lemma 11.5.2. Which shows that the Radon-Nikodym derivative is continuous as a function of policies (under a suitable topology over the policy space).

Lemma 11.5.1 *The space $\mathit{\Gamma}_{\mathsf{WRC}}$ under the weak convergence topology is compact.*

Proof. Note that $\mathit{\Gamma}_{\mathsf{WRC}} \subset \mathcal{P}\left(C([0,T];\mathbb{R}^N)\times\mathcal{P}_\lambda([0,T]\times\mathbb{U})\right)$. Since (in $\mathit{\Gamma}_{\mathsf{WRC}}$) the marginal on $C([0,T];\mathbb{R}^N)$ is fixed and $\mathcal{P}_\lambda([0,T]\times\mathbb{U})$ is compact (via Prohorov's theorem), it follows that $\mathit{\Gamma}_{\mathsf{WRC}}$ is tight. Thus relatively compact by Prohorov's theorem. Since independence is preserved under weak convergence of probability measures (see e.g. the proof of [132, Lemma 2.3] or [346, Theorem 5.6]), thus $\mathit{\Gamma}_{\mathsf{WRC}}$ is also closed, hence compact.

The next lemma shows that the Radon-Nikodym derivative is continuous as a function of policies over $\mathit{\Gamma}_{\mathsf{WRC}}$ (under the topology of weak convergence).

Lemma 11.5.2 *Suppose that Assumptions (A1)–(A2) hold. Then, on $\mathit{\Gamma}_{\mathsf{WRC}}$, the map*

$$U \mapsto \exp\Big[\int_0^T \sigma^{-1}(X_s)b(X_s,U_s)dW_s - \frac{1}{2}\int_0^T |\sigma^{-1}(X_s)b(X_s,U_s)|^2 ds\Big]$$

is continuous in L^1 norm.

Using this continuity property of the Radon-Nikodym derivative as a function of policy, we arrive at the following continuity result. The proof of the following lemma follows from [258, Lemma 2.6] .

Lemma 11.5.3 *Suppose that Assumptions (A1)–(A2) hold. Then, J is continuous in $m \in \Gamma_{\mathsf{WRC}}$ under the weak convergence topology.*

Next theorem proves the existence of an optimal policy in Γ_{WRC}. Also, it shows that the piece-wise constant policies in Γ_{WRC} are near optimal .

Theorem 11.5.4 *[258] Suppose that Assumptions (A1)–(A2) hold. Then, we have*

(i) There exists an optimal control policy in Γ_{WRC}.

(ii) For every $\epsilon > 0$, there exists a piece-wise constant control policy in Γ_{WRC} (thus also non-anticipative) which is ϵ-optimal.

11.5.3 Existence of Optimal Policies for Partially Observed Models

Consider now a partially observed continuous-time process $\{X_t\}$ on $\mathbb{R}^N$, controlled by a control process $\{U_t\}$ taking values in a compact action space $\mathbb{U} \subset \mathbb{R}^L$, and with an associated observation process $\{Y_t\}$ taking values in $\mathbb{R}^M$, where $0 \leq t \leq T$. The evolution of $\{X_t, Y_t\}$ is given by the stochastic differential equations

$$\begin{aligned} dX_t &= b(X_t, U_t)dt + \sigma(X_t)dW_t, \\ dY_t &= g(X_t)dt + dB_t\,, \end{aligned} \tag{11.39}$$

where, $g : \mathbb{R}^N \to \mathbb{R}^M$ is a continuous and bounded function and W and B are independent standard Wiener processes with values in $\mathbb{R}^N$ and $\mathbb{R}^M$, respectively (hence, σ is a $N \times N$-matrix). The objective is to minimize the following cost function

$$E\bigg[\int_0^T c(X_t, U_t)dt + c_T(X_T)\bigg], \tag{11.40}$$

where $c : \mathbb{R}^N \times \mathbb{U} \to [0, \infty)$ and $c_T : \mathbb{R}^N \to [0, \infty)$ are bounded and continuous functions.

The idea is again to first apply Girsanov's transformation so that the measurements Y_t form an independent Wiener process under new probability measure Q . Following Fleming and Pardoux [132, p. 264], we define an admissible control as a probability measure on $C([0,T] \times \mathbb{R}^M) \times \mathcal{P}_\lambda([0,T] \times \mathbb{U})$ with its fixed marginal on $C([0,T] \times \mathbb{R}^M)$ be the Wiener measure. In addition, under the new measure Q, $Y_r - Y_t$ is independent of $\{X_0, W., Y_s, m_s; s \leq t\}$, for any $0 \leq t \leq r \leq T$. Let Γ_{WS} denote the space of such policies, where we endow this space with the weak convergence topology. As inc Lemma 11.5.1, we have Γ_{WS} is compact under weak convergence topology . Without loss of generality, a typical element of Γ_{WS} is denoted by m.

As in Section 11.4, suppose that $g \equiv 0$, that is Y_t is just an independent process. Let us call this probability measure on the processes as $\mathbf{Q} := P_X \times Q_Y$.

Now, consider again our original process where g is non-zero. Let $\mathbf{P}$ be the joint probability measure on the state and the measurement processes. Since $\mathbf{P} \ll \mathbf{Q}$, we have that

$$\begin{aligned} &\int f(X_{[0,t]}, Y_{[0,t]})\mathbf{P}(dX_{[0,t]}, dY_{[0,t]}) \\ &= \int G(X_{[0,t]}, Y_{[0,t]})f(X_{[0,t]}, Y_{[0,t]})\mathbf{Q}(dX_{[0,t]}, dY_{[0,t]}) \end{aligned}$$

for some $\mathbf{Q}$-integrable function G, which is the Radon-Nikodym derivative of $\mathbf{P}$ with respect to $\mathbf{Q}$. Under mild conditions, we have that

$$G_t := G(X_{[0,t]}, Y_{[0,t]}) = \frac{d\mathbf{P}}{d\mathbf{Q}}(X_{[0,t]}, Y_{[0,t]}) \\ = e^{\int_0^t g(X_s)dY_s - \frac{1}{2}\int_0^t |g(X_s)|^2 ds},$$

with $\int_0^t g(X_s)dY_s$ being a stochastic integral, this time with respect to the random measure/process Y_s. This relation allows us to view the partially observed problem as one with independent measurements, with the dependence pushed to the Radon-Nikodym derivative. In particular, we get an equivalent model

$$dX_t = b(X_t, U_t)dt + \sigma(X_t)dW_t, \\ dY_t = dB_t\,. \tag{11.41}$$

Theorem 11.5.5 *[258] Suppose that the drift term b and the diffusion matrix σ satisfy Assumptions (A1)–(A2), uniformly with respect to $y \in \mathbb{R}^M$. Then,*

(i)

$$J(m) = E\left[\frac{d\mathbf{P}}{d\mathbf{Q}}(X_{[0,T]}, Y_{[0,T]})\left(\int_0^T c(X_t, U_t)dt + c_T(X_T)\right)\right], \tag{11.42}$$

is continuous over the space of wide-sense admissible policies Γ_{WS}.

(ii) There exists an optimal control policy in Γ_{WS}.

(iii)For every $\epsilon > 0$, there exists a piece-wise constant control policy in Γ_{WS} which is ϵ-optimal.

Accordingly, we can again establish both existence and discrete approximation results. Once again, the above will allow us to approximate a continuous-time process with a (sampled) discrete-time process and the machinery developed for discrete-time optimal control will be applicable.

Remark 11.16. The utility of this approach was already observed in Section 10.8 (see Remark 10.17). In particular, if one makes the measurements independent, so that the information structure is first static, and then makes the information structure classical by considering the actions at time t measurable on the filtration generated by the past noise processes and actions up to time t; Theorem 10.16, building on [346, Theorem 5.6], can be adapted to show that such a set of measurement-action measures (with fixed marginal on the measurements) that satisfy conditional independence $u_{[0,t]} \leftrightarrow y_{[0,t]} \leftrightarrow y_s - y_t; (x_0, W_[0,T])$ is weakly closed. Furthermore, the value is continuous in the joint measure on $\{x_s; (u,y)_s, s \in [0,T]\}$ and this set of measures is tight. These lead to the compactness-continuity conditions and accordingly an existence result for optimal policies follows. Furthermore, by showing that the set of $\{(u,y)_s, s \geq 0\}$ measures which have quantized support in the measurement variable are dense, one can show also that piece-wise constant control policies are nearly optimal. This allows one to approximate a continuous-time process with a (sampled) discrete-time process and the machinery developed earlier in the lecture notes are applicable. This approach is the essence of Kushner's method [205, p. 278] [203], though stated somewhat differently. This approximation result by discrete-time models also applies for fully-observed models with a similar argument (see Exercise 11.10.2(b)).

Remark 11.17. It may be important to note that Bismut [46] arrived at further existence results, through an approach which avoids separation (and the construction of a belief-MDP), in discrete-time a similar approach is given in Section 10.8.1.

Remark 11.18 (Revisiting the Discrete-time Case). Inspired by the work of Fleming and Pardoux [132], Borkar introduced wide-sense control policies to study discrete-time partially-observed finite state-observation Markov decision processes with average cost criterion (see [58, 60, 62, 63]). We recognize also that Borkar achieves what is in essence equivalent to Witsenhausen's static reduction reviewed earlier in Section 10.4.2. For simplicity, we only consider here the case where state and observation spaces are finite. We consider a discrete-time Markov decision process $\{x_n\}$ on a finite state space $\mathbb{X}$, controlled by a control process $\{u_n\}$ taking values in a compact Borel action space $\mathbb{U}$, and with an associated observation process $\{y_n\}$ taking values in a finite observation space $\mathbb{Y}$, where $n = 0, 1, 2, \ldots$. The evolution of $\{x_n, y_n\}$ is given by

$$P\big(x_{n+1}, y_{n+1} \in \cdot \big| x_m, x_m, u_m, m \le n\big) = \rho(x_{n+1}, y_{n+1} \in \cdot | x_n, u_n),$$

where $\rho : \mathbb{X} \times \mathbb{U} \to \mathcal{P}(\mathbb{X}) \times \mathcal{P}(\mathbb{Y})$ is some transition kernel. To ease the exposition, we assume that ρ is of the following form:

$$\rho(x_{n+1}, y_{n+1} | x_n, u_n) = r(y_{n+1} | x_{n+1}) \otimes p(x_{n+1} | x_n, u_n),$$

where p is the state transition kernel and r is the observation kernel. The initial distribution of x_0 is μ.

A control process $\{u_n\}$ is admissible in classical sense if it is adapted to the filtration $\{\sigma(y_m, m \le n)\}$ generated by observations $\{y_n\}$. In this case, one can write

$$u_n = \pi_n(y_0, \ldots, y_n), n \ge 0, \tag{11.43}$$

for some $\pi_n : \prod_{k=0}^{n} \mathbb{Y} \to \mathbb{U}$. Let us denote $\pi = \{\pi_n\}$.

Note that one can always write the evolution of the state process $\{x_n\}$ as a noise-driven dynamical system

$$x_{n+1} = F(x_n, u_n, w_n), \tag{11.44}$$

where $F : \mathbb{X} \times \mathbb{U} \times [0, 1] \to \mathbb{X}$ is measurable and $\{w_n\}$ are independently and identically distributed uniformly on $[0, 1]$. Using this dynamical system, we now reproduce the above process on a more convenient probability space. This will then enable us to define *wide-sense admissible policies*.

We can thus reduce the problem to an independent static one via Witsenhausen/Girsanov/Borkar, see Borkar's [58, 62] explicit analysis or Witsenhausen's method presented in Section 10.4.2.

Under this reduction, we obtain a new probability space P_0^π under which:

(a) $\{y_n\}$ is i.i.d. uniform on $\mathbb{Y}$ and independent of x_0 and $\{w_n\}$,

(b) $\{u_0, \ldots, u_n, y_0, \ldots, y_n\}$ is independent of $\{w_n\}$, x_0, and $\{y_m, m > n\}$, for all n.

Using these properties, Borkar defines P_0 to be *wide sense admissible* if P_0 satisfies (a) and (b). Such a notion allows for closedness of conditional independence properties under weak convergence of joint probability measures, and thus leads to very general existence results. See [271] for a subtle clarification.

11.5.4 Existence for Models with Decentralized Information

Decentralized Model with Local Measurements

Consider now a continuous-time process $\{X_t\}$ on a Euclidean space $\mathbb{R}^n$, controlled by a collection of control process $\mathbf{U}_t := \{U_t^k, k = 1, \cdots, N\}$ with each U_t^k taking values in a compact Borel action space $\mathbb{U}^k \subset \mathbb{R}^L$, and with an associated observation process $\{Y_t^k\}$ taking values in $\mathbb{R}^M$, where $0 \le t \le T$. Let $\mathbf{Y} = \{Y^1, \cdots, Y^N\}$. The evolution of $\{X_t, Y_t^k, k = 1, \cdots, N\}$ is given by the stochastic differential equations

$$\begin{aligned} dX_t &= b(X_t, U_t^1, \cdots, U_t^N)dt + \sigma(X_t)dW_t, \\ dY_t^i &= g^i(X_t)dt + dB_t^i, \qquad i = 1, \cdots, N. \end{aligned} \tag{11.45}$$

Where, W and $B^i, i = 1, \cdots, N$ are independent standard Wiener processes with values in $\mathbb{R}^n$ and $\mathbb{R}^M$, respectively and $g^i : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous and bounded function . In this section, we assume that the drift term b and the diffusion matrix σ satisfies similar conditions as in (A1) and (A2). In particular, they satisfy the following:

(D1) The function $b \colon \mathbb{R}^n \times (\mathbb{R}^M)^N \times \prod_{k=1}^{N} \mathbb{U}^k \to \mathbb{R}^n$ is jointly continuous and $\sigma = \big[\sigma^{ij}\big] \colon \mathbb{R}^n \times (\mathbb{R}^M)^N \to \mathbb{R}^{n \times n}$ is locally Lipschitz continuous in x (uniformly with respect to the $y \in (\mathbb{R}^M)^N$). In particular, for some constant $C_R > 0$ depending on $R > 0$, we have

$$\|\sigma(x_1, y) - \sigma(x_2, y)\|^2 \leq C_R |x_1 - x_2|^2$$

for all $x_1, x_2 \in B_R$, $y \in (\mathbb{R}^M)^N$, where $\|\sigma\| := \sqrt{(\sigma\sigma^T)}$.

(D2) The functions b and σ are uniformly bounded, i.e., for some constant $C > 0$,

$$\sup_{u \in \prod_{k=1}^N \mathbb{U}^k} |b(x, y, u)| + \|\sigma(x, y)\|^2 \leq C \quad \forall\, x \in \mathbb{R}^d,\, y \in (\mathbb{R}^M)^N.$$

(D3) For some $\hat{C}_1 > 0$, it holds that

$$\sum_{i,j=1}^d a^{ij}(x, y) z_i z_j \geq \hat{C}_1 |z|^2 \quad \forall\, x \in \mathbb{R}^n,\, y \in (\mathbb{R}^M)^N,$$

and for all $z = (z_1, \ldots, z_d) \in \mathbb{R}^n$, where $a := \frac{1}{2}\sigma\sigma^T$.

The objective here is to minimize the following cost function

$$E\bigg[\int_0^T c(X_t, U_t^1, \cdots, U_t^N)dt + c_T(X_T)\bigg], \tag{11.46}$$

where $c : \mathbb{R}^n \times \prod_{k=1}^N \mathbb{U}^k \to [0, \infty)$ and $c_T : \mathbb{R}^n \to [0, \infty)$ are bounded continuous functions. For similar existence analysis the authors in [76, 80–82] assumed that the b, σ are uniformly Lipschitz continuous.

We define the following decoupled measurement model.

$$\begin{aligned} dX_t &= b(X_t, Y_t, U_t^1, \cdots, U^N)dt + \sigma(X_t, \mathbf{Y}_t)dW_t, \\ dY_t^i &= dB_t^i, \qquad i = 1, \cdots, N. \end{aligned} \tag{11.47}$$

Let $\mathbf{Y} = \{Y^1, \cdots, Y^N\}$ and $\mathbf{U} = \{U^1, \cdots, U^N\}$

We let this decoupled measurement model have measure $\mathbf{Q} := P_X \times \prod_{k=1}^N Q^k$. Let $\mathbf{P}$ be the joint probability measure on the state and the measurement processes under a given policy. Since $\mathbf{P} \ll \mathbf{Q}$, we have that

$$\begin{aligned} &\int f(X_{[0,t]}, \mathbf{Y}_{[0,t]})\mathbf{P}(dX_{[0,t]}, d\mathbf{Y}_{[0,t]}) \\ &= \int G(X_{[0,t]}, \mathbf{Y}_{[0,t]}) f(X_{[0,t]}, \mathbf{Y}_{[0,t]})\mathbf{Q}(dX_{[0,t]}, d\mathbf{Y}_{[0,t]}), \end{aligned}$$

for some $\mathbf{Q}$-integrable function G, which is the Radon-Nikodym derivative of $\mathbf{P}$ with respect to $\mathbf{Q}$. Under mild conditions, we have that

$$G_t := \frac{d\mathbf{P}}{d\mathbf{Q}}(X_{[0,t]}, \mathbf{Y}_{[0,t]}) = \prod_{i=1}^N e^{\int_0^t g^i(X_s)dY_s^i - \frac{1}{2}\int_0^t |g^i(X_s)|^2 ds}$$

with $\int_0^t g^i(X_s)dY_s^i$ being a stochastic integral, with respect to the random measure/process Y_s^i. This relation allows us to view the decentralized stochastic control problem as one with independent measurements, with the dependence pushed to the Radon-Nikodym derivative.

We define an admissible control as a probability measure on $C([0, T] \times \mathbb{R}^M) \times \mathcal{P}_\lambda([0, T] \times \mathbb{U}^i)$ with its fixed marginal on $C([0, T] \times \mathbb{R}^M)$ be the Wiener measure. In addition, under the new measure, $Y_r^i - Y_t^i$ is independent of $\{X_0, W_\cdot, Y_s^i, m_s^i; s \leq t\}$, for any $0 \leq t \leq r \leq T$ and independent from all $(m_\cdot^k, Y_\cdot^k), k \neq i$. Let Γ_{DWS} denote the space of such *decentralized wide sense admissible* policies, where we endow this space with the weak convergence topology. Without loss of generality, a typical element of Γ_{DWS} is denoted by $\mathbf{m} = (m^1, \cdots, m^N)$.

Theorem 11.5.6 *[258] Let Assumptions (D1)– (D3) hold. Then,*

(i) Over Γ_{DWS} (wide-sense admissible policies) the function

$$J(\mathbf{m}) = E^{\mathbf{m}}\Big[\prod_{i=1}^{N}(e^{\int_0^t g^i(X_s)dY_s^i - \frac{1}{2}\int_0^t |g^i(X_s)|^2 ds})\Big(\int_0^T c(X_t, U_t^1, \cdots, U_t^N)dt + c_T(X_T)\Big)\Big], \tag{11.48}$$

is continuous.

(ii) There exists an optimal control policy in Γ_{DWS}.

(iii)For every $\epsilon > 0$, there exists a piece-wise constant control policy in Γ_{DWS} which is ϵ-optimal.

Decentralized Model with Coupled Dynamics and Local State

Instead of (11.45) we now consider a collection of N agents with coupled dynamics given as

$$dX_t^i = b^i(X_t^i, U_t^i)dt + b_0^i(\mathbf{X}_t, \mathbf{U}_t)dt + \sigma^i(X_t^i)dB_t^i, \tag{11.49}$$

for $i = 1, \cdots, N$. Here, $b^i : \mathbb{R}^n \times \mathbb{U}^i \to \mathbb{R}^n$, $b_0^i : (\mathbb{R}^n)^N \times \prod_{k=1}^N \mathbb{U}^k \to \mathbb{R}^n$ and $\sigma^i : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ are given functions and B^i are independent standard Wiener processes with values in $\mathbb{R}^n$, $i = 1, \cdots, N$. We assume that b^i, b_0^i, σ^i for $i = 1, \cdots, N$, satisfies the following:

($\hat{D}$1) For $i = 1, \cdots, N$, we have $b^i : \mathbb{R}^n \times \mathbb{U}^i \to \mathbb{R}^n$ and $b_0^i : (\mathbb{R}^n)^N \times \prod_{k=1}^N \mathbb{U}^k \to \mathbb{R}^n$ are jointly continuous and for some constant $C_R > 0$ depending on $R > 0$, we have

$$\big\|\sigma^i(x_1) - \sigma^i(x_2)\big\|^2 \le C_R\,|x_1 - x_2|^2$$

for all $x_1, x_2 \in B_R$.

($\hat{D}$2) The functions b^i, b_0^i and σ^i, $i = 1, \cdots, N$, are uniformly bounded, i.e., for some constant $C > 0$

$$\sup_{u\in\mathbb{U}^i} |b^i(x,u)| + \big\|\sigma^i(x)\big\|^2 \le C \qquad \forall\, x \in \mathbb{R}^n,$$

and $\sup_{u\in\prod_{k=1}^N \mathbb{U}^k} |b_0^i(x,u)| \le C$ for all $x \in (\mathbb{R}^n)^N$.

($\hat{D}$3) For some $\hat{C}_1 > 0$, it holds that

$$\sum_{i,j=1}^{d} a^{k,ij}(x) z_i z_j \ge \hat{C}_1 |z|^2 \qquad \forall\, x \in \mathbb{R}^n,\ k = 1, \cdots, N,$$

and for all $z = (z_1, \ldots, z_d) \in \mathbb{R}^n$, where $a^k := \frac{1}{2}\sigma^k(\sigma^k)$.

In view of ($\hat{D}$3), it is easy to see that $(\sigma^i)^{-1}$ exists and is bounded for all $i = 1, \cdots, N$.

The objective here is to minimize the following cost function

$$E\Big[\int_0^T c(\mathbf{X}_t, U_t^1, \cdots, U^N)dt + c_T(\mathbf{X}_T)\Big], \tag{11.50}$$

where $c : (\mathbb{R}^n)^N \times \prod_{k=1}^N \mathbb{U}^k \to [0,\infty)$ and $c_T : (\mathbb{R}^n)^N \to [0,\infty)$ are bounded continuous functions. We assume that the control policies are only locally measurable, that is U_t^i is measurable with respect to $\sigma(X_{[0,t]}^i)$ for all $t \in [0,T]$.

We define the following decoupled (non-interacting) agent model.

$$dX_t^i = \sigma^i(X_t^i)dW_t^i, \qquad i = 1, \cdots, N; \tag{11.51}$$

Since σ^i is invertible (follows from $(\hat{D3})$), let us have the driving noise process $\Big(W_t^1 + \int_0^t (\sigma^i)^{-1}(X_s^1) \left(b^1(X_s^1, U_s^1) + b_0^1(\mathbf{X}_s, \mathbf{U}_s)\right) ds, \cdots,$ $\int_0^t (\sigma^N)^{-1}(X_s^N) \left(b^N(X_s^N, U_s^N) + b_0^N(\mathbf{X}_s, \mathbf{U}_s)\right) ds \Big)$ have measure μ and the independent process $(W_t^1, \cdots, W_t^N)$ have measure μ_0. Let $\hat{b}^i(\mathbf{X}_s, \mathbf{U}_s) = \left(b^1(X_s^1, U_s^1) + b_0^1(\mathbf{X}_s, \mathbf{U}_s)\right)$. Then, by Girsanov, we know that the density for μ_0 with respect to μ is

$$\frac{d\mu_0}{d\mu} = \prod_{k=1}^{N} e^{(\int_0^T (\sigma^i)^{-1}(X_t)\hat{b}^i(\mathbf{X}_t, \mathbf{U}_t)dW_s^i - \frac{1}{2}\int_0^T (\sigma^i)^{-1}(X_t)\hat{b}^i(\mathbf{X}_t, \mathbf{U}_t)^2 ds)}.$$

As in Section 11.5.2, we define a relaxed wide-sense admissible control policy by first placing the Young topology on the control action space, by viewing the control process to be a probability measure on $C([0,T];\mathbb{R}^d) \times \mathcal{P}_\lambda([0,T] \times \mathbb{U}^i)$ with its fixed marginal on $C([0,T];\mathbb{R}^d)$ to be the Wiener measure . We require also that $m_{[0,t]}^i$ be independent of $W_s^i - W_t^i, s > t$ for every $t \in [0,T]$ and independent from $m^j, W_s^j,\ j \neq i, s \in [0,T]$. We call again such policies decentralized locally wide sense admissible policies, and denote with Γ_{DWS}. Without loss of generality, a typical element of Γ_{DWS} is denoted by $\mathbf{m} = (m^1, \cdots, m^N)$. Also, it is easy to see that in the trasformed model, the measure on the path space is fixed .

Now, following the analysis in Theorem 11.5.4, 11.5.6, we have the following theorem.

Theorem 11.5.7 *[258] Suppose that Assumptions $(\hat{D1})$–$(\hat{D3})$ hold. Then,*

(i) Over the space of wide-sense admissible policies Γ_{DWS} the function

$$J(\mathbf{m}) = E^{\mathbf{m}}\Bigg[\Bigg(\int_0^T c(\mathbf{X}_t, U_t^1, \cdots, U^N)dt + c_T(\mathbf{X}_T)\Bigg) \prod_{i=1}^{N} e^{\int_0^T (\sigma^i)^{-1}(X_t)\hat{b}^i(\mathbf{X}_t, \mathbf{U}_t)dW_s^i - \frac{1}{2}\int_0^T (\sigma^i)^{-1}(X_t)\hat{b}^i(\mathbf{X}_t, \mathbf{U}_t)^2 ds}\Bigg] \tag{11.52}$$

is continuous.

(ii) There exists an optimal control policy in Γ_{DWS}.

(iii)For every $\epsilon > 0$, there exists a piece-wise constant control policy in Γ_{DWS} which is ϵ-optimal.

11.6 Near Optimality of Control Policies Designed for Discrete-time Models via Sampling

In view of the results presented in the previous section, we have that piece-wise constant policies are near optimal. These then lead to discrete-time models whose solutions will be near optimal, and applicable to the original problems. For each of the information structures below, we will consider the following arguments: (i) We obtain a sequence of discrete-time models $\mathcal{T}_n$ (here the subscript n denotes the index of the model sequence, and not the fixed dimension of the state space) arrived from piece-wise constant control policies applied to the true model $\mathcal{T}$. (ii) We show that the solution to the optimal discrete-time model $J(\mathcal{T}_n)$ leads to a solution which is near optimal. The direction, $\lim_{n\to\infty} J(\mathcal{T}_n) \leq J(\mathcal{T}) + \epsilon$ follows from the analysis above and the direction $\lim_{n\to\infty} J(\mathcal{T}_n) \geq J(\mathcal{T})$ follows from the fact that restricting to piece-wise constant policies cannot lead to a better policy when compared with arbitrary admissible policies. (iii) We then show that the policy obtained to solve $J(\mathcal{T}_n)$ can be applied to $\mathcal{T}$ (and thus the approach is constructive) and is near optimal for large n.

11.6.1 Fully Observed Setup

Consider the fully observed setup discussed in Section 11.5.2 with model (11.34) and cost criterion (11.35). That is, with dynamics

$$dX_t = b(X_t, U_t) + \sigma(X_t)dB_t,$$

and cost criterion $J(U) = E[\int_0^T c(X_s, U_s)ds + c_T(X_T)]$.

Discrete-Time Model to be Solved for Near Optimal Solutions. Let X_h, U_h be the solution of the sampled process corresponding to (11.34) with piece-wise constant policies: $U_{h,s} = U_{kh}$ for $kh \leq s < (k+1)h$. We have that

$$X_{(k+1)h} = X_{kh} + \int_{kh}^{(k+1)h} b(X_s, U_{kh})ds + \int_{kh}^{(k+1)h} \sigma(X_s)dW_s \tag{11.53}$$

The cost can be written as, with $N_h = \frac{T}{h}$,

$$E[\sum_{k=0}^{N_h-1} \hat{c}(X_{kh}, U_{kh}) + c_T(X_{N_h h})] \tag{11.54}$$

where $\hat{c}(x,u) = E[\int_0^h c(X_s, U_0)ds | X_0 = x, U_0 = u]$. With $n = \frac{1}{h}$, the above define a discrete-time MDP with transition kernel $\mathcal{T}_n$, cost function $\hat{c}_n$ and total cost $J_n(U)$.

Thus, one defines a discrete-time model in which $X_k := X_{kh}$ and $U_k := U_{kh}$ for $k \in \mathbb{Z}_+$.

The information structure at time t contains the continuous-time measurements. However, since for such fully observed model, Markov policies are optimal, it suffices to the controller to only use the discrete-time measurements.

Theorem 11.6.1 *[258] Suppose that Assumptions (A1)–(A2) hold. Then, the value of the discrete-time model convergences to the value of the original continuous-time model. Moreover, for every $\epsilon > 0$, there exists $h > 0$ so that the solution of the discrete-time approximation gives a policy which is near optimal for the original continuous-time model.*

The above apply also to the partially observed and decentralized models.

Partially Observed Setup

Consider the setup in Section 11.5.3 with dynamics (11.39) and criterion 11.40.

Discrete-Time Model to be Solved for Near Optimal Solutions. Let us have a piece-wise constant control policy with $U_{h,s} = U_{kh}$ for $kh \leq s < (k+1)h$. Let X_h, Y_h, U_h be the solution of the sampled process corresponding to (11.34) with piece-wise constant policies. We have that

$$\begin{aligned} X_{(k+1)h} &= X_{kh} + \int_{kh}^{(k+1)h} b(X_s, U_{kh})ds + \int_{kh}^{(k+1)h} \sigma(X_s)dB_s \\ Y_t &= Y_{kh} + \int_{kh}^{t} g(X_s)ds + \int_{kh}^{t} dB_s, \quad t \in [kh, (k+1)h) \end{aligned} \tag{11.55}$$

Thus, one defines a discrete-time model in which $X_k := X_{kh}$ and $U_k := U_{kh}$ for $k \in \mathbb{Z}_+$, and the path-valued discrete-time measurement $\bar{Y}_k = Y_{[kh,(k+1)h)}$, for $k \in \mathbb{Z}_+$.

The cost can be written as, with $N_h = \frac{T}{h}$,

$$E[\sum_{k=0}^{N_h-1} \hat{c}(X_{kh}, U_{kh}) + c_T(X_{N_h h})] \tag{11.56}$$

with $\hat{c}(x,u) = E[\int_0^h c(X_s, U_0)ds | X_0 = x, U_0 = u]$. With $n = \frac{1}{h}$, the above define a discrete-time POMDP with transition kernel $\mathcal{T}_n$ and cost function $\hat{c}_n$. Following the proof technique as in Theorem 11.6.1 (and Theorem 11.5.4), we obtain the following near- optimality result for partially observable model.

Theorem 11.6.2 *[258] Suppose that the drift term b and the diffusion matrix σ satisfy Assumptions (A1)–(A2), uniformly with respect to $y \in \mathbb{R}^M$. Then, the optimal value of the discrete-time model convergences to the optimal value of the original continuous-time model. Moreover, for every $\epsilon > 0$, there exists $h > 0$ so that the solution of the discrete-time approximation gives a policy which is near optimal for the original continuous-time model.*

The result above, however, while involves discrete-time control and state, requires having access to the path-valued measurements. It would be desirable to obtain a discrete-time model with discrete-time measurements Y_{kh} in the original measurement space. That is, at time kh, we would like to have $U_{kh} = \gamma(Y_{ih}, i \in \{0, 1, \cdots, k\})$ under an admissible policy γ. The following result (refinement to Theorem 11.6.2), building on Lusin's theorem, achieves this (for more details see [258, Section 5.2]).

Theorem 11.6.3 *[258] Suppose that the drift term b and the diffusion matrix σ satisfy Assumptions (A1)–(A2), uniformly with respect to $y \in \mathbb{R}^M$. Then, the optimal value of the discrete-time model (11.55) convergences to the optimal value of the original continuous-time model. Moreover, for every $\epsilon > 0$, there exists $h > 0$ so that the solution of the discrete-time approximation gives a policy (i.e., a policy $U^{\epsilon}_{kh} = \gamma^{\epsilon}(Y_{ih}, i \in \{0, 1, \cdots, k\})$ obtained in the deiscretized molde of (11.41) as in (11.58)) which is near optimal for the original continuous-time model.*

Decentralized Setup

Decentralized Model with Local Measurements. Consider Section 11.5.4 with dynamics (11.45) and cost criterion (11.46). In particular the model is

$$\begin{aligned} dX_t &= b(X_t, U^1_t, \cdots, U^N)dt + \sigma(X_t)dW_t, \\ dY^i_t &= g^i(X_t)dt + dB^i_t, \qquad i = 1, \cdots, N. \end{aligned} \tag{11.57}$$

Discrete-Time Model to be Solved for Near Optimal Solutions. Let us have a piece-wise constant control policy with $U^i_{h,s} = U^i_{kh}$ for $kh \le s < (k+1)h$, $i = 1, \cdots, N$. Let X_h, Y^i_h, U^i_h be the solution of the sampled process corresponding to (11.57) with piece-wise constant policies. We have that

$$\begin{aligned} X_{(k+1)h} =& X_{kh} + \int_{kh}^{(k+1)h} b(X_s, U^1_{kh}, \cdots, U^N_{kh})ds + \int_{kh}^{(k+1)h} \sigma(X_s)dB_s \\ Y^i_t =& Y^i_{kh} + \int_{kh}^{t} g^i(X_s)ds + \int_{kh}^{t} dB^i_s, \quad t \in [kh, (k+1)h). \end{aligned} \tag{11.58}$$

By the similar argument as in the partially observable case, applying Girsanov's change of measure argument to the discretized model (11.58), we can define a new probability measure space in which the measurements of the each individual are independent of the state process. In particular, we obtain the following equivalent discretized model

$$\begin{aligned} X_{(k+1)h} =& X_{kh} + \int_{kh}^{(k+1)h} b(X_s, U^1_{kh}, \cdots, U^N_{kh})ds + \int_{kh}^{(k+1)h} \sigma(X_s)dW_s \\ Y^i_t =& Y^i_{kh} + \int_{kh}^{t} dB^i_s, \quad t \in [kh, (k+1)h), \end{aligned} \tag{11.59}$$

with policy $U^i_{kh} = \gamma^i(Y^i_s, s \le kh)$, by Lusin's theorem, for some continuous function γ^i_c we have $\gamma^i = \gamma^i_c$ on a set of measure $(1 - \epsilon_i)$. Moreover, we have that the process $Y^i_{[0,kh]}$ can be approximated by its piece-wise constant interpolations. Since both running/terminal costs are continuous, the cost (11.60) under a policy $\gamma = (\gamma^1, \ldots, \gamma^N)$ (under continuous-time measurements) and its continuous approximation $\gamma_c = (\gamma^1_c, \ldots, \gamma^N_c)$ (with discrete-time measurements) are close to each other. This enables us to obtain a discrete-time model with discrete-time measurements.

Thus, one defines a discrete-time model in which $X_k := X_{kh}$ and $U_k^i := U_{kh}^i$ for $k \in \mathbb{Z}_+$, and the path-valued discrete-time measurement $\bar{Y}_k^i = Y_{[kh,(k+1)h)}^i$, for $k \in \mathbb{Z}_+$.

The cost can be written as, with $N_h = \frac{T}{h}$,

$$E[\sum_{k=0}^{N_h-1} \hat{c}(X_{kh}, U_{kh}^1, \cdots, U_{kh}^N) + c_T(X_{N_h h})] \tag{11.60}$$

with $\hat{c}(x, \mathbf{u}) = E[\int_0^h c(X_s, U_0^1, \cdots, U_0^N)ds | X_0 = x, \mathbf{U}_0 = \mathbf{u}]$. With $n = \frac{1}{h}$, the above define a discrete-time decentralized POMDP with transition kernel $\mathcal{T}_n$ and cost function $\tilde{c}_n$. Again, for the decentralized model, similar proof technique as in Theorem 11.6.1, gives us the following near- optimality result .

Theorem 11.6.4 *[258] Suppose that Assumptions (D1)–(D3) hold. Then, the optimal value of the discrete-time model (11.58) convergences to the optimal value of the original continuous-time model. Moreover, for every $\epsilon > 0$, there exists $h > 0$ so that the solution of the discrete-time approximation gives a policy (i.e., a policy $U_{kh}^{i,\epsilon} = \gamma^{i,\epsilon}(Y_{rh}, r \in \{0, 1, \cdots, k\})$, $i = 1, \cdots, N$ an optimal solution of (11.59)) which is near optimal for the original continuous-time model.*

Coupled Dynamics and Local State. An analogous result is applicable for the model in Section 11.5.4.

11.6.2 An alternative discrete-time approximation: Euler-Maruyama discretization

In addition to the discrete-time model presented above, one can also consider the following (more standard and elementary) Euler–Maruyama (EM) approximation for SDEs. For the given controlled SDE

$$dX_t = b(X_t, U_t)\, dt + \sigma(X_t)\, dW_t, \qquad X_0 = x,$$

where U_t is an admissible control and W_t is a standard Brownian motion. Fix a time step $h > 0$ and set $t_k = kh$. The (explicit) Euler–Maruyama scheme under a *piecewise-constant-in-time* control $U_t^h = U_{t_k}$ for $t \in [t_k, t_{k+1})$ is given by the recursion

$$\bar{X}_{t_{k+1}}^h \;=\; \bar{X}_{t_k}^h + b\big(\bar{X}_{t_k}^h, U_{t_k}\big)\, h + \sigma\big(\bar{X}_{t_k}^h\big)\, \Delta W_k, \tag{11.61}$$

where $\Delta W_k := W_{t_{k+1}} - W_{t_k}$ and we denote by $\bar{X}^h(\cdot)$ the usual continuous-time interpolation (piecewise constant or piecewise linear as required).

The discrete-time per-stage cost writes as: $c_h(x, \zeta) := c(x, \zeta) \times h$. The associated discounted cost of the approximating discrete-time model under the piece-wise constant control process U^h is given by

$$J_{\alpha,h}(x) = E_x^{U^h}[\sum_{k=0}^{\infty} \beta^k c_h(X_k^h, U_k^h) | X_0^h = x] \tag{11.62}$$

for $x \in \mathbb{R}^d$, where $\beta = e^{-\alpha h}$.

Let v^{h*} be optimal for the discrete-time model (11.61) under cost (11.62).

Then following the arguments as in [257, Theorem 4.3], we then have near optimality of the discrete-time optimal policy v^{h*} obtained for the model (11.61) under cost (11.62) for the continuous time model as the parameter of the discretization approaches to zero .

Theorem 11.19. *Suppose that Assumptions (A1)–(A3) hold. Then we have*

$$\lim_{h \to 0} J_\alpha^{v^{h*}}(x) = J_\alpha^*(x) \quad \textit{a.e. } x \in \mathbb{R}^d, \tag{11.63}$$

where $J_\alpha^{v^{h*}}(x)$ is the expected cost induced by v^{h*} for the true diffusion under the discounted cost criterion, and J_α^* is the value function for the diffusion under the discounted cost criterion.

The same applies for the finite horizon criterion as well.

11.6.3 Borkar's Control Topology and a Partial Differential Equations Approach

Another approach to arrive at near optimality of discretized time and space models is via showing that the cost is continuous on the space of control policies (stationary or Markov, depending on the cost criterion) under an appropriate topology due to Borkar [55], and then to show that time and space quantized policies are dense in the space of all such policies [256].

11.7 Stochastic Stability of Diffusions

Recall that an Itô diffusion is a stochastic process $X_t(\omega)$ satisfying a stochastic differential equation of the form:

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \quad t \geq s, X_s = x$$

where B_t is m-dimensional Brownian motion. Often b, σ satisfy regularity conditions of the form

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq D|x - y|,$$

for some finite D (if one wishes to impose the existence of strong solutions), though this is not a requirement for the analysis to follow. Note that here b, σ only depend on x and not on t. Thus, the process here is time-homogenous.

Continuous-time counterparts of Foster-Lyapunov criteria considered in *Chapters* 3 and 4 exist and are well-developed. We refer the reader to [206], [223], [236] [235] as well as [194]. Dynkin's formula plays a key role in obtaining the continuous-time counterparts of the Foster-Lyapunov criteria developed in *Chapter* 4.

For functions $V : \mathbb{X} \to \mathbb{R}_+$ that are properly defined, as in the Foster-Lyapunov criteria studied in *Chapter 4*, conditions of the form

$$\mathcal{A}V(x) \leq b1_{x\in S}$$
$$\mathcal{A}V(x) \leq -\epsilon + b1_{x\in S}$$
$$\mathcal{A}V(x) \leq -f(x) + b1_{x\in S},$$

will lead to recurrence, positive Harris recurrence and finite expectations, respectively. However, the conditions needed on both V and the Markov process need to be carefully addressed. For example, one needs to ensure that the processes are non-explosive, that is, they do not become unbounded in finite time; and one needs to establish conditions for the strong Markov property. Furthermore, they must lead to a well-defined $\mathcal{A}V(x)$ (see Definition 11.2.7).

In the following, we review related results from Meyn and Tweedie [235, 236]. We consider processes taking values from a locally compact Polish space $\mathbb{X}$.

Let $P^t(x, B) := P_x(X_t \in B)$ for $B \in \mathcal{B}(\mathbb{X})$. Let for any Borel A,

$$\eta_A = \int_0^\infty 1_{\{X_t\in A\}}dt,$$

denote the occupation time. We say that the process X_t is ψ-irreducible if

$$\psi(B) > 0 \implies E_x[\eta_B] > 0, \qquad x \in \mathbb{X},$$

and the process is Harris recurrent if

$$\psi(B) > 0 \implies P_x(\eta_B = \infty) = 1, \qquad x \in \mathbb{X}.$$

Definition 11.20. *A probability measure π on $\mathcal{B}(\mathbb{X})$ is invariant if for every $B \in \mathcal{B}(\mathbb{X})$*

$$\pi(B) = \int \pi(dx) P^t(x, B), \qquad \forall t > 0.$$

A Harris recurrent chain which admits an invariant probability measure is called *positive Harris recurrent*.

Denote by $D(\mathcal{A})$ the set of all functions $V : \mathbb{X} \times \mathbb{R}_+ \to \mathbb{R}$ for which there exists a measurable function $U : \mathbb{X} \times \mathbb{R}_+ \to \mathbb{R}$ such that for each $x \in \mathbb{X}$, $t > 0$,

$$E_x[V(X_t, t)] = V(x, 0) + E_x[\int U(x_s, s) ds]$$

and

$$\int_0^t E_x[|U(x_s, s)|] ds < \infty \tag{11.64}$$

In this case, we have $\mathcal{A}V(x) = U(x)$ and we call $\mathcal{A}$ the extended infinitesimal generator of the process X_t and we say that V is in the domain of $\mathcal{A}$.

In general, it is not easy to know when a function is in the domain of $\mathcal{A}$. One method to enhance the set of functions that are relevant is to consider truncated processes. Let $\{O_m, m \in \mathbb{N}\}$ be a sequence of open bounded sets (with compact closure) for which for every $m \in \mathbb{N}$, $O_m \subset O_{m+1} \subset O_{m+2}$ with $\cup_{m=1}^{\infty} O_m = \mathbb{X}$. Define:

$$T^m = \tau_{O_m^c} := \inf\{t \geq 0 : X_t \in O_m^c\}$$

and let

$$\zeta = \lim_{m \to \infty} T^m.$$

We call X_t non-explosive if $P_x(\zeta = \infty) = 1$ for all $x \in \mathbb{X}$.

Let for $m \in \mathbb{Z}_+$, Δ_m denote a fixed state in O_m^c and define with x^m:

$$x_t^m = x_t 1_{\{t < T^m\}} + \Delta_m 1_{\{t \geq T^m\}}$$

Thus, for a non-explosive process, we can define $x_t^m = x_{\min\{t, T^m\}}$.

For Itô processes, let $\mathcal{A}_m$ denote the extended infinitesimal generator for x_t^m. In this case, $\mathcal{A}_m$ contains C^2 (class of functions on $\mathbb{X} \times \mathbb{R}$ with continuous first and second partial derivatives).

It is important to note that in general, the domain of $\mathcal{A}$ may be smaller than the domain of $\mathcal{A}_m$, in view of the integrability condition stated in (11.64).

Theorem 11.7.1 *[236, Theorem 4.5] Let X_t be non-explosive weak Feller process: that is $P^t g(x) = E[g(X_t)|x_0 = x]$ is continuous in x for every continuous and bounded g, for all $t \geq 0$. Then,*

(i) If

$$\mathcal{A}_m V(x) \leq -\epsilon + b 1_{x \in S}, \quad x \in O_m, m \in \mathbb{N} \tag{11.65}$$

holds for some compact S, then an invariant probability exists.

(ii)

$$\mathcal{A}_m V(x) \leq -f(x) + b 1_{x \in S}, \quad x \in O_m, m \in \mathbb{N} \tag{11.66}$$

holds for compact S and $f : \mathbb{X} \to [1, \infty)$, then under any invariant probability measure π, $E_\pi[f(X)] \leq b$.

Proof. Building on [133] [294, Theorem 2] (see Theorem 3.3.2 for an argument in discrete-time), for a weak Feller process there are two possibilities: either an invariant probability exists or

$$\lim_{T\to\infty} \sup_{\nu} \frac{1}{T} \int_0^T \nu P^s(C) ds = 0,$$

for all compact C, where the supremum is over all initial probability measures on x_0. Condition (i) then implies that the latter cannot take place. The second result follows from the discussion for Theorem 4.2.5 together with (i). $\diamond$

A useful technique in arriving at stochastic stability is to sample the process to obtain a discrete-time Markov chain, whose stability will imply the stability of the original process through a careful construction of invariant probability measures, similar to the discussion on sampled chains in *Chapter* 3: Any invariant measure for the continuous-time process is also invariant for a sampled discrete-time process, and thus the uniqueness of an invariant measure for the sampled process would imply the uniqueness of an invariant measure for the continuous-time process, provided one exists.

Let a be a probability measure on $\mathbb{R}_+$. Define

$$K_a(x, B) = \int P^t(x, B) a(dt)$$

Thus, K_a represents a sampled chain. A Borel set S is called ν_a-petite if ν_a is a non-trivial measure and a is a probability measure on $(0, \infty)$ that satisfies:

$$K_a(x, B) \geq \nu_a(B), \qquad x \in S,$$

for all $B \in \mathcal{B}(\mathbb{X})$. Furthermore, we have the following if S is a petite set. Meyn and Tweedie define a process to be a T-process if for some distribution a, the kernel $K_a(x, A) \geq T(x, A)$ where $(\cdot, A)$ is lower semi-continuous for each Borel A and $T(x, \mathbb{X}) \neq 0$ for each $x \in \mathbb{X}$. Note that strong Feller processes are T-processes as T can be taken to be K_a itself. Recall from Theorem 3.2.8 that for an irreducible T-process, every compact set is petite [236].

Theorem 11.7.2 *[236, Theorem 4.2] Let $\{x_t\}$ be an irreducible non-explosive process and (11.65) hold for S closed and petite, and with V bounded on S. Then, the process is positive Harris recurrent.*

We also refer the reader to [90, Theorem 4.1] and emphasize that, as in *Chapter* 4, irreducibility is not required for the existence of an invariant probability measure. Theorem 11.7.2 then implies the importance of the petiteness condition on S. As we observed earlier in *Chapter* 3, such sets allow for regeneration and hence lead to Harris recurrence and uniqueness of an invariant probability measure.

Let the notation $\{\lim_{t\to\infty} |x_t| = \infty\}$ denote the event that for any compact C, for all t sufficiently large $x_t \notin C$. If

$$P_x(\lim_{t\to\infty} |x_t| = \infty) = 0,$$

x_t is said to be non-evanescent.

Theorem 11.7.3 *[236, Theorem 3.1] Let $\{x_t\}$ satisfy*

$$\mathcal{A}_m V(x) \leq b 1_{x\in S}, \quad x \in O_m, m \in \mathbb{N}$$

for a compact S, $b < \infty$ and where V is a norm-like function (i.e., $\lim_{x\to\infty} V(x) = \infty$). Then,

$$P_x(\lim_{t\to\infty} |x_t| = \infty) = 0$$

for each $x \in \mathbb{X}$.

Further stochastic stability results, beyond the existence of invariant probability measures, have found applications; for these, we refer the reader to [206] and [312]. We state one next.

Theorem 11.7.4 *[206] [312, Prop. 5.5.1] Suppose that there exists a function $V : \mathbb{R}^n \to \mathbb{R}_+$ which is in the domain of $\mathcal{A}_m$ for every m, and satisfies*

$$\mathcal{A}V(x) \leq -\alpha V(x) + b, \qquad x \in O_m, m \in \mathbb{N} \tag{11.67}$$

for some $\alpha, b > 0$. Then,

$$E[V(X_t)] \leq e^{-\alpha t} E[V(X_0)] + \frac{b}{\alpha},$$

provided that $E[V(X_0)] < \infty$.

Exercise 11.7.1 *Prove Theorem 11.7.4. Hint. Apply Dynkin's formula to $V(x_t)e^{\alpha t}$; note $\mathcal{A}(V(x_t)e^{\alpha t}) = \alpha V(x_t)e^{\alpha t} + e^{\alpha t}\mathcal{A}V(x_t)$ and use (11.67).*

11.8 The Wong-Zakai Theorem and Robustness of the Stratonovich Integral

The Brownian noise is an idealization and is not practically achievable. However, it is approximated arbitrarily well by signals which are sufficiently regular, so that with these regular approximations one can define a Riemann integration. This then raises a question of robustness and convergence of approximate integrations. The following establishes such an approximation result.

Let $\{w_n(t), n \in \mathbb{N}\}$ be a sequence of continuous and piece-wise differentiable in t with bounded variation approximations which converge almost surely to a Brownian process, such that there exist random variables n_0, k so that $w_n(t, \omega) \leq k(\omega)$ almost surely and all $t \in [a, b]$ when $n > n_0(\omega)$, and that $w_n(t)$ converges to B_t almost surely. In particular, we ask that $\{w_n(t), n \in \mathbb{N}\}$ be a sequence of continuous and piece-wise differentiable in t approximations which converges uniformly almost surely to a Brownian process.

Theorem 11.8.1 *[333] Let $\sigma(t, x)$ be continuously differentiable in x, t, and let $\{w_n(t), n \in \mathbb{N}\}$ be a sequence of approximations as discussed above of a Brownian process. Then, we have that, almost surely*

$$\lim_{n\to\infty} \int_a^b \sigma(t, w_n(t)) dw_n(t) = \frac{1}{2} \int_a^b \frac{\partial}{\partial t}\sigma(t, B_t)dt + \int \sigma(t, B_t)dB_t$$

where the first two integrations are in the Riemann sense.

Theorem 11.8.2 (Wong-Zakai Theorem) *[333] Let $\mu(t, x), \sigma(t, x)$ be continuously differentiable in x, t and $\mu, \sigma, \frac{\partial}{\partial x}\sigma$ be Lipschitz continuous with constant $k > 0$, and $|\sigma(x, t)| \geq \beta > 0$ with $\frac{\partial}{\partial t}\sigma(x, t) \leq k\sigma^2(x, t)$. Let $\{w_n(t), n \in \mathbb{N}\}$ be a sequence of approximations as discussed above of a Brownian process.*

If for each $n \in \mathbb{N}$, x_n is the solution to the ODE:

$$\frac{dx_n}{dt} = \mu(t, x_n) - \frac{1}{2}\sigma(t, x_n)\frac{\partial}{\partial x}\sigma(t, x_n) + \sigma(t, x_n)\frac{dw_n}{dt}, \quad x_n(a) = x_a,$$

almost everywhere on $[a, b]$, then $x_n(t)$ converges almost surely uniformly in $t \in [a, b]$ to a stochastic process X_t solving the equation

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dB_t, \quad X_a = x_a,$$

as $n \to \infty$.

Note that here, unlike the discrete-time approximations, we are approximating the noise as well. One way to approximate the noise is via a piece-wise linear interpolation of discrete updates:

$$w_n((k + 1)h) = w_n(kh) + \sqrt{h}Z_k,$$

where Z_k is an independent Gaussian with mean zero and variance 1. In this case, notice that $\frac{dw_n}{dt} = \frac{Z}{\sqrt{h}}$ in between the sampling instants.

Note also that in the above, we have the correction term $-\frac{1}{2}\sigma(t,x_n)\frac{\partial}{\partial x}\sigma(t,x_n)$, which disappears when one considers instead of Itô, the Stratonovich integral [175, Theorem 1.2] (see Exercise 11.10.14). With the above, we have the following. Consider

$$dx^n = f(x^n)dw_n$$

where the integration is in the Riemann sense. If $w_n(t) \to B(t)$ as above, then the solution converges to

$$dx = f(x) \circ dB,$$

where the integration is in the Stratonovich sense.

This last observation is yet another motivation for using the Stratonovich integral for certain applications. For related results with control, see [204] and [255].

11.9 Bibliographic Notes

For discounted and average cost problems, analysis based on the theory of partial differential equations can be utilized to obtain more general results [15, Chapter 3]. The regularity conditions on the value function can also be relaxed. For stochastic integration, one can also relax conditions on the functions via Krylov's generalization [198].

For filtering theory, the reader is referred to [204] as well as [25, 239].

11.10 Exercises

Exercise 11.10.1 *a) Solve*

$$dX_t = \mu X_t + \sigma dB_t$$

Hint: Multiply both sides with the integrating factor $e^{-\mu t}$ and work with $d(e^{-\mu t}X_t)$.

b) Solve

$$dX_t = \mu dt + \sigma X_t dB_t$$

Hint: Multiply both sides with the integrating factor $e^{-\sigma B_t + \frac{1}{2}\sigma^2 t}$. Finally, verify by direct computation (via Itô's formula) that

$$X_t = X(0)e^{\sigma B_t + (\mu - \frac{\sigma^2}{2})t}$$

is the solution. The equation in b) above is often used as a model for mathematical finance where μ is called the drift and σ is called the volatility (of the financial environment).

Exercise 11.10.2 (Girsanov's measure transformation / static reduction) *a) Consider*

$$S_n = \sum_{k=1}^{n} X_k$$

where $X_k \sim \mathcal{N}(0,1)$ is i.i.d. Since a sum of Gaussians is Gaussian, $(S_1, \cdots, S_n)$ is a Gaussian random vector with measure, say with measure Q_0, and density

$$C_n e^{-\frac{1}{2}\left(s_1^2 + (s_2 - s_1)^2 + \cdots + (s_n - s_{n-1})^2\right)}$$

for some constant C_n. Now, instead, assume that $S'_n = \sum_{k=1}^n (X_k + \mu_k)$ where μ_k is a sequence of constants. In this case, $(S'_1, \cdots, S'_n)$ is a Gaussian random vector with measure Q and density, for some constant C'_n,

$$C'_n e^{-\frac{1}{2}\left((s_1-\mu_1)^2+(s_2-s_1-\mu_2)^2+\cdots+(s_n-s_{n-1}-\mu_n)^2\right)}$$
$$= C'_n e^{-\frac{1}{2}(s_1^2+(s_2-s_1)^2+\cdots+(s_n-s_{n-1})^2)} e^{(\sum_k \mu_k(S_k-S_{k-1})-\frac{1}{2}\sum_k \mu_k^2)}$$

Thus, the Radon-Nikodym derivative is obtained as

$$\frac{dQ}{dQ_0} = e^{(\sum_k \mu_k(S_k-S_{k-1})-\frac{1}{2}\sum_k \mu_k^2)} \tag{11.68}$$

This is the same derivation we studied in Section 10.4.2.

*With this interpretation, consider the measurement process given in (**??-??**) with $dy_t = h(x_t)dt + dB_t$ and let P be the measure on this process. Let P_0 denote the measure on an independent Brownian motion $dy'_t = dB_t$. Now, by viewing μ_k as the drift term $h(x_t)dt$, compare (11.68) with*

$$Z_T = \exp\left[\int_0^T h(x_s)dy_s - \frac{1}{2}\int_0^T |h(x_s)|^2 ds\right].$$

where

$$\frac{dP}{dP_0} = Z_T.$$

b) Repeat the above with

$$S_n = S_{n-1} + \sigma(S_n)X_n \sim Q_0$$

where $\sigma(\cdot)$ is invertible and $S'_n = S_{n-1} + \mu_k + \sigma(S'_n)X_n \sim Q$ and show that in this case, we have

$$\frac{dQ}{dQ_0} = e^{(\sum_k \sigma^{-1}(S_{k-1})\mu_k(X_k)-\frac{1}{2}\sum_k (\sigma^{-1}(S_{k-1}))^2\mu_k^2)} \tag{11.69}$$

In the context of a diffusion process, $dx_t = h(x_t)dt + \sigma(x_t)dB_t \sim P$ and $dx'_t = \sigma(x'_t)dB_t \sim P_0$ compare the above with

$$\frac{dP}{dP_0} =: Z_T = \exp\left[\int_0^T \sigma^{-1}(x_s)h(x_s)dB_s - \frac{1}{2}\int_0^T |\sigma^{-1}(x_s)h(x_s)|^2 ds\right].$$

c) In part a) if x_t is a controlled process and in part b) if h takes in control as an input, the above measure transformations then lead to stochastic analysis where one can study the problem in a new probability space where the control does not impact the flow of the process x'_t or B_t, as in static reduction, but it is dependent on them. This approach facilitates various continuity, compactness and approximation results, leading to very general optimality results.

Exercise 11.10.3 (Feynman-Kac Formula: Expected hitting time to a boundary) *Let $S \subset \mathbb{R}^n$ be a bounded open set with smooth boundary ∂S. The following partial differential equation for the notation of the Laplacian of a function f given with $\Delta f := \sum_i \frac{\partial^2 f}{\partial (x^i)^2} f(x)$)*

$$-\frac{1}{2}\Delta u = 1, \quad u \in S$$
$$u = 0 \quad u \in \partial S$$

is known to admit a solution $u(x)$. Now, for any initial point $x \in S$, consider the Brownian motion B_t with $B_0 = x$. Define

$$\tau_S^x = \inf\{t \geq 0 : B_t \in \partial S\}$$

Show that

$$u(x) = E[\tau_S^x]$$

Hint. For the equation $X_t = B_t$, the generator satisfies the relation $\mathcal{A}(f) = \frac{1}{2}\Delta f$. Then, with $\min(N, \tau_S^x) = \tau^N$ via Dynkin's formula

$$E[u(X_{\tau^N})] = E[u(X_0)] + E[\int_0^{\tau^N} \frac{1}{2}\Delta u(X_s)ds]$$

Since $-\frac{1}{2}\Delta u = 1$ until the stopping time and u is bounded, we have that $\lim_{N\to\infty} E[\tau^N]$ is bounded and τ_S^x has finite expectation. As a result,

$$u(x) = E[u(X_0)|X_0 = x] = -E_x[\int_0^{\tau^N} \frac{1}{2}\Delta u(X_s)ds] + E_x[u(X_{\tau^N})] \to_{N\to\infty} E_x[\tau_S^x]$$

Finally, conclude with observing that for $x \in \partial S$ $u(x) = 0$ (and by the above for x inside S, $\Delta u(x) = -1$).

Exercise 11.10.4 (White Noise Property of the Brownian Noise) *Let us view/define the Fourier transform of dB_t to be*[3] *defined sample path wise:*

$$a_k(\omega) = \int_0^1 dB_t(\omega)e^{-i2\pi kt}dt.$$

Show that $a_k, k \in \mathbb{Z}$ is Gaussian, and i.i.d.

Exercise 11.10.5 *Prove the Itô isometry property.*

Exercise 11.10.6 *Complete the details for the solution to the optimal portfolio selection problem given in Example 11.13.*

Exercise 11.10.7 *Solve an average-cost version of the linear quadratic regulator problem and identify conditions on the cost function that leads to a cost that is independent of the initial condition.*

Exercise 11.10.8 *Consider a Brownian process in $\mathbb{R}^d$. Show that this process is recurrent for $d = 1, 2$ but transient for $d \geq 3$.*

Exercise 11.10.9 *Construct an example of a control-free stochastic differential equation with additive Brownian noise such that, while in the absence of the noise term the (deterministic) process is unstable, the presence of the noise makes the system stochastically stable.*

Exercise 11.10.10 (Finite state continuous-time Markov processes and discrete-time representation via uniformization) *Let X_t be a time-homogenous continuous-time Markov process with a finite state space $\mathbb{X}$. Let for any $a, b \in \mathbb{X}$ and $t \geq 0$:*

$$P(X_t = b|X_0 = x) =: P_t(x, b),$$

denote the continuous-time transition kernel. Define

$$\lim_{t\downarrow 0} \frac{P_t - P_0}{h} =: \Lambda$$

to be a transition rate intensity matrix. Define $P(X_t \in \cdot) = \mu_t(\cdot) = (\mu_0 P_t)(\cdot)$. Then, it follows (show this as an exercise) that

[3] This integral is a well-defined Riemann-Stieltjes integral via the Young integral [339]. Note that this is not a contradiction with Theorem 11.2.1 since what Theorem 11.2.1 implies is that the Riemann-Stieltjes integral may not be well-defined for an arbitrary function, but if the function is sufficiently regular one can still define a Riemann-Stieltjes integration with respect to the Brownian for a given sample path, as is the case here. See [339].

$$\frac{d\mu_t}{dt} = \lim_{h \downarrow 0} \frac{P_{t+h} - P_t}{h} = \mu_t \Lambda$$

and thus

$$\mu_t = \mu_0 e^{\Lambda t}$$

Now, define

$$\lambda = \max_{i \in \mathbb{X}} \sum_{j \in \mathbb{X}, \neq i} \lambda_{i,j}$$

and define

$$R = I + \frac{\Lambda}{\lambda}$$

It can be shown that R is a (discrete-time) stochastic matrix since the row-sum of Λ is zero.

Show that, if Z_t is a discrete-time time-homogenous Markov chain with (discrete-time) transition kernel R, then

$$P(Z_t = \cdot) = P(X_{N_t} = \cdot),$$

where N_t is an independent Poisson process with rate λ.

Hint. Show that

$$\mu_t = \mu_0 e^{\Lambda t} = \mu_0 e^{\lambda R t} e^{-\lambda t} = \sum_{k \geq 0} \frac{R^k t^k}{k!} e^{-\lambda t} = \sum_{k \geq 0} \left(\frac{t^k}{k!} e^{-\lambda t} \right) R^k$$

Then, observe that $P(N_t = k) = \left(\frac{\lambda^k t^k}{k!} e^{-\lambda t} \right)$ and that N_t is independent from X_t, to conclude that $\mu_t = P(X_{N_t} \in \cdot)$.

Exercise 11.10.11 *For a system with (11.9), arrive at (11.10) for the case where p is invariant. This relation is important for the study of stochastic learning/simulated annealing problems [84].*

Exercise 11.10.12 *Read [218, Chapters 4 and 5] to study the maximum principle and viscosity solutions for the deterministic-setup.*

Exercise 11.10.13 *Consider the diffusion process*

$$dX_t = \nabla f(X_t) dt + \sqrt{2} dB_t.$$

Show, via the Fokker-Planck equation, that $p(x) = K e^{f(x)}$ is an invariant probability measure where K is a normalizing constant.

Exercise 11.10.14 *[Itô vs. Stratonovich Equivalence] Consider*

$$dX_t = b(X_t, U_t) dt + \sigma(X_t) \circ dB_t$$

By considering the construction of the Stratonovich integral, study that [175, Theorem 1.2]

$$\begin{aligned} \sigma(X_t) \circ dB_t &= \sigma(X(t)) dB_t + \frac{1}{2} d\sigma(X(t)) dB_t \\ &= \sigma(X(t)) dB_t + \frac{1}{2} (\frac{d}{dx} \sigma(X(t))) dX(t) dB_t = \sigma(X(t)) dB_t + \frac{1}{2} (\frac{d}{dx} \sigma(X(t))) \sigma(X_t) dt \end{aligned} \tag{11.70}$$

Thus, arrive at an equivalence relation between the two integrals, so that we have

$$dX_t = \tilde{b}(X_t, U_t) dt + \sigma(X_t) dB_t,$$

where the integral is now in the Itô sense with

$$\tilde{b}(X_t, U_t) = b(X_t, U_t) + \frac{1}{2}(\frac{d}{dx}\sigma(X(t)))\sigma(X_t)$$

12

Robustness to Incorrect Models and Learning

In many applications, typically only an ideal model (controlled transition kernel) is assumed and the control design is based on the given model, raising the problem of performance loss due to the mismatch between the assumed model and the actual model. In this chapter, we study continuity properties of discrete-time stochastic control problems with respect to system models (i.e., controlled transition kernels) and robustness of optimal control policies designed for incorrect models applied to the true system.

The chapter studies both fully observed and partially observed setups under an infinite horizon discounted expected cost criterion as well as the average cost criterion. The results for the discounted cost criterion will also imply the identical results for finite horizon criteria.

We will show that continuity can be established under total variation convergence of the transition kernels under mild assumptions and with further restrictions on the dynamics and observation model under weak and setwise convergence of the transition kernels. Using these continuity properties, we establish convergence results and error bounds due to mismatch that occurs by the application of a control policy which is designed for an incorrectly estimated system model to a true model, thus establishing positive and negative results on robustness. These findings entail positive implications on empirical learning in (data-driven) stochastic control since often system models are learned through empirical training data where typically a weak convergence criterion applies but stronger convergence criteria do not.

The chapter can also be viewed as a generalization of the approximations framework presented in Section 8.2. This connection will be made explicit later in the chapter.

12.1 Introduction

We will discuss both the partially and fully observed setups. Let $\mathbb{X} \subset \mathbb{R}^m$ denote a Borel set which is the state space of a partially observed controlled Markov process. Let $\mathbb{Y} \subset \mathbb{R}^n$ be a Borel set denoting the observation space of the model, and let the state be observed through an observation channel Q. As before in the notes, the observation channel, Q, is defined as a stochastic kernel (regular conditional probability) from $\mathbb{X}$ to $\mathbb{Y}$, such that $Q(\,\cdot\,|x)$ is a probability measure on the (Borel) σ-algebra $\mathcal{B}(\mathbb{Y})$ of $\mathbb{Y}$ for every $x \in \mathbb{X}$, and $Q(A|\,\cdot\,) : \mathbb{X} \to [0,1]$ is a Borel measurable function for every $A \in \mathcal{B}(\mathbb{Y})$. A decision maker (DM) is located at the output of the channel Q, and hence it only sees the observations $\{Y_t,\, t \in \mathbb{Z}_+\}$ and chooses its actions from $\mathbb{U}$, the action space which is a Borel subset of some Euclidean space. As discussed earlier, an *admissible policy* γ is a sequence of control functions $\{\gamma_t,\, t \in \mathbb{Z}_+\}$ such that γ_t is measurable with respect to the σ-algebra generated by the information variables

$$I_t = \{Y_{[0,t]}, U_{[0,t-1]}\}, \quad t \in \mathbb{N}, \qquad I_0 = \{Y_0\},$$

where

$$U_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_+, \tag{12.1}$$

are the $\mathbb{U}$-valued control actions and

$$Y_{[0,t]} = \{Y_s,\, 0 \le s \le t\}, \quad U_{[0,t-1]} = \{U_s,\, 0 \le s \le t-1\}.$$

We define Γ to be the set of all such admissible policies. The update rules of the system are determined by (12.1) and the following:

$$\Pr\big((X_0, Y_0) \in B\big) = \int_B P(dx_0)Q(dy_0|x_0), \quad B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}),$$

where P is the (prior) distribution of the initial state X_0, and

$$\begin{aligned}&\Pr\Big((X_t, Y_t) \in B \,\Big|\, (X, Y, U)_{[0,t-1]} = (x, y, u)_{[0,t-1]}\Big)\\ &= \int_B \mathcal{T}(dx_t|x_{t-1}, u_{t-1})Q(dy_t|x_t), B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}), t \in \mathbb{N},\end{aligned}$$

where $\mathcal{T}$ is the transition kernel of the model. The objective of the agent (decision maker) is the minimization of the infinite horizon discounted cost,

$$J_\beta(c, \mathcal{T}, \gamma) = E_P^{\mathcal{T},\gamma}\left[\sum_{t=0}^{\infty} \beta^t c(X_t, U_t)\right]$$

for some discount factor $\beta \in (0, 1)$, over the set of admissible policies $\gamma \in \Gamma$, where $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is a Borel-measurable stage-wise cost function and $E_P^{\mathcal{T},\gamma}$ denotes the expectation with initial state probability measure P and transition kernel $\mathcal{T}$ under policy γ. Note that we write the infinite horizon discounted cost as a function of the transition kernels and the stage-wise cost function since we will analyze the cost under the changes on those variables.

We define the optimal cost for the discounted infinite horizon setup as a function of the stage-wise cost function and the transition kernels as

$$J_\beta^*(c, \mathcal{T}) = \inf_{\gamma \in \Gamma} J_\beta(c, \mathcal{T}, \gamma).$$

Problem P1: Continuity of $J_\beta^*(c, \mathcal{T})$ under the Convergence of the Models. Let $\{\mathcal{T}_n, n \in \mathbb{N}\}$ be a sequence of transition kernels which converges in some sense to another transition kernel $\mathcal{T}$ and $\{c_n, n \in \mathbb{N}\}$ be a sequence of stage-wise cost functions corresponding to $\mathcal{T}_n$ which converge in some sense to another cost function c. Does that imply that

$$J_\beta^*(c_n, \mathcal{T}_n) \to J_\beta^*(c, \mathcal{T})?$$

Problem P2: Robustness to Incorrect Models. A problem of major practical importance is robustness of an optimal controller to modeling errors. Suppose that an optimal policy is constructed according to a model which is incorrect: how does the application of the control to the true model affect the system performance and does the error decrease to zero as the models become closer to each other? In particular, suppose that γ_n^* is an optimal policy designed for $\mathcal{T}_n$ and c_n, an incorrect model for a true model $\mathcal{T}$ and c. Is it the case that if $\mathcal{T}_n \to \mathcal{T}$ and $c_n \to c$, then

$$J_\beta(c, \mathcal{T}, \gamma_n^*) \to J_\beta^*(c, \mathcal{T})?$$

Problem P3: Empirical Consistency of Learned Probabilistic Models and Data-Driven Stochastic Control. Let $\mathcal{T}(\cdot|x, u)$ be a transition kernel given previous state and action variables $x \in \mathbb{X}, u \in \mathbb{U}$, which is unknown to the decision maker (DM). Suppose the DM builds a model for the transition kernels, $\mathcal{T}_n(\cdot|x, u)$, for all possible $x \in \mathbb{X}, u \in \mathbb{U}$ by collecting training data (e.g. from the evolving system). Do we have that the cost calculated under $\mathcal{T}_n$ converges to the true cost (i.e., do we have that the cost obtained from applying the optimal policy for the empirical model converges to the true cost as the training length increases)?

Problem P4: Approximation by Finite MDPs as an Instance of Robustness to Incorrect Models. Can we view the approximation problem of a continuous space MDP model with a finite model, as studied in Section 8.2, as an instance of the robustness problem?

We will study the above for the average cost criterion as well. For the average cost criterion, we have

$$J_\infty(c,\mathcal{T},\gamma)=\limsup_{N\to\infty}\frac{1}{N}E_{x_0}^{\mathcal{T},\gamma}\left[\sum_{t=0}^{N-1}c(X_t,U_t)\right]$$

To denote the explicit dependence of the optimal cost in the transition kernel, we use the notation

$$J_\infty^*(c,\mathcal{T})=\inf_{\gamma\in\Gamma}J_\infty(c,\mathcal{T},\gamma).$$

12.1.1 Some Examples and Convergence Criteria for Transition Kernels

Convergence Criteria for Transition Kernels.

Before introducing the convergence criteria to be presented in the chapter, we refer the reader to Appendix D.

Definition 12.1.1 *For a sequence of transition kernels $\{\mathcal{T}_n, n\in\mathbb{N}\}$, we say that*

- *$\mathcal{T}_n\to\mathcal{T}$ weakly if $\mathcal{T}_n(\cdot|x,u)\to\mathcal{T}(\cdot|x,u)$ weakly, for all $x\in\mathbb{X}$ and $u\in\mathbb{U}$,*
- *$\mathcal{T}_n\to\mathcal{T}$ setwise if $\mathcal{T}_n(\cdot|x,u)\to\mathcal{T}(\cdot|x,u)$ setwise, for all $x\in\mathbb{X}$ and $u\in\mathbb{U}$,*
- *$\mathcal{T}_n\to\mathcal{T}$ under the total variation distance if $\mathcal{T}_n(\cdot|x,u)\to\mathcal{T}(\cdot|x,u)$ under total variation for all $x\in\mathbb{X}$ and $u\in\mathbb{U}$.*

Examples [187].

Let a controlled model be given as $x_{t+1}=F(x_t,u_t,w_t)$, where $\{w_t\}$ is an i.i.d. noise process. The uncertainty on the transition kernel for such a system may arise from lack of information on F or the i.i.d. noise process w_t or both:

(i) Let $\{F_n\}$ denote an approximating sequence for F, so that $F_n(x,u,w)\to F(x,u,w)$ pointwise. Assume that the probability measure of the noise is known. Then, corresponding kernels $\mathcal{T}_n$ converge weakly to $\mathcal{T}$: If we denote the probability measure of w with μ, for any $g\in C_b(\mathbb{X})$ and for any $(x_0,u_0)\in\mathbb{X}\times\mathbb{U}$ using the dominated convergence theorem we have

$$\lim_{n\to\infty}\int g(x_1)\mathcal{T}_n(dx_1|x_0,u_0)=\lim_{n\to\infty}\int g(F_n(x_0,u_0,w))\mu(dw)$$
$$=\int g(F(x_0,u_0,w))\mu(dw)=\int g(x_1)\mathcal{T}(dx_1|x_0,u_0).$$

(ii) Much of the robust control literature deals with deterministic systems where the nominal model is a deterministic perturbation of the actual model (see e.g. [281]). The considered model is in the following form: $\tilde{F}(x_t,u_t)=F(x_t,u_t)$
$+\Delta F(x_t,u_t)$, where F represents the nominal model and ΔF is the model uncertainty satisfying some norm bounds. For such deterministic systems, pointwise convergence of $\tilde{F}$ to the nominal model F, i.e. $\Delta F(x_t,u_t)\to 0$, can be viewed as weak convergence for deterministic systems by the discussion in (i). It is evident, however, that total variation convergence would be too strong for such a convergence criterion, since $\delta_{\tilde{F}(x_t,u_t)}\to\delta_{F(x_t,u_t)}$ weakly but $\|\delta_{\tilde{F}(x_t,u_t)}-\delta_{F(x_t,u_t)}\|_{TV}=2$ for all $\Delta F(x_t,u_t)\neq 0$ where δ denotes the Dirac measure.

(iii) Let $F(x_t,u_t,w_t)=f(x_t,u_t)+w_t$ be such that the function f is known and $w_t\sim\mu$ is not known correctly and an incorrect model μ_n is assumed. If $\mu_n\to\mu$ weakly, setwise, or in total variation, then the corresponding transition kernels $\mathcal{T}_n$ converge in the same sense to $\mathcal{T}$. Observe the following:

$$\int g(x_1)\mathcal{T}_n(dx_1|x_0,u_0)-\int g(x_1)\mathcal{T}(dx_1|x_0,u_0)$$

$$= \int g(w_0 + f(x_0, u_0))\mu_n(dw_0) - \int g(w_0 + f(x_0, u_0))\mu(dw_0). \tag{12.2}$$

(a) Suppose $\mu_n \to \mu$ weakly. If g is a continuous and bounded function, then $g(\cdot + f(x_0, u_0))$ is a continuous and bounded function for all $(x_0, u_0) \in \mathbb{X} \times \mathbb{U}$. Thus, (12.2) goes to 0. Note that f does not need to be continuous. (b) Suppose $\mu_n \to \mu$ setwise. If g is a measurable and bounded function, then $g(\cdot + f(x_0, u_0))$ measurable and bounded for all $(x_0, u_0) \in \mathbb{X} \times \mathbb{U}$. Thus, (12.2) goes to 0. (c) Finally, assume $\mu_n \to \mu$ in total variation. If g is bounded, (12.2) converges to 0, as in item (b). As a special case, assume that μ_n and μ admit densities h_n and h, respectively; then the pointwise convergence of h_n to h implies the convergence of μ_n to μ in total variation by Scheffé's theorem.

(iv) Suppose now neither F nor the probability model of w_t is known perfectly. It is assumed that w_t admits a measure μ_n and $\mu_n \to \mu$ weakly. For the function F we again have an approximating sequence $\{F_n\}$. If $F_n(x, u, w_n) \to F(x, u, w)$ for all $(x, u) \in \mathbb{X} \times \mathbb{U}$ and for any $w_n \to w$, then the transition kernel $\mathcal{T}_n$ corresponding to the model F_n converges weakly to the one of F, $\mathcal{T}$: For any $g \in C_b(\mathbb{X})$,

$$\lim_{n\to\infty} \int g(x_1)\mathcal{T}_n(dx_1|x_0, u_0) = \lim_{n\to\infty} \int g(F_n(x_0, u_0, w))\mu_n(dw)$$
$$= \int g(F(x_0, u_0, w))\mu(dw) = \int g(x_1)\mathcal{T}(dx_1|x_0, u_0).$$

(v) Let again $\{F_n\}$ denote an approximating sequence for F and suppose now $F_{x_0,u_0,n}(\cdot) := F_n(x_0, u_0, \cdot) : \mathbb{W} \to \mathbb{X}$ is invertible for all $x_0, u_0 \in \mathbb{X} \times \mathbb{U}$ and $F^{-1}_{(x_0,u_0),n}(\cdot)$ denotes the inverse for fixed (x_0, u_0). It is assumed that $F^{-1}_{(x_0,u_0),n}(x_1) \to F^{-1}_{x_0,u_0}(x_1)$ pointwise for all (x_0, u_0). Suppose further that the noise process w_t admits a continuous density $f_W(w)$. The Jacobian matrix, $\frac{\partial x_1}{\partial w}$, is the matrix whose components are the partial derivatives of x_1, i.e. with $x_1 \in \mathbb{X} \subset \mathbb{R}^m$ and $w \in \mathbb{W} \subset \mathbb{R}^m$, it is an $m \times m$ matrix with components $\frac{\partial (x_1)_i}{\partial w_j}$, $1 \leq i, j \leq m$. If the Jacobian matrix of derivatives $\frac{\partial x_1}{\partial w}(w)$ is continuous in w and nonsingular for all w, then we have that the density of the state variables can be written as

$$f_{X_1,n,(x_0,u_0)}(x_1) = f_W(F^{-1}_{x_0,u_0,n}(x_1))\Big|\frac{\partial x_1}{\partial w}(F^{-1}_{x_0,u_0,n}(x_1))\Big|^{-1},$$
$$f_{X_1,(x_0,u_0)}(x_1) = f_W(F^{-1}_{x_0,u_0}(x_1))\Big|\frac{\partial x_1}{\partial w}(F^{-1}_{x_0,u_0}(x_1))\Big|^{-1}.$$

With the above, $f_{X_1,n,(x_0,u_0)}(x_1) \to f_{X_1,(x_0,u_0)}(x_1)$ pointwise for all fixed (x_0, u_0). Therefore, by Scheffé's theorem, the corresponding kernels $\mathcal{T}_n(\cdot|x_0, u_0) \to \mathcal{T}(\cdot|x_0, u_0)$ in total variation for all (x_0, u_0).

(vi) These examples will be utilized in Section 12.5.1, where data-driven stochastic control problems will be considered where estimated models are obtained through empirical measurements of the state action variables.

12.1.2 Summary

We now introduce the main assumptions that will be occasionally used for the technical results in the chapter.

Assumption 12.1.1 *The following hold.*

(a) The sequence of transition kernels $\mathcal{T}_n$ satisfies the following: $\{\mathcal{T}_n(\cdot|x_n, u_n), n \in \mathbb{N}\}$ converges weakly to $\mathcal{T}(\cdot|x, u)$ for any sequence $\{x_n, u_n\} \subset \mathbb{X} \times \mathbb{U}$ and $x, u \in \mathbb{X} \times \mathbb{U}$ such that $(x_n, u_n) \to (x, u)$.

(b) The stochastic kernel $\mathcal{T}(\cdot|x, u)$ is weakly continuous in (x, u).

(c) The sequence of stage-wise cost functions c_n satisfies the following: $c_n(x_n, u_n) \to c(x, u)$ for any sequence $\{x_n, u_n\} \subset \mathbb{X} \times \mathbb{U}$ and $x, u \in \mathbb{X} \times \mathbb{U}$ such that $(x_n, u_n) \to (x, u)$.

(d) The stage-wise cost function $c(x, u)$ is non-negative, bounded, and continuous on $\mathbb{X} \times \mathbb{U}$.

(e) $\mathbb{U}$ *is compact.*

The following is Assumption 6.3.1(ii).

Assumption 12.1.2 *The observation channel* $Q(\cdot|x)$ *is continuous in total variation i.e., if* $x_n \to x$*, then* $Q(\cdot\,|x_n) \to Q(\cdot\,|x)$ *in total variation (only for partially observed models).*

Assumption 12.1.3 *The following hold.*

(a) The sequence of transition kernels $\mathcal{T}_n$ *satisfies the following:* $\{\mathcal{T}_n(\cdot|x,u_n), n \in \mathbb{N}\}$ *converges setwise to* $\mathcal{T}(\cdot|x,u)$ *for any sequence* $\{u_n\} \subset \mathbb{U}$ *and* $x, u \in \mathbb{X} \times \mathbb{U}$ *such that* $u_n \to u$*.*

(b) The stochastic kernel $\mathcal{T}(\cdot|x,u)$ *is setwise continuous in* u*.*

(c) The sequence of stage-wise cost functions c_n *satisfies the following:* $c_n(x,u_n) \to c(x,u)$ *for any sequence* $\{u_n\} \subset \mathbb{U}$ *and* $x, u \in \mathbb{X} \times \mathbb{U}$ *such that* $u_n \to u$*.*

(d) The stage-wise cost function $c(x,u)$ *is non-negative, bounded, and continuous on* $\mathbb{U}$*.*

(e) $\mathbb{U}$ *is compact.*

Assumption 12.1.4 *The following hold.*

(a) The sequence of transition kernels $\mathcal{T}_n$ *satisfies the following:* $\|\mathcal{T}_n(\cdot|x,u_n) - \mathcal{T}(\cdot|x,u)\|_{TV} \to 0$ *for any sequence* $\{u_n\} \subset \mathbb{U}$ *and* $x, u \in \mathbb{X} \times \mathbb{U}$ *such that* $u_n \to u$*.*

(b) The stochastic kernel $\mathcal{T}(\cdot|x,u)$ *is continuous in total variation in* u*.*

(c) The sequence of stage-wise cost functions c_n *satisfies the following:* $c_n(x,u_n) \to c(x,u)$ *for any sequence* $\{u_n\} \subset \mathbb{U}$ *and* $x, u \in \mathbb{X} \times \mathbb{U}$ *such that* $u_n \to u$*.*

(d) The stage-wise cost function $c(x,u)$ *is non-negative, bounded, and continuous on* $\mathbb{U}$*.*

(e) $\mathbb{U}$ *is compact.*

In Sections 12.2.1 and 12.2.2 we study continuity (Problem P1) and robustness (Problem P2) for partially observed models. In particular we show the following:

(a) Continuity and robustness do not hold in general under weak convergence of kernels (Theorem 12.2.1).

(b) Under Assumptions 12.1.1 and 12.1.2, continuity and robustness hold (Theorem 12.2.4, Theorem 12.2.8).

(c) Continuity and robustness do not hold in general under setwise convergence of the kernels (Theorem 12.2.5).

(d) Continuity and robustness do not hold in general under total variation convergence of the kernels (Example 12.1).

(e) Under Assumption 12.1.4, continuity and robustness hold (Theorem 12.2.6, Theorem 12.2.7).

In Section 12.3, we study continuity (Problem P1) and robustness (Problem P2) for fully observed models. In particular we show the following

(a) Continuity and robustness do not hold in general under weak convergence of kernels (Theorem 12.3.1, Example 12.1).

(b) Under Assumption 12.1.1, continuity holds (Theorem 12.3.2), under Assumption 12.1.1, robustness holds if the optimal policies for every initial point are identical (Theorem 12.3.3).

(c) Continuity and robustness do not hold in general under setwise convergence of the kernels (Theorem 12.3.4, Theorem 12.3.6).

(d) Under Assumption 12.1.3, continuity holds (Theorem 12.3.5), and under Assumption 12.1.3, robustness holds if the optimal policies for every initial point are identical (Theorem 12.3.7).

(e) Continuity and robustness do not hold in general under total variation convergence of the kernels (Example 12.1).

(f) Under Assumption 12.1.4, continuity and robustness hold (subsection 12.3.3).

In Section 12.5, we study applications to empirical learning (in Section 12.5.1) where we establish the positive relevance of Theorem 12.3.2, and then applications to finite model approximations under the perspective of robustness in in Section 12.5.2. Here, we restrict the analysis to the case with weakly continuous kernels.

12.2 Continuity and Robustness of Optimal Cost in Convergence of Models (POMDP Case)

12.2.1 Continuity of Optimal Cost in Convergence of Models (POMDP Case)

We now study continuity of the optimal discounted cost under the convergence of transition kernels and cost functions.

Weak Convergence

Absence of Continuity under Weak Convergence. The following shows that the optimal cost may not be continuous under weak convergence of transition kernels.

Theorem 12.2.1 *[187]. Let $\mathcal{T}_n \to \mathcal{T}$ weakly, then it is not necessarily true that $J^*_\beta(c,\mathcal{T}_n) \to J^*_\beta(c,\mathcal{T})$ even when the prior distributions are the same, the measurement channel Q is continuous in total variation, and $c(x,u)$ is continuous and bounded on $\mathbb{X}\times\mathbb{U}$.*

We prove the result with a counterexample [187]. Letting $\mathbb{X}=\mathbb{U}=\mathbb{Y}=[-1,1]$ and $c(x,u)=(x-u)^2$, the observation channel is chosen to be uniformly distributed over [-1,1], $Q\sim U([-1,1])$, the initial distributions of the state variable are chosen to be same as $P\sim\delta_1$, where $\delta_x(A):=1_{\{x\in A\}}$ for Borel A, and the transition kernels are:

$$\begin{aligned}\mathcal{T}(\cdot|x,u)&=\delta_{-1}(x)[\frac{1}{2}\delta_1(\cdot)+\frac{1}{2}\delta_{-1}(\cdot)]+\delta_1(x)[\frac{1}{2}\delta_1(\cdot)+\frac{1}{2}\delta_{-1}(\cdot)]\\&\quad+(1-\delta_{-1}(x))(1-\delta_1(x))\delta_0(\cdot)\\\mathcal{T}_n(\cdot|x,u)&=\delta_{-1}(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot)+\frac{1}{2}\delta_{(-1+1/n)}(\cdot)]+\delta_1(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot)\\&\quad+\frac{1}{2}\delta_{(-1+1/n)}(\cdot)]+(1-\delta_{-1}(x))(1-\delta_1(x))\delta_0(\cdot).\end{aligned}$$

It can be seen that $\mathcal{T}_n\to\mathcal{T}$ weakly according to Definition 12.1.1(i). Note that the cost function is continuous, and the measurement channel is continuous in total variation. The optimal discounted costs can be found as

$$\begin{aligned}J^*_\beta(c,\mathcal{T})&=\sum_{k=1}^{\infty}E_P^{\mathcal{T}}[\beta^kX_k^2]=\sum_{k=1}^{\infty}\beta^k=\frac{\beta}{1-\beta}\\J^*_\beta(c,\mathcal{T}_n)&=\sum_{k=1}^{\infty}E_P^{\mathcal{T}_n}[\beta^kX_k^2]=\beta[\frac{1}{2}(1-\frac{1}{n})^2+\frac{1}{2}(-1+\frac{1}{n})^2].\end{aligned}$$

Then we have $J^*_\beta(c,\mathcal{T}_n)\to\beta\neq\frac{\beta}{1-\beta}$.

A Sufficient Condition for Continuity under Weak Convergence

In the following, we will establish and utilize some regularity properties for the optimal cost with respect to the convergence of transition kernels.

Assumption 12.2.1 *(a) The stochastic kernel $\mathcal{T}(\cdot|x,u)$ is weakly continuous in (x,u), i.e. if $(x_n,u_n)\to(x,u)$, then $\mathcal{T}(\cdot|x_n,u_n)\to\mathcal{T}(\cdot|x,u)$ weakly.*

(b) The observation channel $Q(\cdot|x)$ is continuous in total variation, i.e., if $x_n\to x$, then $Q(\cdot\,|x_n)\to Q(\cdot\,|x)$ in total variation.

(c) The stage-wise cost function $c(x,u)$ is non-negative, bounded and continuous on $\mathbb{X}\times\mathbb{U}$

(d) $\mathbb{U}$ is compact.

As we have see in *Chapter* 6, any POMDP can be reduced to a (completely observable) MDP, whose states are the posterior state distributions or *beliefs* of the observer; that is, the state at time t is $Z_t(\cdot) := \mathsf{Pr}\{X_t\in\cdot\,|Y_0,\ldots,Y_t,U_0,\ldots,U_{t-1}\}\in\mathcal{P}(\mathbb{X})$. We call this equivalent MDP the belief-MDP . The belief-MDP has state space $\mathcal{Z}=\mathcal{P}(\mathbb{X})$ and action space $\mathbb{U}$. Under the topology of weak convergence, since $\mathbb{X}$ is a Borel space, $\mathcal{Z}$ is metrizable with the Prokhorov metric which makes $\mathcal{Z}$ a Borel space [249]. The transition probability η (6.3) of the belief-MDP was earlier constructed through non-linear filtering equations.

The one-stage cost function c of the belief-MDP is given by $\tilde{c}(z,u) := \int_{\mathbb{X}} c(x,u)z(dx)$. Under the regularity of the belief-MDP, we have shown that the *discounted cost optimality operator* $\mathbb{T}: C_b(\mathcal{Z})\to C_b(\mathcal{Z})$

$$(T(f))(z)=\min_u(\tilde{c}(z,u)+\beta E[f(z_1)|z_0=z,u_0=u]) \tag{12.3}$$

is a contraction from $C_b(\mathcal{Z})$ to itself under the supremum norm. As a result, there exists a fixed point, the value function, and an optimal control policy exists. In view of this existence result, in the following we will consider optimal policies.

The following result is key to proving the main result of this section whose detailed analysis can be found in [187].

Theorem 12.2.2 *Suppose we have a uniformly bounded family of functions $\{f_n^\gamma:\mathbb{X}\to\mathbb{R},\gamma\in\mathit{\Gamma},n>0\}$ such that $\|f_n^\gamma\|_\infty<C$ for all $\gamma\in\mathit{\Gamma}$ and for all $n>0$ for some $C<\infty$.*

Further suppose we have another uniformly bounded family of functions $\{f^\gamma:\mathbb{X}\to\mathbb{R},\gamma\in\mathit{\Gamma}\}$ such that $\|f^\gamma\|_\infty<C$ for all $\gamma\in\mathit{\Gamma}$ for some $C<\infty$. Under the following assumptions,

(i) For any $x_n\to x$

$$\sup_{\gamma\in\mathit{\Gamma}}\left|f_n^\gamma(x_n)-f^\gamma(x)\right|\to 0,\quad \sup_{\gamma\in\mathit{\Gamma}}\left|f^\gamma(x_n)-f^\gamma(x)\right|\to 0,$$

(ii) $\sup_\gamma\rho(\mu_n^\gamma,\mu^\gamma)\to 0$ where ρ is some metric for the weak convergence topology,

we have

$$\sup_{\gamma\in\mathit{\Gamma}}\left|\int f_n^\gamma(x)\mu_n^\gamma(dx)-\int f^\gamma(x)\mu^\gamma(dx)\right|\to 0.$$

Theorem 12.2.3 *Under Assumptions 12.1.1 and 12.1.2,*

$$\sup_{\gamma\in\mathit{\Gamma}}|J_\beta(c_n,\mathcal{T}_n,\gamma)-J_\beta(c,\mathcal{T},\gamma)|\to 0.$$

Proof sketch.

$$\begin{aligned}&\sup_{\gamma\in\mathit{\Gamma}}|J_\beta(c_n,\mathcal{T}_n,\gamma)-J_\beta(c,\mathcal{T},\gamma)|\\ =&\sup_{\gamma\in\mathit{\Gamma}}\left|\sum_{t=0}^{\infty}\beta^t\left(E_P^{\mathcal{T}_n}\left[c_n\big(X_t,\gamma(Y_{[0,t]})\big)\right]-E_P^{\mathcal{T}}\left[c\big(X_t,\gamma(Y_{[0,t]})\big)\right]\right)\right|\end{aligned}$$

$$\leq \sum_{t=0}^{\infty} \beta^t \sup_{\gamma\in\Gamma} \left| E_P^{\mathcal{T}_n}\left[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\right] - E_P^{\mathcal{T}}\left[c\big(X_t, \gamma(Y_{[0,t]})\big)\right]\right|.$$

Recall that an *admissible policy* γ is a sequence of control functions $\{\gamma_t,\, t \in \mathbb{Z}_+\}$. In the last step above, we make a slight abuse of notation; the $\sup$ at the first step is over all sequence of control functions $\{\gamma_t,\, t \in \mathbb{Z}_+\}$ whereas the $\sup$ at the last step is over all sequence of control functions $\{\gamma_{t'},\, t' \leq t\}$, but we will use the same notation, γ, in the rest of the proof.

For any $\epsilon > 0$, we choose a $K < \infty$ such that $\sum_{t=K+1}^{\infty} \beta^k 2\|c\|_\infty \leq \epsilon/2$. For the chosen K, we choose an $N < \infty$ such that

$$\sup_{\gamma\in\Gamma} \left| E_P^{\mathcal{T}_n}\left[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\right] - E_P^{\mathcal{T}}\left[c\big(X_t, \gamma(Y_{[0,t]})\big)\right]\right| \leq \epsilon/2K$$

for all $t \leq K$ and for all $n > N$. We note that in [187] a fixed c function was considered, but by considering the additional term

$$\sup_{\gamma\in\Gamma} \left| E_P^{\mathcal{T}_n}\left[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\right] - E_P^{\mathcal{T}}\left[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\right]\right|$$

and noting that $\sup_\gamma |\int Q(dy|x_n)c_n(x_n, \gamma(y)) - \int Q(dy|x)c(x, \gamma(y))| \to 0$, for every $x_n \to x$, by a generalized dominated convergence theorem as Q is continuous in total variation, a triangle inequality argument shows that the same result applies. This follows from a generalized dominated convergence theorem as stated in Theorem 12.2.2 whose detailed analysis can be found in [187]. Thus, $\sup_{\gamma\in\Gamma} \left|J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma) \to 0$ as $n \to \infty$. ◇

Theorem 12.2.4 *[183, 187] Suppose the conditions of Theorem 12.2.3 hold. Then*

$$\lim_{n\to\infty} |J_\beta^*(c_n, \mathcal{T}_n) - J_\beta^*(c, \mathcal{T})| = 0.$$

Proof sketch. We start with the following bound:

$$\begin{aligned} &|J_\beta^*(c_n, \mathcal{T}_n) - J_\beta^*(c, \mathcal{T})| \\ &\leq \max\Big(J_\beta(c_n, \mathcal{T}_n, \gamma^*) - J_\beta(c, \mathcal{T}, \gamma^*), J_\beta(c, \mathcal{T}, \gamma_n^*) - J_\beta(c_n, \mathcal{T}_n, \gamma_n^*)\Big), \end{aligned} \tag{12.4}$$

where γ^* and γ_n^* are the optimal policies, respectively, for $\mathcal{T}$ and $\mathcal{T}_n$. Both terms go to 0 by Theorem 12.2.3. ◇

Absence of Continuity under Setwise Convergence

We now show that continuity of optimal costs may fail under the setwise convergence of transition kernels. Theorem 12.3.4 in the next section establishes this result for fully observed models, which serves as a proof for this setup also.

Theorem 12.2.5 *[183, 187] Let $\mathcal{T}_n \to \mathcal{T}$ setwise. Then, it is not true in general that*

$$J_\beta^*(c, \mathcal{T}_n)$$

$\to J_\beta^*(c, \mathcal{T})$*, even when $\mathbb{X}, \mathbb{Y}$, and $\mathbb{U}$ are compact and $c(x, u)$ is continuous and bounded in $\mathbb{X} \times \mathbb{U}$.*

Continuity under Total Variation

Theorem 12.2.6 *[183, 187] Under Assumption 12.1.4, $J_\beta^*(c_n, \mathcal{T}_n) \to J_\beta^*(c, \mathcal{T})$.*

Proof sketch. We start with the following bound:

$$|J^*_\beta(c_n, \mathcal{T}_n) - J^*_\beta(c, \mathcal{T})| \leq \max\Bigg(J_\beta(c_n, \mathcal{T}_n, \gamma^*) - J_\beta(c, \mathcal{T}, \gamma^*), J_\beta(c_n, \mathcal{T}_n, \gamma^*_n) - J_\beta(c, \mathcal{T}, \gamma^*_n)\Bigg),$$

where γ^* and γ^*_n are the optimal policies, respectively, for $\mathcal{T}$ and $\mathcal{T}_n$.

We now study the following:

$$\begin{aligned}
&\sup_{\gamma\in\Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \\
&= \sup_{\gamma\in\Gamma} \left| \sum_{t=0}^{\infty} \beta^t \left(E_P^{\mathcal{T}_n}\Big[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\Big] - E_P^{\mathcal{T}}\Big[c\big(X_t, \gamma(Y_{[0,t]})\big)\Big]\right)\right| \\
&\leq \sum_{t=0}^{\infty} \beta^t \sup_{\gamma\in\Gamma} \left| E_P^{\mathcal{T}_n}\Big[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\Big] - E_P^{\mathcal{T}}\Big[c\big(X_t, \gamma(Y_{[0,t]})\big)\Big]\right|.
\end{aligned}$$

It can be shown that ([187])

$$\sup_{\gamma\in\Gamma} \left| E_P^{\mathcal{T}_n}\Big[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\Big] - E_P^{\mathcal{T}}\Big[c\big(X_t, \gamma(Y_{[0,t]})\big)\Big]\right| \to 0. \tag{12.5}$$

This was shown in [187] for fixed c. The extension to varying c_n follows from a triangle inequality step with the assumption that $\mathcal{T}_n(\cdot|x, u_n) \to \mathcal{T}(\cdot|x, u)$ setwise, and $c_n(x, u_n) \to c(x, u)$ for any $u_n \to u$. Therefore, using identical steps as in the proof of Theorem 12.2.3 we have $\sup_{\gamma\in\Gamma} \left|J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)\right| \to 0$. ◇

12.2.2 Robustness to Incorrect Models (POMDP Case)

Here, we consider the robustness problem **P2**: Suppose we design an optimal policy, γ^*_n, for a transition kernel, $\mathcal{T}_n$ and a cost function c_n, assuming they are the correct model and apply the policy to the true model whose transition kernel is $\mathcal{T}$ and whose cost function is c. We study the robustness of the sub-optimal policy γ^*_n.

Total Variation

The next theorem gives an asymptotic robustness result.

Theorem 12.2.7 *Under Assumption 12.1.4*

$$|J_\beta(c_n, \mathcal{T}, \gamma^*_n) - J^*_\beta(c, \mathcal{T})| \to 0,$$

where γ^*_n is the optimal policy designed for the kernel $\mathcal{T}_n$.

Proof sketch. We write the following:

$$|J_\beta(c, \mathcal{T}, \gamma^*_n) - J^*_\beta(c, \mathcal{T})| \leq |J_\beta(c, \mathcal{T}, \gamma^*_n) - J^*_\beta(c_n, \mathcal{T}_n)| + |J^*_\beta(c_n, \mathcal{T}_n) - J^*_\beta(c, \mathcal{T})|.$$

Both terms can be shown to go to 0 using (12.5). ◇

Setwise Convergence

Theorem 12.3.6 in the next section establishes the lack of robustness under the setwise convergence of kernels. As we note later, a fully observed system can be viewed as a partially observed system with the measurement being the state itself, (see (12.6)).

Weak Convergence

Theorem 12.2.8 *[183, 187] Under Assumptions 12.1.1 and 12.1.2,* $|J_\beta(c,\mathcal{T},\gamma_n^*) - J_\beta^*(c,\mathcal{T})| \to 0$*, where* γ_n^* *is the optimal policy designed for the transition kernel* $\mathcal{T}_n$*.*

Proof sketch. We write

$$\begin{aligned}|J_\beta(c,\mathcal{T},\gamma_n^*) - J_\beta^*(c,\mathcal{T})| \leq & |J_\beta(c,\mathcal{T},\gamma_n^*) - J_\beta(c_n,\mathcal{T}_n,\gamma_n^*)| + |J_\beta(c_n,\mathcal{T}_n,\gamma_n^*) \\ & - J_\beta(\mathcal{T},\gamma^*)|.\end{aligned}$$

The first term goes to 0 by Theorem 12.2.3. For the second term we use Theorem 12.2.4. ◇

12.3 Continuity and Robustness in the Fully Observed Case

In this section, we consider the fully observed case where the controller has direct access to the state variables. We present the results for this case separately, since here we cannot utilize the regularity properties of measurement channels which allows for stronger continuity and robustness results. Under measurable selection conditions due to weak or strong (setwise) continuity of transition kernels [165, Section 3.3], for infinite horizon discounted cost problems optimal policies can be selected from those which are stationary and deterministic. Therefore we will restrict the policies to be stationary and deterministic so that $U_t = \gamma(X_t)$ for some measurable function γ. Notice also that fully observed models can be viewed as partially observed with the measurement channel thought to be

$$Q(\cdot|x) = \delta_x(\cdot), \tag{12.6}$$

which is only weakly continuous, thus it does not satisfy Assumption 12.1.2.

12.3.1 Weak Convergence

Absence of Continuity under Weak Convergence.

We start with a negative result.

Theorem 12.3.1 *[183, 187] For* $\mathcal{T}_n \to \mathcal{T}$ *weakly, it is not necessarily true that* $J_\beta^*(c,\mathcal{T}_n) \to J_\beta^*(c,\mathcal{T})$ *even when the prior distributions are the same and* $c(x,u)$ *is continuous and bounded in* $\mathbb{X}\times\mathbb{U}$*.*

Proof. We prove the result with a counterexample, similar to the model used in the proof of Theorem 12.2.1 Letting $\mathbb{X} = [-1,1]$, $\mathbb{U} = \{-1,1\}$ and $c(x,u) = (x-u)^2$, the initial distributions are given by $P \sim \delta_1$, that is, $X_0 = 1$, and the transition kernels are

$$\begin{aligned}\mathcal{T}(\cdot|x,u) = & \delta_{-1}(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] \\ & + (1-\delta_{-1}(x))(1-\delta_1(x))\delta_0(\cdot),\end{aligned}$$

$$\begin{aligned}\mathcal{T}_n(\cdot|x,u) =&\delta_{-1}(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) + \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot)\\ &+ \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + (1-\delta_{-1}(x))(1-\delta_1(x))\delta_0(\cdot).\end{aligned}$$

It can be seen that $\mathcal{T}_n \to \mathcal{T}$ weakly according to Definition 12.1.1(i). Under this setup we can calculate the optimal costs as follows:

$$J^*_\beta(c,\mathcal{T}_n) = \frac{1}{n^2} + \sum_{k=2}^{\infty}\beta^k = \frac{1}{n^2} + \frac{\beta^2}{1-\beta},$$

and $J^*_\beta(c,\mathcal{T}) = 0$. Thus, continuity does not hold. ◇

We now present another counter example emphasizing the importance of continuous convergence in the actions. The following counter example shows that without the continuous convergence and regularity assumptions on the kernel $\mathcal{T}$, continuity fails even when $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ pointwise (for x,u) in total variation (also setwise and weakly) and even when the cost function $c(x,u)$ is continuous and bounded. Notice that this example also holds for both setwise and weak convergence.

Example 12.1. Assume that the kernels are given by

$$\begin{aligned}&\mathcal{T}_n(\cdot|x,u) \sim U([u^n, 1+u^n]),\\ &\mathcal{T}(\cdot|x,u) \sim \begin{cases} U([0,1]) & \text{if } u \neq 1,\\ U([1,2]) & \text{if } u = 1,\end{cases}\end{aligned}$$

where $\mathbb{U} = [0,1]$ and $\mathbb{X} = \mathbb{R}$. We note first that $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ in total variation for every fixed x and u.

The cost function is in the following form:

$$c(x,u) = \begin{cases} 2 & \text{if } x \leq \frac{1}{e},\\ 2 - \frac{x-\frac{1}{e}}{0.1} & \text{if } \frac{1}{e} < x \leq 0.1 + \frac{1}{e},\\ 1 & \text{if } 0.1 + \frac{1}{e} < x \leq 1 + \frac{1}{e} - 0.1,\\ 2 - \frac{1+\frac{1}{e}-x}{0.1} & \text{if } 1 + \frac{1}{e} - 0.1 < x \leq 1 + \frac{1}{e},\\ 2 & \text{if } 1 + \frac{1}{e} < x.\end{cases}$$

Notice that $c(x,u)$ is a continuous function.

With this setup, $\gamma^*(x) = 0$ is an optimal policy for $\mathcal{T}$ since on the $[0,1]$ interval the induced cost is less than the cost induced on the $[1,2]$ interval. The cost under this policy is

$$J^*_\beta(c,\mathcal{T}) = \sum_{t=0}^{\infty}\beta^t\left(2\times\frac{1}{e} + \frac{0.3}{2} + 0.9 - \frac{1}{e}\right) = \frac{1}{1-\beta}\left(1.05 + \frac{1}{e}\right).$$

For $\mathcal{T}_n$, $\gamma^*_n(x) = e^{-\frac{1}{n}}$ is an optimal policy for every n as $e^{-\frac{1}{n}\times n} = \frac{1}{e}$ and thus the state is distributed between $\frac{1}{e} < x \leq 1 + \frac{1}{e}$ in which interval the cost is the least. Hence, we can write

$$\begin{aligned}\lim_{n\to\infty} J_\beta(c,\mathcal{T}_n,\gamma^*_n) &= \sum_{t=0}^{\infty}\beta^t\left(0.3 + 1 - 0.2\right) = \frac{1.1}{1-\beta} \neq \frac{1}{1-\beta}\left(1.05 + \frac{1}{e}\right)\\ &= J^*_\beta(c,\mathcal{T}).\end{aligned}$$

◇

A Sufficient Condition for Continuity under Weak Convergence.

We will now establish that if the kernels and the model components have some further regularity, continuity does hold. The assumptions of the following result are the same as the assumptions for the partially observed case (Theorem 12.2.4) except for the assumption on the measurement channel Q.

Theorem 12.3.2 *[183, 187] Under Assumption 12.1.1, $J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \to J_\beta(c, \mathcal{T}, \gamma^*)$ for any initial state x_0, as $n \to \infty$.*

Proof. We will use the successive approximations for an inductive argument.

Recall the discounted cost optimality operator $T : C_b(Z) \to C_b(Z)$ from (12.3)

$$(T(v))(x) = \inf_{u \in \mathbb{U}} \Big(c(x,u) + \beta E[v(x_1)|x_0 = x, u_0 = u] \Big),$$

which is a contraction from $C_b(\mathbb{X})$ to itself under the supremum norm and has a fixed point, the value function. For the kernel $\mathcal{T}$, we will denote the approximation functions by

$$v^k(x) = T(v^{k-1})(x),$$

and for the kernel $\mathcal{T}_n$ we will use $v_n^k(x)$ to denote the approximation functions, notice that the operator T also depends on n for the model $\mathcal{T}_n$, but we will continue using it as T in what follows.

We wish to show that the approximation functions for $\mathcal{T}_n$ continuously converge to the ones for $\mathcal{T}$. Then, for the first step of the induction we have

$$v^1(x) = c(x, u^*), \quad v_n^1(x_n) = c_n(x_n, u_n^*),$$

and thus we can write,

$$|v^1(x) - v_n^1(x_n)| \leq \sup_{u \in \mathbb{U}} \big| c(x,u) - c_n(x_n, u) \big|$$

since $c_n(x_n, u_n) \to c(x,u)$ for all $(x_n, u_n) \to (x,u)$ and the action space, $\mathbb{U}$, is compact, the first step of the induction holds, i.e. $\lim_{n\to\infty} |v^1(x) - v_n^1(x_n)| = 0$.

For the k^{th} step we have

$$\begin{aligned}
v^k(x) &= T(v^{k-1})(x) = \inf_u \big[c(x,u) + \beta \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x,u) \big], \\
v_n^k(x_n) &= T(v_n^{k-1})(x_n) = \inf_u \big[c_n(x_n,u) + \beta \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n,u) \big].
\end{aligned}$$

Note that the assumptions of the theorem satisfy the measurable selection criteria and hence we can choose minimizing selectors ([165, Section 3.3]). If we denote the selectors by u^* and u_n^*, we can write

$$\begin{aligned}
&|v^k(x) - v_n^k(x_n)| \\
&\leq \max \Big(\Big[|c(x,u^*) - c_n(x_n,u^*)| \\
&\quad + \beta \Big| \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x,u^*) - \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n,u^*) \Big| \Big], \\
&\qquad \Big[|c(x,u_n^*) - c_n(x_n,u_n^*)| \\
&\quad + \beta \Big| \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x,u_n^*) - \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n,u_n^*) \Big| \Big] \Big).
\end{aligned}$$

Hence, we can write

$$\begin{aligned} &|v^k(x) - v_n^k(x_n)| \\ \leq &\sup_{u\in\mathbb{U}} \Big[|c(x,u) - c_n(x_n,u)| \\ &\qquad + \beta \Big| \int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}(dx^1|x,u) - \int_{\mathbb{X}} v_n^{k-1}(x^1)\mathcal{T}_n(dx^1|x_n,u) \Big| \Big], \end{aligned} \tag{12.7}$$

above, the first term goes to 0 as $c_n(x_n,u_n) \to c(x,u)$ for all $(x_n,u_n) \to (x,u)$ and the action space, $\mathbb{U}$, is compact. For the second term we write,

$$\begin{aligned} &\sup_{u\in\mathbb{U}} \Big| \int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}(dx^1|x,u) - \int_{\mathbb{X}} v_n^{k-1}(x^1)\mathcal{T}_n(dx^1|x_n,u) \Big| \\ \leq &\sup_{u\in\mathbb{U}} \Big| \int_{\mathbb{X}} \big(v^{k-1}(x^1) - v_n^{k-1}(x^1)\big)\mathcal{T}_n(dx^1|x_n,u) \Big| \\ &+ \sup_{u\in\mathbb{U}} \Big| \int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}(dx^1|x,u) - \int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}_n(dx^1|x_n,u) \Big| \end{aligned}$$

above, for the first term, by the induction argument for any $x_n^1 \to x^1$, $\big|v^{k-1}(x^1) - v_n^{k-1}(x_n^1)\big| \to 0$ (i.e., we have continuous convergence).

We also have that $\mathcal{T}_n(\cdot|x_n,u) \to \mathcal{T}(\cdot|x,u)$ weakly uniformly over $u \in \mathbb{U}$ as $\mathbb{U}$ is compact. Therefore, using Theorem 12.2.2 the first term goes to 0. For the second term we again use that $\mathcal{T}_n(\cdot|x_n,u)$ converges weakly to $\mathcal{T}(\cdot|x,u)$ uniformly over $u \in \mathbb{U}$. With an almost identical induction argument it can also be shown that $v^{k-1}(x^1)$ is continuous in x^1, thus the second term also goes to 0.

So far, we have showed that for any $k \in \mathbb{N}$, $\lim_{n\to\infty} \big|v_n^k(x_n) - v^k(x)\big| = 0$ for any $x_n \to x$, in particular it is also true that $\lim_{n\to\infty} \big|v_n^k(x) - v^k(x)\big| = 0$ for any x.

As we have stated earlier, it can be shown that the approximation operator, T is a contractive operator under supremum norm with modulus β and it converges to a fixed point which is the value function. Thus, we have

$$\big|J_\beta(c,\mathcal{T},\gamma^*) - v^k(x)\big| \leq \|c\|_\infty \frac{\beta^k}{1-\beta}, \quad \big|J_\beta^*(c_n,\mathcal{T}_n,\gamma_n^*) - v_n^k(x)\big| \leq \|c\|_\infty \frac{\beta^k}{1-\beta}. \tag{12.8}$$

Combining the results,

$$\begin{aligned} |J_\beta(c_n,\mathcal{T}_n,\gamma_n^*) - |J_\beta(c,\mathcal{T},\gamma^*)| \leq &|J_\beta(c_n,\mathcal{T}_n,\gamma_n^*) - v_n^k(x)| + |v_n^k(x) - v^k(x)| \\ &+ |J_\beta(c,\mathcal{T},\gamma^*) - v^k(x)|. \end{aligned}$$

Note that the first and the last term can be made arbitrarily small since (12.8) holds for all $k \in \mathbb{N}$; the second term goes to 0 with an inductive argument for all $k \in \mathbb{N}$. ◇

A Sufficient Condition for Robustness under Weak Convergence.

We now present a result that establishes robustness if the optimal policies for every initial point are identical. That is, for every n, γ_n^* is optimal for every $x_0 \in \mathbb{X}$ (under the model $\mathcal{T}_n$). A sufficient condition for this property is that γ_n^* solves the discounted cost optimality equation (DCOE) for every initial point.

A policy $\gamma^* \in \Gamma$ solves the discounted cost optimality equation and is optimal if it satisfies

$$J_\beta^*(c,\mathcal{T},x) = c(x,\gamma^*(x)) + \beta \int J_\beta^*(c,\mathcal{T},x_1)\mathcal{T}(dx_1|x,\gamma^*(x)).$$

Thus, a policy is optimal for every initial point if it satisfies the DCOE for all initial points $x \in \mathbb{X}$. The following generalizes [187].

Theorem 12.3.3 *Under Assumption 12.1.1, $J_\beta(c_n, \mathcal{T}, \gamma_n^*) \to J_\beta(c, \mathcal{T}, \gamma^*)$ for any initial point x_0 if γ_n^* is optimal for any initial point for the kernel $\mathcal{T}_n$ and for the stage-wise cost function c_n.*

Remark 12.2. For the partially observed case, the proof approach we use makes use of policy exchange (e.g. (12.4)) and for this approach the total variation continuity of channel $Q(\cdot|x)$ is a key step to deal with the uniform convergence over policies. As we stated before, the channel for fully observed models can be considered in the form of (12.6) which is only weakly continuous and not continuous in total variation. Thus, obtaining a result uniformly over all policies may not be possible. However, for the fully observed models we can reach continuity and robustness (Theorem 12.3.2, Theorem 12.3.3) using a value iteration approach. With this approach, instead of exchanging policies and analyzing uniform convergence over all policies, we can exchange control actions (e.g. (12.7)) and analyze uniform convergence over the action space $\mathbb{U}$ by using the discounted optimality operator (12.3). Hence, we are only able to show convergence over optimal policies for the fully observed case, i.e. $J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \to J_\beta(c, \mathcal{T}, \gamma^*)$ or $J_\beta(c, \mathcal{T}, \gamma_n^*) \to J_\beta(c, \mathcal{T}, \gamma^*)$ where γ_n^* and γ^* are optimal policies, whereas, for partially observed models, regularity of the channel allows us to show convergence over any sequence of policies, i.e. $\sup_{\gamma \in \Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \to 0$.

Remark 12.3. As we have discussed in subsection 12.2.1, a partially observed model can be reduced to a fully observed process where the state process (beliefs) becomes probability measure valued. Consider the partially observed models with transition kernels $\mathcal{T}_n$ and $\mathcal{T}$ (with a channel Q) and their corresponding fully observed transition kernels η_n and η: following the discussions and techniques in [128] and [184], one can show that η_n and η satisfy the conditions of Theorem 12.3.3 and Theorem 12.3.2 that is $\eta_n(\cdot|z_n, u_n) \to \eta(\cdot|z, u)$ for any $(z_n, u_n) \to (z, u)$ under the following set of assumptions

- $\mathcal{T}_n(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x, u)$ for any $(x_n, u_n) \to (x, u)$,
- $Q(\cdot|x)$ is continuous on total variation in x.

We remark that these conditions also agree with the conditions presented for continuity and robustness of the partially observed models (Theorem 12.2.4 and Theorem 12.2.8).

12.3.2 Setwise Convergence

Absence of Continuity under Setwise Convergence.

We give a negative result similar to Theorem 12.2.5, via Example 12.1:

Theorem 12.3.4 *Letting $\mathcal{T}_n \to \mathcal{T}$ setwise, then it is not necessarily true that $J_\beta^*(c, \mathcal{T}_n) \to J_\beta^*(c, \mathcal{T})$ even when $c(x, u)$ is continuous and bounded in $\mathbb{X} \times \mathbb{U}$.*

A Sufficient Condition for Continuity under Setwise Convergence.

Theorem 12.3.5 *Under Assumption 12.1.3 $J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \to J_\beta(c, \mathcal{T}, \gamma^*)$, for any initial state x_0, as $n \to \infty$.*

Proof. We use the same value iteration technique that we used to prove Theorem 12.3.2. See [187]. ◇

Absence of Robustness under Setwise Convergence.

Now, we give a result showing that even if the continuity holds under the setwise convergence of the kernels, the robustness may not be satisfied (see [187, Theorem 4.7]).

Theorem 12.3.6 *Supposing $\mathcal{T}_n(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ setwise for every $x \in \mathbb{X}$ and $u \in \mathbb{U}$ and $(x_n,u_n) \to (x,u)$, then it is not true in general that $J_\beta(c,\mathcal{T},\gamma_n^*) \to J_\beta(c,\mathcal{T},\gamma^*)$, even when $\mathbb{X}$ and $\mathbb{U}$ are compact and $c(x,u)$ is continuous and bounded in $\mathbb{X}\times\mathbb{U}$.*

A Sufficient Condition for Robustness under Setwise Convergence.

We now present a similar result to Theorem 12.3.3 that is we show that under the conditions of Theorem 12.3.5, if further for every n, γ_n^* is optimal for every $x_0 \in \mathbb{X}$ (under the model $\mathcal{T}_n$) then robustness holds under setwise convergence.

Theorem 12.3.7 *Supposing Assumption 12.1.3 holds, if further we have that for every n, γ_n^* is optimal for every $x_0 \in \mathbb{X}$ (under the model $\mathcal{T}_n$) then $J_\beta(c,\mathcal{T},\gamma_n^*) \to J_\beta(c,\mathcal{T},\gamma^*)$.*

12.3.3 Total Variation

The continuity result in Theorem 12.2.6 and the robustness result in Theorem 12.2.7 apply to this case since the fully observed model may be viewed as a partially observed model with the measurement channel Q given in (12.6).

12.4 The Average Cost Case

The results above also apply to the average cost setup by adding an ergodicity condition, as we have seen in *Chapter* 7.

In the following we will denote the set of all stationary policies by Γ_s . For the transitions under some stationary policy γ, we will use the following notation: $\mathcal{T}(\cdot|x,\gamma) := \mathcal{T}(\cdot|x,\gamma(x))$.

We also define the t-step transition kernel $\mathcal{T}^t(\cdot|x,\gamma)$ in an iterative fashion as follows:

$$\mathcal{T}^t(\cdot|x,\gamma) := \int \mathcal{T}(\cdot|x_{t-1},\gamma)\mathcal{T}^{t-1}(dx_{t-1}|x,\gamma),$$

where $\mathcal{T}^1(\cdot|x,\gamma) = \mathcal{T}(\cdot|x,\gamma)$.

We will use the following ergodicity condition for some of the results.

Assumption 12.4.1 *For every stationary policy γ, the transition kernels $\mathcal{T}$ and $\mathcal{T}_n$ lead to positive Harris recurrent chains and in particular admit invariant measures π_γ and π_γ^n, and for these invariant measures uniformly for every initial point $x \in X$ we have:*

$$\lim_{t\to\infty} \sup_{\gamma\in\Gamma_s} \|\mathcal{T}^t(\cdot|x,\gamma) - \pi_\gamma(\cdot)\|_{TV} = 0$$
$$\lim_{t\to\infty} \sup_{n} \sup_{\gamma\in\Gamma_s} \|\mathcal{T}_n^t(\cdot|x,\gamma) - \pi_\gamma^n(\cdot)\|_{TV} = 0.$$

We have seen the following earlier in *Chapter* 7, repeated in a concise form for reader's convenience:

For our continuity and robustness results, it will be instrumental to work with stationary policies. This will be without any loss under mild conditions to be presented in this subsection. An approach for average cost problems is to make use of average cost optimality equation (ACOE). To work with ACOE one usually needs contraction properties of the transition kernel. The following result provides further alternative sufficient conditions on existence of optimal policies (which turn out to be stationary) for infinite horizon average cost problems.

Assumption 12.4.2 *The following hold.*

(A) Assumption 7.2.2 holds,

(B) The action space U is compact,

(C) $c(x,u)$ is bounded and continuous in (x,u),

(C') $c(x,u)$ is bounded and continuous in u for every fixed x,

(D) $\mathcal{T}(\cdot|x,u)$ is weakly continuous in (x,u),

(D') $\mathcal{T}(\cdot|x,u)$ is setwise continuous in u for every x.

Proposition 12.4. *Suppose Assumption 12.4.2 A, B, and, either C and D, or C' and D', hold. Then $J_\infty(\mathcal{T},\gamma)$ admits an optimal stationary policy.* ◇

12.4.1 Approximation by finite horizon cost

We denote the t-step finite horizon cost function under a stationary policy γ and a transition model $\mathcal{T}$ by $J_t(\mathcal{T},\gamma)$ and the corresponding optimal cost is denoted by $J_t^*(\mathcal{T})$:

$$J_t(\mathcal{T},\gamma) = \sum_{i=0}^{t-1} E_\gamma^{\mathcal{T}}[c(X_i,U_i)]$$
$$J_t^*(\mathcal{T}) = \inf_{\gamma\in\Gamma} J_t(\mathcal{T},\gamma).$$

The following result shows that the infinite horizon average cost induced by a stationary policy can be approximated by a finite cost under the same stationary policies with proper ergodicity conditions.

Lemma 12.5. *[182] Under Assumption 12.4.1, if the cost function c is bounded then for every initial state we have*

$$\sup_{\gamma\in\Gamma}\left|\frac{J_t(\mathcal{T},\gamma)}{t} - J_\infty(\mathcal{T},\gamma)\right| \to 0,$$
$$\sup_{\gamma\in\Gamma}\sup_n\left|\frac{J_t(\mathcal{T}_n,\gamma)}{t} - J_\infty(\mathcal{T}_n,\gamma)\right| \to 0.$$

Proof. We have that $J_\infty(\mathcal{T},\gamma) = \int c(x,\gamma(x))\pi^\gamma(dx)$. Thus, we can write

$$\begin{aligned}
&\left|\frac{J_t(\mathcal{T},\gamma)}{t} - J_\infty(\mathcal{T},\gamma)\right| \\
&= \left|\frac{1}{t}\sum_{i=0}^{t-1} E_\gamma^{\mathcal{T}}[c(X_i,U_i)] - \int c(x,\gamma(x))\pi^\gamma(dx)\right| \\
&\le \frac{1}{t}\sum_{i=0}^{t-1}\left|\int c(x_i,\gamma(x_i))\mathcal{T}^i(dx_i|x_0,\gamma) - \int c(x,\gamma(x))\pi^\gamma(dx)\right| \\
&\le \frac{1}{t}\sum_{i=0}^{t-1}\|c\|_\infty\|\mathcal{T}^i(\cdot|x_0,\gamma) - \pi^\gamma\|_{TV}.
\end{aligned}$$

We now fix an $\epsilon > 0$ and choose a $t_\epsilon < \infty$ such that $\|\mathcal{T}^i(\cdot|x_0,\gamma) - \pi^\gamma\|_{TV} < \epsilon$ for all $i > t_\epsilon$. We also choose another T_ϵ with $\frac{2t_\epsilon}{t} < \epsilon$ for all $t > T_\epsilon$. With this setup, we have

$$\frac{1}{t}\sum_{i=0}^{t-1}\|\mathcal{T}^i(\cdot|x_0,\gamma) - \pi^\gamma\|_{TV}$$

$$\leq \frac{1}{t}\sum_{i=0}^{t_\epsilon-1} \|\mathcal{T}_\gamma^i(\cdot|x_0) - \pi^\gamma\|_{TV} + \frac{1}{t}\sum_{i=t_\epsilon}^{t} \|\mathcal{T}_\gamma^i(\cdot|x_0) - \pi^\gamma\|_{TV}$$
$$\leq \frac{2t_\epsilon}{t} + \epsilon \leq 2\epsilon, \qquad \forall t > T_\epsilon.$$

We have shown that for any fixed $\epsilon > 0$, we can choose a $T_\epsilon < \infty$, independent of γ, such that

$$\left|\frac{J_t(\mathcal{T},\gamma)}{t} - J_\infty(\mathcal{T},\gamma)\right| < \epsilon, \quad \forall t > T_\epsilon.$$

Hence the result is complete for $\mathcal{T}$.

For $\mathcal{T}_n$ the result follows from the same steps since we can again choose such t_ϵ and T_ϵ due to the uniformity over n and γ in Assumption 12.4.1. ◇

The next result from [160, Corollary 4.11] shows that the optimal infinite horizon cost can be approximated by an optimal finite horizon cost induced by the same transition kernel.

Lemma 12.6. *Suppose the cost function c is bounded and either Assumption 12.4.2 A, B, C, D or Assumption 12.4.2 A, B, C', D' hold (for $\mathcal{T}$ and $\mathcal{T}_n$). Then, we have*

$$\lim_{t\to\infty}\left|J_\infty^*(\mathcal{T}) - \frac{J_t^*(\mathcal{T})}{t}\right| \to 0,$$
$$\lim_{t\to\infty}\sup_n\left|J_\infty^*(\mathcal{T}_n) - \frac{J_t^*(\mathcal{T}_n)}{t}\right| \to 0.$$

12.4.2 Continuity under the convergence of transition kernels

Theorem 12.7. *[182] We have that $|J_\infty^*(\mathcal{T}_n) - J_\infty^*(\mathcal{T})| \to 0$, under*

c1. Assumption 12.4.2 A, B, C and D if $\mathcal{T}_n(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ weakly for any $(x_n,u_n)\to(x,u)$.

c2. Assumption 12.4.2 A, B, C′ and D′ if $\mathcal{T}_n(\cdot|x,u_n) \to \mathcal{T}(\cdot|x,u)$ setwise for any $u_n \to u$ for every fixed x.

Proof. We use the following bound:

$$|J_\infty^*(\mathcal{T}_n) - J_\infty^*(\mathcal{T})|$$
$$\leq \left|J_\infty^*(\mathcal{T}_n) - \frac{J_t^*(\mathcal{T}_n)}{t}\right|$$
$$+ \left|\frac{J_t^*(\mathcal{T}_n)}{t} - \frac{J_t^*(\mathcal{T})}{t}\right| + \left|\frac{J_t^*(\mathcal{T})}{t} - J_\infty^*(\mathcal{T})\right|.$$

The first and the last terms above can be made arbitrarily small by choosing t large enough uniformly over n using Lemma 12.6 under suitable assumptions. For the second term, we can use continuity results for finite time problems for the fixed t as the assumptions cover the requirements studied earlier: see the proofs of Theorem 12.3.2 (for weak convergence) and Theorem 12.3.5 (for setwise convergence). ◇

12.4.3 Robustness to Incorrect Controlled Transition Kernel Models

In this section, we investigate robustness for infinite horizon average cost problems. We first restate the problem: Consider an MDP with transition kernel $\mathcal{T}_n$, and assume that an optimal control policy for this MDP under the average cost criterion is γ_n^*, that is

$$\inf_{\gamma\in\Gamma} J_\infty(\mathcal{T}_n,\gamma) = J_\infty(\mathcal{T}_n,\gamma_n^*).$$

Now, consider another MDP with transition kernel $\mathcal{T}$ whose optimal cost is denoted by $J_\infty^*(\mathcal{T})$. If the controller does not know the true transition kernel $\mathcal{T}$ and calculates an optimal policy assuming the transition kernel is $\mathcal{T}_n$, then the incurred cost by this policy is $J_\infty(\mathcal{T},\gamma_n^*)$. The focus of this section is to find sufficient conditions such that as $\mathcal{T}_n \to \mathcal{T}$,

$$J_\infty(\mathcal{T},\gamma_n^*) \to J_\infty(\mathcal{T},\gamma^*).$$

Suppose that the MDP with kernel $\mathcal{T}_n$ admits two different optimal policies γ_n^1 and γ_n^2. Although, the cost incurred by these policies under the kernel $\mathcal{T}_n$ are the same, under the kernel $\mathcal{T}$ they may have different cost values. That is, even though we have that

$$J_\infty(\mathcal{T}_n,\gamma_n^1) = J_\infty(\mathcal{T}_n,\gamma_n^2) = J_\infty^*(\mathcal{T}_n),$$

we may have $J_\infty(\mathcal{T},\gamma_n^1) \neq J_\infty(\mathcal{T},\gamma_n^2)$. An example is as follows: Consider a system with state space $X=[-1,1]$, control action space $U=\{-1,0,1\}$, the cost function $c(x,u)=(x-u)^2$ and the transition models are given as

$$\begin{aligned}\mathcal{T}_n(\cdot|x,u) &= \frac{1}{2}\delta_1(\cdot)+\frac{1}{2}\delta_{-1}(\cdot)\\ \mathcal{T}(\cdot|x,u) &= \delta_0(\cdot).\end{aligned}$$

Notice that two optimal policies for $\mathcal{T}_n$ are

$$\gamma_n^1(x)=\begin{cases}1 & \text{if } x=1,\\ -1 & \text{if } x=-1,\\ 0 & \text{else.}\end{cases} \quad \gamma_n^2(x)=\begin{cases}1 & \text{if } x\geq 0,\\ -1 & \text{if } x<0.\end{cases}$$

However, if the initial point is $x_0=0$, we have that $J_\infty(\mathcal{T},\gamma_n^1)=0\neq 1=J_\infty(\mathcal{T},\gamma_n^2)$.

In what follows, we show that under total variation convergence of $\mathcal{T}_n\to\mathcal{T}$, this issue does not cause a problem so that we have $J_\infty(\mathcal{T},\gamma_n^*)\to J_\infty(\mathcal{T},\gamma^*)$ for any stationary optimal policy γ_n^*. However, under weak or setwise convergence of the transition models, we establish the same result under some particularly constructed optimal polices γ_n^*, namely we focus on the policies that solve the average cost optimality equation (ACOE) (in analogy with the corresponding results under the discounted cost criterion: Theorem 12.3.3 under weak continuity, Theorem 12.3.7 for setwise continuity).

The following summarize some of the relevant findings in Section 7.2.

Proposition 12.8. *[160] Suppose the cost function c is bounded. Under Assumption 12.4.1, there exists a $\beta<1$ such that the following holds:*

(i) $sp\left(Tv-Tw\right)\leq \beta\, sp(u-w)$, for any $v,w\in B(X)$ where T is the operator defined in (7.21).

(ii) Since T is a contraction under the span norm, it admits a fixed point $v^*\in B(X)$ such that

$$j^*+v^*(x)=\inf_{u\in U}\left(c(x,u)+\int_X v^*(y)\mathcal{T}(dy|x,u)\right),$$

for some constant j^*.

(iii)For any initial point $x_0\in X$, the constant j^* defined in (ii) is the optimal infinite horizon average cost for the kernel $\mathcal{T}$, that is

$$j^*=J_\infty^*(\mathcal{T},x_0)=\inf_{\gamma\in\Gamma}J_\infty(\mathcal{T},\gamma,x_0)$$

for every $x_0\in X$.

(iv) If there exists a policy $\gamma^* \in \Gamma$ satisfying the ACOE, then this stationary policy is an optimal policy for the average infinite horizon cost problem; that is, if γ^* satisfies

$$j^* + v^*(x) = c(x, \gamma^*(x)) + \int_X v^*(y)\mathcal{T}(dy|x, \gamma^*(x)),$$

then $J_\infty(\mathcal{T}, \gamma^*, x_0) = J^*_\infty(\mathcal{T}, x_0)$.

Theorem 12.9. *[182] We have that*

$$J_\infty(\mathcal{T}, \gamma^*_n, x) \to J^*_\infty(\mathcal{T}, x)$$

for any $x \in X$, where γ^*_n is the optimal policy for the transition kernel $\mathcal{T}_n$ that satisfies the ACOE, under Assumption 12.4.2 A, B, C and D if $\mathcal{T}_n(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x, u)$ weakly for any $(x_n, u_n) \to (x, u)$.

Proof. Consider the following two ACOEs for the kernels $\mathcal{T}_n$ and $\mathcal{T}$ with their fixed points v^*_n and v^*:

$$j^*_n + v^*_n(x) = \inf_{u\in U}\left[c(x, u) + \int v^*_n(y)\mathcal{T}_n(dy|x, u)\right] \tag{12.9}$$

$$j^* + v^*(x) = \inf_{u\in U}\left[c(x, u) + \int v^*(y)\mathcal{T}(dy|x, u)\right] \tag{12.10}$$

We now show that, for all $x_n \to x$,

$$v^*_n(x_n) - v^*(x) \to c \tag{12.11}$$

for some constant c with $|c| < \infty$. To show this, we first write

$$\begin{aligned} &v^*_n(x_n) - v^*(x) \\ &= \big(v^*_n(x_n) - v^t_n(x_n)\big) + \big(v^t_n(x_n) - v^t(x)\big) + \big(v^t(x) - v^*(x)\big) \end{aligned}$$

where v^t_n and v^t are the results of operator (7.21) applied to the 0-function, t times for kernels the $\mathcal{T}_n$ and $\mathcal{T}$. Notice that v^t_n and v^t are the value functions for t-step cost problem and by the assumptions ([187, Theorem 4.4]) we have that $|v^t_n(x_n) - v^t(x)| \to 0$ for every fixed t. For the first and the last terms, we use the fact that the operator (7.21) is a contraction under Assumption 7.2.2 for the span semi-norm and hence both terms go to some constants as $t \to \infty$ uniformly for all n, that is $v^*_n(x_n) - v^t_n(x_n) \to c_1$ and $v^t(x) - v^*(x) \to c_2$ for some $|c_1|, |c_2| < \infty$. Thus, we have that (12.11) holds for some $c < \infty$.

Since U is compact, for every $x_n \to x$, $\gamma_n(x_n)$ has a convergent subsequence which converges to say some $u^* \in U$. If we take the limit along this subsequence for (12.9), using the assumptions that $\mathcal{T}_n(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x, u)$ weakly, the fact that $\lim_{n\to\infty}\big(v^*_n(x_n) - v^*(x)\big) = c$, and that $j^*_n \to j^*$ (continuity results from Theorem 12.7) we get

$$\begin{aligned} &\lim_k \left(j^*_{n_k} + v^*_{n_k}(x_{n_k})\right) \\ &= \lim_k c(x, \gamma^*_{n_k}(x_{n_k})) + \int v^*_{n_k}(y)\mathcal{T}_{n_k}(dy|x_{n_k}, \gamma^*_{n_k}(x_{n_k})) \\ &= j^* + v^*(x) + c = c(x, u^*) + \int v^*(y)\mathcal{T}(dy|x, u^*) + c. \end{aligned}$$

Therefore, u^* satisfies the ACOE for the kernel $\mathcal{T}$ and thus, any convergent subsequence of $\gamma^*_n(x_n)$ is an optimal action for x for the kernel $\mathcal{T}$.

Now consider the following operator $\hat{T}_n$, for the kernel $\mathcal{T}$ and the policy γ^*_n which is optimal for $\mathcal{T}_n$

$$\hat{T}_n\hat{v}_n(x) = c(x, \gamma^*_n(x)) + \int \hat{v}_n(y)\mathcal{T}(dy|x, \gamma^*_n(x)). \tag{12.12}$$

One can show that this operator is also a contraction under span semi-norm and admits a fixed point $\hat{v}_n^*$, such that

$$\hat{j}_n + \hat{v}_n^*(x) = c(x, \gamma_n^*(x)) + \int \hat{v}_n^*(y)\mathcal{T}(dy|x, \gamma_n^*(x))$$

where $\hat{j}_n = J_\infty(\mathcal{T}, \gamma_n^*, x)$ for all x. Hence, we need to show that $\hat{j}_n \to j^*$ to complete the proof. To show this, in [182], it has been proven that

$$\lim_{n\to\infty} \hat{v}_n^*(x_n) - v^*(x) = \hat{c}, \tag{12.13}$$

for any $x_n \to x$ for some constant $\hat{c} < \infty$.

Now, assume that $\lim_n \hat{j}_n \neq j^*$ and that there exists a subsequence $\hat{j}_{n_k}$ and an $\epsilon > 0$ such that $|\hat{j}_{n_k} - j^*| > \epsilon$ for every k. We will show that this cannot hold, by establishing the existence of a further subsequence $\hat{j}_{n_{k_l}}$ which converges to j^* in the following.

We first note that $\lim_{n\to\infty} \hat{v}_n^*(x_n) - v^*(x) = \hat{c}$. Hence, Theorem D.3.1 yields that $\int \hat{v}_{n_{k_l}}^*(y)\mathcal{T}(dy|x, \gamma_{n_{k_l}}^*(x)) \to \int v^*(y)\mathcal{T}(dy|x, u^*) + \hat{c}$ where $\hat{c}$ also satisfies $\left(\hat{v}_{n_{k_l}}^*(x) - v^*(x)\right) \to \hat{c}$.

Therefore, taking the limit along this subsequence,

$$\begin{aligned}
&\lim_{l\to\infty} \hat{j}_{n_{k_l}} \\
&= \lim_{l\to\infty} c(x, \gamma_{n_{k_l}}^*(x)) + \int \hat{v}_{n_{k_l}}^*(y)\mathcal{T}(dy|x, \gamma_{n_{k_l}}^*(x)) - \hat{v}_{n_{k_l}}^*(x) \\
&= c(x, u^*) + \int v^*(y)\mathcal{T}(dy|x, u^*) - v^*(x) = j^*.
\end{aligned}$$

This contradicts to $|\hat{j}_{n_k} - j^*| > \epsilon$, hence we conclude that $\hat{j}_n \to j^*$. ◇

We now obtain the same under setwise convergence, without proof.

Theorem 12.10. *[182] We have that $J_\infty(\mathcal{T}, \gamma_n^*, x) \to J_\infty^*(\mathcal{T}, x)$ for any $x \in X$, where γ_n^* is the optimal policy for the transition kernel $\mathcal{T}_n$ that satisfies the ACOE, under Assumption 12.4.2 A, B, C and D if $\mathcal{T}_n(\cdot|x, u_n) \to \mathcal{T}(\cdot|x, u)$ setwise for any $u_n \to u$.*

For total variation, a more direct result follows.

Theorem 12.11. *[182] We have that $|J_\infty(\mathcal{T}, \gamma_n^*) - J_\infty^*(\mathcal{T})| \to 0$ for any stationary optimal policy γ_n^* for $\mathcal{T}_n$, under Assumption 12.4.2 A, B, C′ and D′ if $\mathcal{T}_n(\cdot|x, u_n) \to \mathcal{T}(\cdot|x, u)$ in total variation for any $u_n \to u$ for every fixed x.*

Proof. We write:

$$\begin{aligned}
&|J_\infty(\mathcal{T}, \gamma_n^*) - J_\infty^*(\mathcal{T})| \\
&\quad \leq |J_\infty^*(\mathcal{T}_n) - J_\infty^*(\mathcal{T})| + |J_\infty(\mathcal{T}, \gamma_n^*) - J_\infty^*(\mathcal{T}_n)|
\end{aligned}$$

the first term goes to by Theorem 12.7. For the second term we write

$$\begin{aligned}
&|J_\infty(\mathcal{T}, \gamma_n^*) - J_\infty^*(\mathcal{T}_n)| \\
&\leq \left| J_\infty(\mathcal{T}, \gamma_n^*) - \frac{J_t(\mathcal{T}_n, \gamma_n^*)}{t} \right| \\
&+ \left| \frac{J_t(\mathcal{T}_n, \gamma_n^*)}{t} - \frac{J_t(\mathcal{T}, \gamma_n^*)}{t} \right| + \left| \frac{J_t(\mathcal{T}, \gamma_n^*)}{t} - J_\infty(\mathcal{T}, \gamma_n^*) \right|
\end{aligned}$$

The first and the last terms above again can be made arbitrarily small by choosing t large enough uniformly over n using Lemma 12.5. For the second term we use [187, Section A.2] where it is shown that under the stated assumptions $\sup_{\gamma\in\Gamma}|J_t(\mathcal{T}_n,\gamma)-J_t(\mathcal{T},\gamma)|\to 0$. Hence the proof is complete. ⋄

12.5 Applications to Data-Driven Learning and Finite Model Approximations

12.5.1 Application of Robustness Results to Data-Driven Learning

In practice, one may estimate the kernel of a controlled Markov chain using empirical data.

Let us briefly review the basic case where an i.i.d. sequence of random variables is repeatedly observed, but its probability measure is not known apriori. Let $\{(X_i),\, i\in\mathbb{N}\}$ be an $\mathbb{X}$-valued i.i.d. random variable sequence generated according to some distribution μ. Defining for every (fixed) Borel $B\subset\mathbb{X}$, and $n\in\mathbb{N}$, the empirical occupation measures $\mu_n(B)=\frac{1}{n}\sum_{i=1}^n 1_{\{X_i\in B\}}$, one has $\mu_n(B)\to\mu(B)$ almost surely by the strong law of large numbers. It then follows that $\mu_n\to\mu$ weakly with probability one ([111], Theorem 11.4.1). However, μ_n does not converge to μ in total variation or setwise, in general. On the other hand, if we know that μ admits a density, we can find estimators to estimate μ under total variation [104, Chapter 3]. In the previous sections, we established robustness results under the convergence of transition kernels in the topology of weak convergence and total variation. We build on these observations.

Corollary 12.12 (to Theorem 12.2.6 and Theorem 12.2.7). *Suppose we are given the following dynamics for finite state space, $\mathbb{X}$, and finite action space, $\mathbb{U}$,*

$$x_{t+1}=f(x_t,u_t,w_t),\qquad y_t=g(x_t,v_t)$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d.noise processes and the noise models are unknown. Suppose that there is an initial training period so that under some policy, every x,u pair is visited infinitely often if training were to continue indefinitely, but that the training ends at some finite time. Let us assume that, through this training, we empirically learn the transition dynamics such that for every (fixed) Borel $B\subset\mathbb{X}$, for every $x\in\mathbb{X}$, $u\in\mathbb{U}$ and $n\in\mathbb{N}$, the empirical occupation measures are

$$\mathcal{T}_n(B|x_0=x,u_0=u)=\frac{\sum_{i=1}^n 1_{\{X_i\in B,X_{i-1}=x,U_{i-1}=u\}}}{\sum_{i=1}^n 1_{\{X_{i-1}=x,U_{i-1}=u\}}}.$$

Then we have that $J^*_\beta(\mathcal{T}_n)\to J^*_\beta(\mathcal{T})$ and $J_\beta(\mathcal{T},\gamma^*_n)\to J^*_\beta(\mathcal{T})$, where γ^*_n is the optimal policy designed for $\mathcal{T}_n$. Since the channel model g has no restrictions, this result also applies to the fully observed model setup by taking $g(x_t,v_t)=x_t$.

Proof. We have that $\mathcal{T}_n(\cdot|x,u)\to\mathcal{T}(\cdot|x,u)$ weakly for every $x\in\mathbb{X}$, $u\in\mathbb{U}$ almost surely by law of large numbers. Since the spaces are finite, we also have $\mathcal{T}_n(\cdot|x,u)\to\mathcal{T}(\cdot|x,u)$ under total variation. By Theorem 12.2.6 and Theorem 12.2.7, the results follow. ⋄

The following holds for more general spaces.

Corollary 12.13 (to Theorems 12.2.8, 12.2.4, 12.3.2 and 12.3.3). *Suppose we are given the following dynamics with state space $\mathbb{X}$ and action space $\mathbb{U}$,*

$$x_{t+1}=f(x_t,u_t,w_t),\qquad y_t=g(x_t,v_t),$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d.noise processes and the noise models are unknown. Suppose that $f(x,u,\cdot):\mathbb{W}\to\mathbb{X}$ is invertible for all fixed (x,u) and $f(x,u,w)$ is continuous and bounded on $\mathbb{X}\times\mathbb{U}\times\mathbb{W}$. We construct the empirical measures for the noise process w_t such that for every (fixed) Borel $B\subset\mathbb{W}$, and for every $n\in\mathbb{N}$, the empirical occupation measures are

$$\mu_n(B)=\frac{1}{n}\sum_{i=1}^n 1_{\{f^{-1}_{x_{i-1},u_{i-1}}(x_i)\in B\}} \tag{12.14}$$

where $f^{-1}_{x_{i-1},u_{i-1}}(x_i)$ denotes the inverse of $f(x_{i-1},u_{i-1},w):\mathbb{W}\to\mathbb{X}$ for given (x_{i-1},u_{i-1}). Using the noise measurements, we construct the empirical transition kernel estimates for any (x_0,u_0) and Borel B as

$$\mathcal{T}_n(B|x_0,u_0)=\mu_n(f^{-1}_{x_0,u_0}(B)).$$

(i) If the measurement channel (represented by the function g) is continuous in total variation then $J^*_\beta(\mathcal{T}_n)\to J^*_\beta(\mathcal{T})$ and $J_\beta(\mathcal{T},\gamma^*_n)\to J^*_\beta(\mathcal{T})$, where γ^*_n is the optimal policy designed for $\mathcal{T}_n$ for all initial points.

(ii) If the measurement channel is in the form $g(x_t,v_t)=x_t$ (i.e. fully observed) then $J^*_\beta(\mathcal{T}_n)\to J^*_\beta(\mathcal{T})$ and if further for every n, γ^*_n is optimal for every $x_0\in\mathbb{X}$ (under the model $\mathcal{T}_n$) then $J_\beta(\mathcal{T},\gamma^*_n)\to J^*_\beta(\mathcal{T})$.

Proof. We have $\mu_n\to\mu$ weakly with probability one where μ is the model. We claim that the transition kernels are such that $\mathcal{T}_n(\cdot|x_n,u_n)\to\mathcal{T}(\cdot|x,u)$ weakly for any $(x_n,u_n)\to(x,u)$. To see that observe the following for $h\in C_b(\mathbb{X})$

$$\begin{aligned}&\int h(x_1)\mathcal{T}_n(dx_1|x_n,u_n)-\int h(x_1)\mathcal{T}(dx_1|x,u)\\&=\int h(f(x_n,u_n,w))\mu_n(dw)-\int h(f(x,u,w))\mu(dw)\to 0,\end{aligned}$$

where μ_n is the empirical measure for w_t and μ is the true measure again. For the last step, we used that $\mu_n\to\mu$ weakly and $h(f(x_n,u_n,w))$ continuously converge to $h(f(x,u,w))$ i.e. $h(f(x_n,u_n,w_n))\to h(f(x,u,w)$ for some $w_n\to w$ since f and g are continuous functions. Similarly, it can be also shown that $\mathcal{T}_n(\cdot|x,u)$ and $\mathcal{T}(\cdot|x,u)$ are weakly continuous on (x,u). Thus, for the case where the channel is continuous in total variation by Theorem 12.2.8 and Theorem 12.2.4 if $c(x,u)$ is bounded and $\mathbb{U}$ is compact the result follows.

For the fully observed case, $J^*_\beta(\mathcal{T}_n)\to J^*_\beta(\mathcal{T})$ by Theorem 12.3.2 and $J_\beta(\mathcal{T},\gamma^*_n)\to J^*_\beta(\mathcal{T})$ by Theorem 12.3.3. ⋄

Remark 12.14. We note here that the moment estimation method can also lead to consistency. Suppose that the distribution of W is determined by its moments, such that estimate models W_n have moments of all orders and $\lim_n = E[W_n^r] = E[W^r]$ for all $r\in\mathbb{Z}_+$. Then, we have that [43, Thm 30.2] $W_n\to W$ weakly and thus $\mathcal{T}_n(\cdot|x_n,u_n)\to\mathcal{T}(\cdot|x,u)$ weakly for any $(x_n,u_n)\to(x,u)$ under the assumptions of above corollary. Hence, we reach continuity and robustness using the same arguments as in the previous result (Corollary 12.13).

Now, we give a similar result with the assumption that the noise process of the dynamics admits a continuous probability density function.

Corollary 12.15 (to Theorem 12.2.6 and Theorem 12.2.7). *Suppose we are given the following dynamics for real vector state space $\mathbb{X}$ and action space $\mathbb{U}$*

$$x_{t+1}=f(x_t,u_t,w_t),\qquad y_t=g(x_t,v_t),$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d.noise processes and the noise models are unknown but it is known that the noise w_t admits a continuous probability density function. Suppose that $f(x,u,\cdot):\mathbb{W}\to\mathbb{X}$ is invertible for all (x,u). We collect i.i.d. samples of $\{w_t\}$ as in (12.14) and use them to construct an estimator, $\tilde{\mu}_n$, as described in [104] which consistently estimates μ in total variation. Using these empirical estimates, we construct the empirical transition kernel estimates for any (x_0,u_0) and Borel B as

$$\mathcal{T}_n(B|x_0,u_0)=\tilde{\mu}_n(f^{-1}_{x_0,u_0}(B)).$$

Then independent of the channel, $J^*_\beta(\mathcal{T}_n)\to J^*_\beta(\mathcal{T})$ and $J_\beta(\mathcal{T},\gamma^*_n)\to J^*_\beta(\mathcal{T})$, where γ^*_n is the optimal policy designed for $\mathcal{T}_n$. Since the channel model g has no restrictions, this result also applies to the fully observed model setup by taking $g(x_t,v_t)=x_t$.

Proof. By [104] we can estimate μ in total variation so that almost surely $\lim_{n\to\infty}\|\tilde{\mu}_n-\mu\|_{TV}=0$. We claim that the convergence of $\tilde{\mu}_n$ to μ under total variation metric implies the convergence of $\mathcal{T}_n$ to $\mathcal{T}$ in total variation uniformly over all $x\in\mathbb{X}$ and $u\in\mathbb{U}$ i.e. $\lim_{n\to\infty}\sup_{x,u}\|\mathcal{T}_n(\cdot|x,u)-\mathcal{T}(\cdot|x,u)\|_{TV}=0$. Observe the following:

$$\begin{aligned}
&\sup_{x,u} \|\mathcal{T}_n(\cdot|x,u) - \mathcal{T}(\cdot|x,u)\|_{TV} \\
=&\sup_{x,u} \sup_{||h||_\infty \leq 1} \Big| \int h(x_1)\mathcal{T}_n(dx_1|x,u) - \int h(x_1)\mathcal{T}(dx_1|x,u) \Big| \\
=&\sup_{x,u} \sup_{||h||_\infty \leq 1} \Big| \int h(f(x,u,w))\tilde{\mu}_n(dw) - \int h(f(x,u,w))\mu(dw) \Big| \\
\leq&\|\tilde{\mu}_n - \mu\|_{TV} \to 0.
\end{aligned}$$

Thus, by Theorem 12.2.6 and Theorem 12.2.7, the result follows. ◇

The following example presents some system and channel models which satisfy the requirements of the above corollaries.

Example 12.16. Let $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{U}$ be real vector spaces with

$$x_{t+1} = f(x_t, u_t) + w_t, \qquad y_t = h(x_t, v_t)$$

for unknown i.i.d. noise processes $\{w_t\}$ and $\{v_t\}$.

1. Suppose the channel is in the following form; $y_t = h(x_t, v_t) = x_t + v_t$ where v_t admits a density (e.g. Gaussian density). It can be shown by an application of Scheffé's theorem that the channels in this form are continuous in total variation. If further $f(x_t, u_t)$ is continuous and bounded then the requirements of Corollary 12.13 hold for partially observed models.
2. If the channel is in the following form; $x_t = h(x_t, v_t)$ then the system is fully observed. If further $f(x_t, u_t)$ is continuous and bounded then the requirements of Corollary 12.13 holds for fully observed models.
3. Suppose the function $f(x_t, u_t)$ is known, if the noise process w_t admits a continuous density, then one can estimate the noise model in total variation in a consistent way (see [104]). Hence, the conditions of Corollary 12.15 holds independent of the channel model.

◇

12.5.2 Application to Approximations of MDPs and POMDPs with Weakly Continuous Kernels

We now discuss **Problem P4**, that is whether approximation of an MDP model with a standard Borel space with a finite MDPs can be viewed an instance of robustness problem to incorrect models and whether our results can be applied.

In Section 8.2, we presented conditions under which finite state/action models are asymptotically optimal. Here, we view those approximation results as an instance of robustness. We will focus on the weakly continuous model setup.

By Section 8.2.1 we know that finite quantization policies are nearly optimal under mild weak continuity conditions (see Assumption 8.2.1). Thus, to make the presentation shorter, we will either assume that the action set is finite, or it has been approximated by a finite action space through the construction above. Assuming finite action sets will help us avoid measurability issues ([275, p. 6-7]) as well as issues with existence of optimal policies.

One can write the following fixed point equation for the finite MDP

$$J^n_\beta(x) = \min_{a \in \mathbb{U}} \left\{ c_n(x,a) + \beta \sum_{x_1 \in \mathbb{X}_n} J^n_\beta(x_1)\mathcal{T}_n(x_1|x,a) \right\}$$

where $\mathcal{T}_n$ is the transition model for the finite MDP and c_n is the cost function defined on the finite model. Since the acton space is finite, we can find an optimal policy, say f^*_n for this fixed point equation. One can also simply extend J^n_β and f^*_n, which are defined on $\mathbb{X}_n$ to the entire state space $\mathbb{X}$ by taking them constant over the quantization bins $\mathbb{S}_{n,i}$. If we call the extended versions $\hat{J}^n_\beta$ and $\hat{f}_n$, the following result holds, which is a re-statement of Theorem **??**:

Theorem 12.5.1 *[275, Theorem 2.2 and 4.1] Suppose Assumption 8.2.1 holds. Then, for any $\beta \in (0,1)$ the discounted cost of the deterministic stationary policy $\hat{f}_n$, obtained by extending the discounted optimal policy f_n^* of f-MDP$_m$ to $\mathbb{X}$ (i.e., $\hat{f}_n = f_n^* \circ Q_n$), converges to the discounted value function J^* of the compact-state MDP:*

$$\lim_{n\to\infty} \|\hat{J}_\beta^n(\cdot) - J_\beta^*(\cdot)\| = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \|J_\beta(\hat{f}_n, \cdot) - J_\beta^*\| = 0. \tag{12.15}$$

Theorem 12.5.1 shows that under Assumption 8.2.1, an optimal solution can be approximated via the solutions of finite models. We now show that the above approximation scheme can be viewed in relation to our robustness results.

Proof sketch of Theorem 12.5.1 via results from Section 12.3. With the introduced setup, one can see that the extended value function and optimal policy for the finite model satisfy the following:

$$\hat{J}_\beta^n(x) = \min_{a\in\mathbb{U}} \left\{ \hat{c}_n(x,u) + \beta \int \hat{J}_\beta^n(x_1) \hat{\mathcal{T}}_n(dx_1|x,u) \right\}$$

where $\hat{c}_n$ is the extended version of c_n to the state space $\mathbb{X}$ by making it constant over the quantization bins $\{\mathbb{S}_{n,i}\}_i$ and $\hat{\mathcal{T}}_n$ is such that for any function f

$$\int f(x_1) \hat{\mathcal{T}}_n(dx_1|x,u) := \int_{x_1\in\mathbb{X}} \int_{z\in\mathbb{S}_{n,i}} f(x_1) \mathcal{T}(dx_1|z,u) \psi_{n,i}(dz)$$

where $\mathbb{S}_{n,i}$ is the quantization bin that x belongs to.

With this setup, one can see that for any $x_n \to x$ we have $\hat{c}_n(x_n,u) \to c(x,u)$ and for any continuous and bounded f

$$\begin{aligned} &\int f(x_1) \hat{\mathcal{T}}_n(dx_1|x_n,u) := \int_{x_1\in\mathbb{X}} \int_{z\in\mathbb{S}_{n,i}} f(x_1) \mathcal{T}(dx_1|z,u) \psi_{n,i}(dz) \\ &\to \int f(x_1) \mathcal{T}(dx_1|x,u). \end{aligned}$$

Hence, Assumption 12.1.1 holds under Assumption 8.2.1, and we can conclude the proof using Theorem 12.3.3 and Theorem 12.3.2. ◇

12.6 Bibliographic Notes

In this chapter, we studied regularity properties of optimal stochastic control on the space of transition kernels, and applications to robustness of optimal control policies designed for an incorrect model applied to an actual system. We also presented applications to data-driven learning and related the robustness problem to finite MDP approximation techniques. For the problems presented in this chapter, our focus was on infinite horizon discounted cost setup. However, we note that the results can be extended to the infinite horizon average cost setup under various forms of ergodicity properties on the state process.

Robustness is a desired property for the optimal control of stochastic or deterministic systems when a given model does not reflect the actual system perfectly, as is usually the case in practice. This is a classical problem, and there is a very large literature on robust stochastic control and its application to learning-theoretic methods; see e.g. [24, 31, 52, 118, 148, 170, 337], [6, 18, 115, 119, 179, 186, 187, 211, 240, 250, 254, 298, 311]. Studies on robustness via minimax methods include [177, 245]. A comprehensive literature review is presented in [183, 187]. For empirical learning methods and their stability properties, see [113, 149]

This chapter primarily builds on [182, 183, 187, 192].

12.7 Exercises

Exercise 12.7.1

A

Basics of Function Spaces

A.1 Normed Linear (Vector) Spaces and Metric Spaces

Definition A.1.1 *A linear (vector) space $\mathbb{X}$ is a space which is closed under addition and scalar multiplication: In particular, we define an addition operation, $+$ and a scalar multiplication operation $\cdot$ such that*

$$+ : \mathbb{X} \times \mathbb{X} \to \mathbb{X}$$

$$\cdot : \mathbb{C} \times \mathbb{X} \to \mathbb{X}$$

with the following properties (we note that we may take the scalars to be either real or complex numbers). The following are satisfied for $x, y \in \mathbb{X}$ and α, β scalars:

(i) $x + y = y + x$

(ii) $(x + y) + z = x + (y + z)$.

(iii) $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y$.

(iv) $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$.

(v) There is a null vector $\underline{0}$ such that $x + \underline{0} = x$.

(vi) $\alpha \cdot (\beta \cdot x) = (\alpha\beta) \cdot x$

(vii) For every $x \in \mathbb{X}$, $1 \cdot x = x$

(viii) For every $x \in \mathbb{X}$, there exists an element, called the (additive) inverse of x and denoted with $-x$ with the property $x + (-x) = \underline{0}$.

Example A.1. (i) The space $\mathbb{R}^n$ is a linear space. The null vector is $\underline{0} = (0, 0, \cdots, 0) \in \mathbb{R}^n$.
(ii) Consider the interval $[a, b]$. The collection of real-valued continuous functions on $[a, b]$ is a linear space. The null element $\underline{0}$ is the function which is identically 0. This space is called the space of real-valued continuous functions on $[a, b]$
(iii) The set of all infinite sequences of real numbers having only a finite number of terms not equal to zero is a vector space. If one adds two such sequences, the sum also belongs to this space. This space is called the space of finitely many non-zero sequences.
(iv) The collection of all polynomial functions defined on an interval $[a, b]$ with complex coefficients forms a complex linear space. Note that the sum of polynomials is another polynomial.

Definition A.1.2 *A non-empty subset M of a (real) linear vector space $\mathbb{X}$ is called a subspace of $\mathbb{X}$ if*

$$\alpha x + \beta y \in M, \quad \forall x, y \in M \quad \textit{and} \quad \alpha, \beta \in \mathbb{R}.$$

In particular, the null element $\underline{0}$ is an element of every subspace. For M, N two subspaces of a vector space $\mathbb{X}$, $M \cap N$ is also a subspace of $\mathbb{X}$.

Definition A.1.3 *A normed linear space X is a linear vector space on which a map from X to* $\mathbb{R}$*, that is a member of* $\Gamma(X;\mathbb{R})$*) called norm is defined such that:*

- $||x|| \geq 0 \quad \forall x \in X, \quad ||x|| = 0$ *if and only if* x *is the null element (under addition and multiplication) of X.*
- $||x + y|| \leq ||x|| + ||y||$
- $||\alpha x|| = |\alpha|||x||, \quad \forall \alpha \in \mathbb{R}, \quad \forall x \in X$

Definition A.1.4 *In a normed linear space X, an infinite sequence of elements* $\{x_n\}$ *converges to an element x if the sequence* $\{||x_n - x||\}$ *converges to zero.*

Example A.2. a) The normed linear space $C([a,b])$ consists of continuous functions on $[a,b]$ together with the norm $||x|| = \max_{\{a \leq t \leq b\}} |x(t)|$.

b) $l_p(\mathbb{Z}_+;\mathbb{R}) := \{x \in \Gamma(\mathbb{Z}_+;\mathbb{R}) : ||x||_p = \left(\sum_{i\in\mathbb{Z}_+} |x(i)|^p\right)^{\frac{1}{p}} < \infty\}$ is a normed linear space for all $1 \leq p < \infty$. c) Recall that if S is a set of real numbers bounded above, then there is a smallest real number y such that $x \leq y$ for all $x \in S$. The number y is called the *least upper bound* or *supremum* of S. If S is not bounded from above, then the supremum is ∞. In view of this, for $p = \infty$, we define

$$l_\infty(\mathbb{Z}_+;\mathbb{R}) := \{x \in \Gamma(\mathbb{Z}_+;\mathbb{R}) : ||x||_\infty = \sup_{i\in\mathbb{Z}_+} |x(i)| < \infty\}$$

d) $L_p([a,b];\mathbb{R}) = \{\{x \in \Gamma([a,b];\mathbb{R}) : ||x||_p = \left(\int_a^b |x(i)|^p\right)^{\frac{1}{p}} < \infty\}$ is a normed linear space. For $p = \infty$, we typically write: $L_\infty([a,b];\mathbb{R}) := \{x \in \Gamma([a,b];\mathbb{R}) : ||x||_\infty = \sup_{t\in[a,b]} |x(t)| < \infty\}$. However, for $1 \leq p < \infty$, to satisfy the condition that $||x||_p = 0$ implies that $x(t) = 0$, we need to assume that functions which are equal to zero almost everywhere are equivalent; for $p = \infty$ the definition is often revised with essential supremum instead of supremum so that

$$||x||_\infty = \inf_{y(t)=x(t)\text{a.e.}} \sup_{t\in[a,b]} |y(t)|$$

To show that l_p defined above is a normed linear space, we need to show that $||x + y||_p \leq ||x||_p + ||y||_p$.

Theorem A.1.1 (Minkowski's Inequality) *For* $1 \leq p \leq \infty$

$$||x + y||_p \leq ||x||_p + ||y||_p$$

The proof of this result uses a very important inequality, known as Hölder's inequality.

Theorem A.1.2 (Hölder's Inequality)

$$\sum x(k)y(k) \leq ||x||_p||y||_q,$$

with $1/p + 1/q = 1$ *and* $1 \leq p, q \leq \infty$.

Definition A.1.5 *A metric defined on a set X, is a function* $d : X \times X \to \mathbb{R}$ *such that:*

- $d(x,y) \geq 0, \quad \forall x, y \in X$ *and* $d(x,y) = 0$ *if and only if* $x = y$.
- $d(x,y) = d(y,x), \quad \forall x, y \in X.$
- $d(x,y) \leq d(x,z) + d(z,y), \quad \forall x, y, z \in X.$

Definition A.1.6 *A metric space (X, d) is a set equipped with a metric d.*

A normed linear space is also a metric space, with metric

$$d(x, y) = ||x - y||.$$

An important class of normed spaces that are widely used in optimization and engineering problems are Banach spaces:

A.1.1 Banach Spaces

Definition A.1.7 *A sequence $\{x_n\}$ in a normed space X is Cauchy if for every ϵ, there exists an N such that $||x_n - x_m|| \leq \epsilon$, for all $n, m \geq N$.*

The important observation on Cauchy sequences is that, every converging sequence is Cauchy, however, not all Cauchy sequences are convergent: This is because the limit might not live in the original space where the sequence elements take values in. This brings the issue of completeness:

Definition A.1.8 *A normed linear space X is complete, if every Cauchy sequence in X has a limit in X. A complete normed linear space is called Banach.*

Banach spaces are important for many reasons including the following one: In optimization problems, sometimes we would like to see if a sequence converges, for example if a solution to a minimization problem exists, without knowing what the limit of the sequence could be. Banach spaces allow us to use Cauchy sequence arguments to claim the existence of optimal solutions. If time allows, we will discuss how this is used by using *contraction* and *fixed point* arguments for transformations.

In applications, we will also discuss completeness of a subset. A subset of a Banach space X is complete if and only if it is closed. If it is not closed, one can provide a counterexample sequence which does not converge. If the set is closed, every Cauchy sequence in this set has a limit in X and this limit should be a member of this set, hence the set is complete.

Exercise A.1.1 *The space of bounded functions $\{x : [0, 1] \to \mathbb{R}, sup_{t\in[0,1]}|x(t)| < \infty\}$ is a Banach space.*

The above space is also denoted by $L_\infty([0, 1]; \mathbb{R})$ or $L_\infty([0, 1])$.

Theorem A.1.3 $l_p(\mathbb{Z}_+; \mathbb{R}) := \{x \in f(\mathbb{Z}_+; \mathbb{R}) : ||x||_p = \left(\sum_{i\in\mathbb{N}_+} |x(i)|^p \right)^{\frac{1}{p}} < \infty\}$ *is a Banach space for all* $1 \leq p \leq \infty$.

Sketch of Proof: The proof is completed in three steps.

(i) Let $\{x_n\}$ be Cauchy. This implies that for every $\epsilon > 0$, $\exists N$ such that for all $n, m \geq N$ $||x_n - x_m|| \leq \epsilon$. This also implies that for all $n > N$, $||x_n|| \leq ||x_N|| + \epsilon$. Now let us denote x_n by the vector $\{x_1^n, x_2^n, x_3^n \dots, \}$. It follows that for every k the sequence $\{x_k^n\}$ is also Cauchy. Since $x_k^n \in \mathbb{R}$, and $\mathbb{R}$ is complete, $x_k^n \to x_k$ for some x_k. Thus, the sequence x_n pointwise converges to some vector x_*.

(ii) Is $x \in l_p(\mathbb{Z}_+; \mathbb{R})$? Define $x_{n,K} = \{x_1^n, x_2^n, \dots, x_{K-1}^n, x_K^n, 0, 0, \dots\}$, that is vector which truncates after the Kth coordinate. Now, it follows that

$$||x_{n,K}|| \leq ||x_N|| + \epsilon,$$

for every $n \geq N$ and K and

$$\lim_{n\to\infty} ||x_{n,K}||^p = \lim_{n\to\infty} \sum_{i=1}^{K} |x_i^n|^p = \sum_{i=1}^{K} |x_i|^p,$$

since there are only finitely many elements in the summation. The question now is whether $||x_\infty|| \in p(\mathbb{Z}_+;\mathbb{R})$. Now,

$$||x_{n,K}|| \leq ||x_N|| + \epsilon,$$

and thus

$$\lim_{n\to\infty} ||x_{n,K}|| = ||x_K|| \leq ||x_N|| + \epsilon,$$

Let us take another limit, by the monotone convergence theorem (Recall that this theorem says that a monotonically increasing sequence which is bounded has a limit).

$$\lim_{K\to\infty} ||x_{*,K}||^p = \lim_{K\to\infty} \sum_{i=1}^{K} |x_i|^p = ||x_\infty||_p^p \leq ||x_N|| + \epsilon.$$

(iii) The final question is: Does $||x_n - x_*|| \to 0$? Since the sequence is Cauchy, it follows that for $n, m \geq N$

$$||x_n - x_m|| \leq \epsilon$$

Thus,

$$||x_{n,K} - x_{m,K}|| \leq \epsilon$$

and since K is finite

$$\lim_{m\to\infty} ||x_{n,K} - x_{m,K}|| = ||x_{n,K} - x_{*,K}|| \leq \epsilon$$

Now, we take another limit

$$\lim_{K\to\infty} ||x_{n,K} - x_{*,K}|| \leq \epsilon$$

By the monotone convergence theorem again,

$$\lim_{K\to\infty} ||x_{n,K} - x_{*,K}|| = ||x_n - x|| \leq \epsilon$$

Hence, $||x_n - x|| \to 0$. ◇

The above spaces are also denoted $l_p(\mathbb{Z}_+)$, when the range space is clear from context.

The following is a useful result.

Theorem A.1.4 (Hölder's Inequality)

$$\sum x(t)y(t) \leq ||x||_p ||y||_q,$$

with $1/p + 1/q = 1$ *and* $1 \leq p, q \leq \infty$.

Remark: A brief remark for notations: When the range space is $\mathbb{R}$, the notation $l_p(\Omega)$ denotes $l_p(\Omega;\mathbb{R})$ for a discrete-time index set Ω and likewise for a continuous-time index set Ω, $L_p(\Omega)$ denotes $L_p(\Omega;\mathbb{R})$. ◇

A.1.2 Hilbert Spaces

We first define pre-Hilbert spaces.

Definition A.1.9 *A pre-Hilbert space X is a linear vector space where an inner product is defined on $X \times X$. Corresponding to each pair $x, y \in X$ the inner product $\langle x, y\rangle$ is a scalar (that is real-valued or complex-valued). The inner product satisfies the following axioms:*

1. $\langle x, y\rangle = \langle y, x\rangle^*$ *(the superscript denotes the complex conjugate) (we will also use* $\overline{\langle y, x\rangle}$ *to denote the complex conjugate)*
2. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$

3. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$

4. $\langle x, x \rangle \geq 0$*, equals* 0 *iff* x *is the null element.*

The following is a crucial result in such a space, known as the Cauchy-Schwarz inequality, the proof of which was presented in class:

Theorem A.1.5 *For* $x, y \in X$*,*

$$\langle x, y \rangle \leq \sqrt{\langle x, x \rangle} \sqrt{\langle y, y \rangle},$$

where equality occurs if and only if $x = \alpha y$ *for some scalar* α*.*

Exercise A.1.2 *In a pre-Hilbert space* $\langle x, x \rangle$ *defines a norm:* $||x|| = \sqrt{\langle x, x \rangle}$

The proof for the result requires one to show that $\sqrt{\langle x, x \rangle}$ satisfies the triangle inequality, that is

$$||x + y|| \leq ||x|| + ||y||,$$

which can be proven by an application of the Cauchy-Schwarz inequality.

Not all spaces admit an inner product. In particular, however, $l_2(\mathbb{N}_+; \mathbb{R})$ admits an inner product with $\langle x, y \rangle = \sum_{t \in \mathbb{N}_+} x(n) y(n)$ for $x, y \in l_2(\mathbb{N}_+; \mathbb{R})$. Furthermore, $||x|| = \sqrt{\langle x, x \rangle}$ defines a norm in $l_2(\mathbb{N}_+; \mathbb{R})$.

The inner product, in the special case of $\mathbb{R}^N$, is the usual inner vector product; hence $\mathbb{R}^N$ is a pre-Hilbert space with the usual inner-product.

Definition A.1.10 *A complete pre-Hilbert space, is called a Hilbert space.*

Hence, a Hilbert space is a Banach space, endowed with an inner product, which induces its norm.

Proposition A.1.1 *The inner product is continuous: if* $x_n \to x$*, and* $y_n \to y$*, then* $\langle x_n, y_n \rangle \to \langle x, y \rangle$ *for* x_n, y_n *in a Hilbert space.*

Proposition A.1.2 *In a Hilbert space* X*, two vectors* $x, y \in X$ *are orthogonal if* $\langle x, y \rangle = 0$*. A vector x is orthogonal to a set* $S \subset X$ *if* $\langle x, y \rangle = 0 \quad \forall y \in S$*.*

Theorem A.1.6 (Projection Theorem:) *Let H be a Hilbert space and B a closed subspace of H. For any vector* $x \in H$*, there is a unique vector* $m \in B$ *such that*

$$||x - m|| \leq ||x - y||, \forall y \in B.$$

A necessary and sufficient condition for $m \in B$ *to be the minimizing element in* B *is that,* $x - m$ *is orthogonal to* B*.*

A.1.3 Separability

Definition A.1.11 *Given a normed linear space* X*, a subset* $D \subset X$ *is dense in* X*, if for every* $x \in X$*, and each* $\epsilon > 0$*, there exists a member* $d \in D$ *such that* $||x - d|| \leq \epsilon$*.*

Definition A.1.12 *A set is countable if every element of the set can be associated with an integer via an ordered mapping.*

Examples of countables spaces are finite sets and the set $\mathbb{Q}$ of rational numbers. An example of uncountable sets is the set $\mathbb{R}$ of real numbers.

Theorem A.1.7 *a) A countable union of countable sets is countable. b) Finite Cartesian products of countable sets is countable. c) Infinite Cartesian products of countable sets may not be countable. d)* $[0,1]$ *is not countable.*

Cantor's diagonal argument and the triangular enumeration are important steps in proving the theorem above.

Since rational numbers are the ratios of two integers, one may view rational numbers as a subset of the product space of countable spaces; thus, rational numbers are countable.

Definition A.1.13 *A space* X *is separable, if it contains a countable dense set.*

Separability informs us that for approximation purposes it suffices to work with a countable set, when the set is uncountable. Examples of separable sets are $\mathbb{R}$, and the set of continuous and bounded functions on a compact set metrized with the maximum distance between the functions.

Complete, separable and metrizable spaces form a very broad class of signal spaces. Such spaces are called *Polish metric* spaces when a metric is defined apriori. Borel subsets of such spaces are called *standard Borel* spaces.

B

On the Convergence of Random Variables

B.1 Limit Events and Continuity of Probability Measures

Given $A_1, A_2, \ldots, A_n, \cdots \in \mathcal{F}$, define:

$$\limsup_n A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$$

$$\liminf_n A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$$

For the superior limit, an element is in this set, if it is in infinitely many A_ns. For the inferior case, an element is in the limit, if it is in almost except for a finite number of A_ns. The limit of a sequence of sets exists if the above limits are equal. We have the following result:

Theorem B.1.1 *For a sequence of events A_n:*

$$P(\liminf_n A_n) \leq \liminf_n P(A_n) \leq \limsup_n P(A_n) \leq P(\limsup_n A_n)$$

We have the following regarding continuity of probability measures:

Theorem B.1.2 *(i)For a sequence of events A_n with $A_n \subset A_{n+1}$ for all n,*

$$\lim_{n\to\infty} P(A_n) = P(\cup_{n=1}^{\infty} A_n)$$

(ii) For a sequence of events A_n with $A_{n+1} \subset A_n$ for all n,

$$\lim_{n\to\infty} P(A_n) = P(\cap_{n=1}^{\infty} A_n)$$

B.2 Borel-Cantelli Lemma

Theorem B.2.1 *(i) If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\limsup_n A_n) = 0$. (ii) If $\{A_n\}$ are independent and if $\sum P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.*

Proof sketch. (i) For every $M \in \mathbb{N}$: $P(\limsup_n A_n) = P(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k) \leq P(\cup_{k=M}^{\infty} A_k) \leq \sum_{k=M}^{\infty} P(A_k)$. Therefore, as $M \to \infty$, the sum on the right is less than any $\epsilon > 0$. Since this ϵ is arbitrary, the result follows. (ii) Note that for every $M \in \mathbb{N}$, $P((\cup_{k=M}^{\infty} A_k)^c) = P(\cap_{k=M}^{\infty} A_k^c) = \prod_{k=M}^{\infty} P(A_k^c) = \prod_{k=M}^{\infty}(1 - P(A_k))$, where we use independence of the events. Thus, by Exercise 3.5.7, we have that $P((\cup_{k=M}^{\infty} A_k)^c) = 0$. Since for every M, $P(\cup_{k=M}^{\infty} A_k) = 1$, and the sets $\cup_{k=M}^{\infty} A_k$ are non-expanding, the result follows from Theorem B.1.2. $\diamond$

Exercise B.2.1 *Let $\{A_n\}$ be a sequence of independent events where A_n is the event that the nth coin flip is head. What is the probability that there are infinitely many heads if $P(A_n) = 1/n^2$?*

An important application of the above is the following:

Theorem B.2.2 *Let $Z_n, n \in \mathbb{N}$ and Z be random variables and for every $\epsilon > 0$,*

$$\sum_n P(|Z_n - Z| \geq \epsilon) < \infty.$$

Then,

$$P(\{\omega : Z_n(\omega) = Z(\omega)\}) = 1.$$

That is Z_n converges to Z with probability 1.

B.3 Convergence of Random Variables

B.3.1 Convergence almost surely (with probability 1)

Definition B.3.1 *A sequence of random variables X_n converges almost surely to a random variable X if $P(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$.*

B.3.2 Convergence in Probability

Definition B.3.2 *A sequence of random variables X_n converges in probability to a random variable X if $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon\}) = 0$ for every $\epsilon > 0$.*

B.3.3 Convergence in Mean-square

Definition B.3.3 *A sequence of random variables X_n converges in the mean-square sense to a random variable X if $\lim_{n\to\infty} E[|X_n - X|^2] = 0$.*

B.3.4 Convergence in Distribution

Definition B.3.4 *Let X_n be a random variable with cumulative distribution function F_n, and X be a random variable with cumulative distribution function F. A sequence of random variables X_n converges in distribution (or weakly) to a random variable X if $\lim_{n\to\infty} F_n(x) = F(x)$ for all points of continuity of F.*

Theorem B.3.1 *a) Convergence in almost sure sense implies in probability. b) Convergence in mean-square sense implies convergence in probability. c) If $X_n \to X$ in probability, then $X_n \to X$ in distribution.*

We also have partial converses for the above results:

Theorem B.3.2 *a) If $P(|X_n| \leq Y) = 1$ for some random variable Y with $E[Y^2] < \infty$, and if $X_n \to X$ in probability, then $X_n \to X$ in mean-square. b) If $X_n \to X$ in probability, there exists a subsequence X_{n_k} which converges to X almost surely. c) If $X_n \to X$ and $X_n \to Y$ in probability, mean-square, or almost surely; then $P(X = Y) = 1$.*

A sequence of random variables is uniformly integrable if:

$$\lim_{K\to\infty} \sup_n E[|X_n|1_{|X_n|\geq K}] = 0.$$

A sufficient condition for a sequence of random variables to be uniformly integrable is that there exists a function $g : \mathbb{R} \to \mathbb{R}$ with the property as $t \to \infty$, $\frac{g(t)}{t} \uparrow \infty$, so that $\sup_n E[g(X_n)] < \infty$: (to see this, note that $g(|X_n|) = \frac{|X_n|g(|X_n|)}{|X_n|} \geq \frac{g(K)}{K}|X_n|$, for $|X_n| \geq K$. Thus,

$$\lim_{K\to\infty} \sup_n E[|X_n|1_{|X_n|\geq K}] \leq \lim_{K\to\infty} \sup_n \frac{E[g(X_n)1_{|X_n|\geq K}]}{g(K)/K} \leq \lim_{K\to\infty} \sup_n \frac{E[g(X_n)]}{g(K)/K} \to 0.$$

Note also that, if $\{X_n\}$ is uniformly integrable, then, $\sup_n E[|X_n|] < \infty$.

Theorem B.3.3 *Under uniform integrability, convergence in almost sure sense implies convergence in mean-square.*

Theorem B.3.4 *If $X_n \to X$ in probability, there exists some subsequence X_{n_k} which converges to X almost surely.*

A further useful result is the following.

Theorem B.3.5 *[Skorohod's representation theorem] Let $X_n \to X$ in distribution. Then, there exists a sequence of random variables Y_n and Y such that, X_n and Y_n have the same cumulative distribution functions; X and Y have the same cumulative distribution functions and $Y_n \to Y$ almost surely.*

With the above, we can prove the following result.

Theorem B.3.6 *The following are equivalent: i) X_n converges to X in distribution. ii) $E[f(X_n)] \to E[f(X)]$ for all continuous and bounded functions f. iii) The characteristic functions $\Phi_n(u) := E[e^{iuX_n}]$ converge pointwise for every $u \in \mathbb{R}$.*

C

Some Remarks on Measurable Selections

As we observe in *Chapter 5*, in stochastic control measurability issues arise extensively both for the measurability of control policies as well as that of value functions/optimal costs. Theorem 5.1.1 and 5.2.1, and Lemma 5.2.4 are some examples where these were crucially utilized. In addition, we observed that the theory of martingales and filtration, the measurability properties are essential.

One particular aspect is to ensure that maps of the form:

$$J(x) := \inf_{u \in \mathbb{U}} c(x, u) \tag{C.1}$$

are measurable or at least Lebesgue-integrable.

Theorem C.0.1 *[Kuratowski Ryll-Nardzewski Measurable Selection Theorem] [202] [283] and [169, Theorem 2] Let $\mathbb{X}, \mathbb{U}$ be Polish spaces and $\Gamma = \{(x, \psi(x)), x \in \mathbb{X}\}$ where $\psi(x) \subset \mathbb{U}$ be such that, $\psi(x)$ is closed for each $x \in \mathbb{X}$ and Γ be a Borel measurable set in $\mathbb{X} \times \mathbb{U}$. Then, there exists at least one measurable function $f : \mathbb{X} \to \mathbb{U}$ such that $\{(x, f(x)), x \in \mathbb{X}\} \subset \Gamma$.*

A proof sketch is as follows for the case with $\mathbb{U} = \mathbb{R}_+$ and $\psi(x)$ is compact valued. With $n \in \mathbb{N}$, consider the infinite sequence of rationals $\{k/n; k \in \mathbb{Z}_+\}$. Consider $\psi^{-1}([\frac{k}{n}, \frac{k+1}{n}))$. Define the Borel set

$$X_{(k,n)} = \psi^{-1}([\frac{k}{n}, \frac{k+1}{n})) \setminus \cup_{m=1}^{k-1} \psi^{-1}([\frac{k}{n}, \frac{k+1}{n})).$$

Then, define a multi-function:

$$\psi^n(x) = \psi(x) \cap [\frac{k}{n}, \frac{k+1}{n}),$$

whenever $x \in X_{(k,n)}$. Now, each ψ^n is a multi-function. Take $n \to \infty$, in this case, since $\psi(x)$ is closed, each converging subsequence $u_{n_k} \in X_{(k,n)}$ is so that $\lim_{n_k \to \infty} u_{n_k} \in \psi(x)$ (by the closed property). Therefore, for each x, the limit $\lim_{n \to \infty} \psi^n(x)$ is well-defined and single-valued, and is placed in $\psi(x)$. This approach can be generalized.

This result was utilized in *Chapter 5* (see Lemma 5.2.4). Recall also the relationship of the argument with that in the proof of Theorem 5.1.1.

In the following, we assume that the spaces considered are Polish. A function f is μ-measurable if there exists a Borel measurable function g which agrees with f μ-a.e. A function that is μ-measurable for every probability measure is called universally measurable.

A measurable image of a Borel set is called an *analytic* set [117].

Fact C.0.1 *The image of a Borel set under a measurable function, and hence an analytic set, is universally measurable.*

Remark C.1. We note that in some texts, an analytic set is defined as the continuous image of a Borel set. However, as [117] notes, one could always express the image of a Borel set A under a measurable function $f : \mathbb{X} \to \mathbb{Y}$ as a projection (which is a continuous map) of $(A, f(A))$ onto $\mathbb{Y}$.

The integral of a universally measurable function is well-defined and is equal to the integral of a Borel measurable function which is μ-almost equal to that function. While applying dynamic programming, we often seek to establish the existence of measurable functions through the operation:

$$J_t(x_t) = \inf_{u \in \mathbb{U}(x_t)} \left(c(x,u) + \int J_{t+1}(x_{t+1}) Q(dx_{t+1}|x_t, u) \right)$$

However, we need a stronger condition that universal measurability for the recursions to be well-defined. A function f is called lower semi-analytic if $\{x : f(x) < c\}$ is analytic for each scalar c.

Theorem C.0.2 *[117] Let $i : \mathbb{X} \to 2^{\mathbb{S}}$ (that is, i maps $\mathbb{X}$ to subsets of $\mathbb{S}$) be such that i^{-1} is Borel measurable, and $f : \mathbb{S} \to \mathbb{R}$ be measurable. Then:*

$$v(x) = \inf_{z : z \in i(x)} f(z)$$

is lower semi-analytic.

Observe that (see p. 85 of [117])

$$\{x : v(x) < c\} = i^{-1}(\{z : f(z) < c\})$$

The set $\{z : f(z) < c\}$ is Borel, and thus if i^{-1} is also Borel, it follows that v is lower semi-analytic. We require then that $i^{-1} : \mathbb{S} \to \mathbb{X}$ to be Borel. Consider now the following application.

Theorem C.0.3 *Consider $G = \{(x,u) : u \in \mathbb{U}(x)\}$ which is a Borel measurable set. The map,*

$$v(x) = \inf_{(x,z) \in G} v(x,z),$$

is lower semi-analytic.

Proof. The graph G is measurable. It follows that$\{x : v(x) < c\} = i^{-1}(\{(x,z) : v(x,z) < c\})$, where i^{-1} is the projection of G onto $\mathbb{X}$, which is a continuous operation; the image may not be measurable but as a measurable mapping of a Borel set, it is analytic. As a result v is lower semi-analytic. ◇

Theorem C.0.4 *Lower semi-analytic functions are universally measurable.*

Implication: Dynamic programming can be carried out for such expressions. In particular, the following is due to Bertsekas and Shreve [37, Chapter 7]:

Theorem C.0.5 *The following hold:*

(i) Let E_1, E_2 be Borel and $g : E_1 \times E_2 \to \mathbb{R}$ be lower semi-analytic. Then,

$$h(e_1) = \inf_{e_2 \in E_2} g(e_1, e_2)$$

is lower semi-analytic.

(ii) Let E_1, E_2 be Borel and $g : E_1 \times E_2 \to \mathbb{R}$ be lower semi-analytic. Let $Q(de_2|e_1)$ be a stochastic kernel. Then,

$$f(e_1) := \int g(e_2) Q(de_2|e_1)$$

is lower semi-analytic.

We note that the second result would not be correct if g is only taken to be universally measurable. The result above ensures that we can follow the dynamic programming arguments in an inductive manner under conditions that are less restrictive than the conditions stated in the measurable selection conditions. These then imply the existence of ϵ-optimal solutions (possibly universally measurable) [288].

Building on this discussion, and the material in *Chapter 5*, we summarize three useful results in the following.

Fact C.0.2 *Consider (C.1).*

(i) If c is continuous on $\mathbb{X} \times \mathbb{U}$ and $\mathbb{U}$ is compact, then J is continuous and there exists an optimal measurable policy.

(ii) If the measurable c is continuous on $\mathbb{U}$ for every x, and $\mathbb{U}$ is compact, then J is measurable (Prop D.5 in [165] and Himmelberg and Schäl [283]); see Theorem 5.2.4. Furthermore, there exists an optimal measurable policy.

(iii) [37, Prop. 7.47 and 7.50] If c is measurable on $\mathbb{X} \times \mathbb{U}$ and $\mathbb{U}$ is Borel, then J is lower semi-analytic. Furthermore, there exists a near optimal universally measurable function.

Compactness of $\mathbb{U}$ is a crucial component for some of these results. However, as discussed in Section 5.2, $\mathbb{U}(x)$ may be allowed to depend on x, for item (ii) under the assumption that the graph $G = \{(x, u) : u \in \mathbb{U}(x)\}$ defined above is Borel, and $\mathbb{U}(x)$ is compact for every x; see p. 182 in [165] (and also [169], [283], [125] and [202], among others). For (i), this relaxation also requires that the set valued map $\mathbb{U}(x)$ is upper semi-continuous: let $x_n \to x$, then for every sequence $u_n \in \mathbb{U}(x_n)$, there exists a subsequence which converges to some u where every such limit u satisfies $u \in \mathbb{U}(x)$.

D

On Spaces of Probability Measures

In this section we present various topologies, and when applicable, several metrics on the sets of probability measures.

D.1 Convergence of Sequences of Probability Measures

Let $\mathbb{X}$ be a Polish space and let $\mathcal{P}(\mathbb{X})$ denote the family of all probability measures on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. Let $\{\mu_n,\, n \in \mathbb{N}\}$ be a sequence in $\mathcal{P}(\mathbb{X})$.

The sequence $\{\mu_n\}$ is said to converge to $\mu \in \mathcal{P}(\mathbb{X})$ *weakly* if

$$\int_{\mathbb{X}} c(x)\mu_n(dx) \to \int_{\mathbb{X}} c(x)\mu(dx) \tag{D.1}$$

for every continuous and bounded $c : \mathbb{X} \to \mathbb{R}$. [1]

On the other hand, $\{\mu_n\}$ is said to converge to $\mu \in \mathcal{P}(\mathbb{X})$ *setwise* if

$$\int_{\mathbb{X}} c(x)\mu_n(dx) \to \int_{\mathbb{X}} c(x)\mu(dx)$$

for every measurable and bounded $c : \mathbb{X} \to \mathbb{R}$. Setwise convergence can also be defined through pointwise convergence on Borel subsets of $\mathbb{X}$ (see, e.g., [168]), that is

$$\mu_n(A) \to \mu(A), \quad \text{for all } A \in \mathcal{B}(\mathbb{X})$$

since the space of simple functions are dense in the space of bounded and measurable functions under the supremum norm.

For two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$, the *total variation* metric is given by

$$\begin{aligned} \|\mu - \nu\|_{TV} &:= 2 \sup_{B \in \mathcal{B}(\mathbb{X})} |\mu(B) - \nu(B)| \\ &= \sup_{f:\, \|f\|_\infty \le 1} \left| \int f(x)\mu(dx) - \int f(x)\nu(dx) \right|, \end{aligned} \tag{D.2}$$

where the supremum is over all measurable real f such that $\|f\|_\infty = \sup_{x \in \mathbb{X}} |f(x)| \le 1$. A sequence $\{\mu_n\}$ is said to converge to $\mu \in \mathcal{P}(\mathbb{X})$ in total variation if $\|\mu_n - \mu\|_{TV} \to 0$.

[1] It is important to emphasize that what is typically studied in probability as weak convergence is not the exact weak convergence notion used in functional analysis: The topological dual space of the set of probability measures does not only consist of expectations of continuous and bounded functions. However, the dual space of the space of continuous and bounded functions with the supremum norm does admit a representation in terms of expectations [220]; hence, the weak convergence here is in actuality the weak* convergence in analysis and distribution theory.

Setwise convergence is equivalent to pointwise convergence on Borel sets whereas total variation requires uniform convergence on Borel sets. Thus these three convergence notions are in increasing order of strength: convergence in total variation implies setwise convergence, which in turn implies weak convergence.

On the other hand, total variation is a stringent notion for convergence. For example a sequence of discrete probability measures never converges in total variation to a probability measure which admits a density function with respect to the Lebesgue measure and such a space is not separable. Setwise convergence also induces a topology on the space of probability measures and channels which is not easy to work with since the space under this convergence is not metrizable [143, p. 59].

However, the space of probability measures on a complete, separable, metric (Polish) space endowed with the topology of weak convergence is itself a complete, separable, metric space [42].

There are various ways to metrize weak convergence. One immediate metric builds on the following reasoning: One can construct (since the space of continuous functions on a compact set is separable under the supremum norm) a countable collection of continuous functions $\{c_k, k \in \mathbb{N}\}$ such that it suffices to only consider these functions in (D.1) to establish weak-convergence. We can thus use these *weak-convergence determining* functions (see e.g. [120, Theorem 3.4.5]) to define a countable collection of semi-norms $d_k(\mu,\nu) := |\int c_k(x)\mu(dx) - \int c_k(x)\nu(dx|$, and from these we can construct a locally convex space which is metrizable. Thus, we have the following metric which metrizes the weak topology:

$$\rho(\mu,\nu) = \sum_{m=1}^{\infty} 2^{-(m+1)} \left| \int_S f_m(x)\mu(dx) - \int_S f_m(x)\nu(dx) \right|, \tag{D.3}$$

where $\{f_m\}_{m\geq 1}$ is an appropriate sequence of continuous and bounded functions such that $\|f_m\|_\infty \leq 1$ for all $m \geq 1$ (see [249, Theorem 6.6, p. 47]).

The Prohorov metric [42] also can be used to metrize this convergence topology.

As a more practical metric, the Wasserstein metric can also be used (for compact $\mathbb{X}$) to metrize the weak convergence space topology.

Definition D.1.1 (Wasserstein metric) *The* Wasserstein metric *of order p, $1 \leq p < \infty$, for two distributions $\mu, \nu \in \mathcal{P}(\mathbb{X})$ with finite pth moments (thus defined only on such a subset of $\mathcal{P}(\mathbb{X})$) is defined as*

$$W_p(\mu,\nu) = \inf_{\eta \in \mathcal{H}(\mu,\nu)} \left(\int_{\mathbb{X}\times\mathbb{X}} \eta(dx,dy)\|x-y\|^p \right)^{\frac{1}{p}},$$

where $\mathcal{H}(\mu,\nu)$ denotes the set of probability measures on $\mathbb{X}\times\mathbb{X}$ with first marginal μ and second marginal ν, and $\|\cdots\|$ is a norm.

For compact $\mathbb{X}$, the Wasserstein distance of order p metrizes the weak topology on the set of probability measures on $\mathbb{X}$ (see [320, Theorem 6.9]; one can also see the connection via Theorem B.3.5). For non-compact $\mathbb{X}$, weak convergence combined with convergence of moments up to order p (that is of $\int \mu_n(dx)\|x\|^p \to \int \mu(dx)\|x\|^p$) is equivalent to convergence in W_p. Finally, the bounded-Lipschitz metric ρ_{BL} [320, p.109] can also be used to metrize weak convergence:

$$\rho_{BL}(\mu,\nu) = \sup_{\|f\|_{BL}\leq 1} \left| \int_{\mathbb{X}} f(e)\mu(de) - \int_{\mathbb{X}} f(e)\nu(de) \right|, \tag{D.4}$$

where

$$\|f\|_{BL} := \|f\|_\infty + \sup_{e\neq e'} \frac{f(e)-f(e')}{d_{\mathbb{X}}(e,e')},$$

and $d_{\mathbb{X}}$ is the metric on $\mathbb{X}$.

We note that W_1 can equivalently be written as [320, Remark 6.5]:

$$W_1(\mu,\nu) := \sup_{\|f\|_{Lip}\leq 1}\left|\int_{\mathbb{X}} f(e)\mu(de) - \int_{\mathbb{X}} f(e)\nu(de)\right|,$$

where

$$\|f\|_{Lip} := \sup_{e\neq e'} \frac{f(e)-f(e')}{d_{\mathbb{X}}(e,e')}.$$

Comparing this with (D.4), it follows that

$$\rho_{BL} \leq W_1. \tag{D.5}$$

Another important distance measure (though not a metric) that is commonly used is relative entropy:

Definition D.1.2 *For two probability measures P and Q, relative entropy is defined as $D(P\|Q) = \int \log \frac{dP}{dQ} dP = \int \frac{dP}{dQ} \log \frac{dP}{dQ} dQ$ where $P \ll Q$ and $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative of P with respect to Q.*

Total variation is related to relative entropy via Pinsker's inequality [94]: $\|P-Q\|_{TV} \leq \sqrt{\frac{2}{\log(e)} D(P\|Q)}$. This also shows that convergence in relative entropy implies that under total variation.

Weak convergence is very important in applications of stochastic control and probability in general. Prohorov's theorem [111] provides a way to characterize compactness properties under weak convergence.

D.2 Some Measurability Results on Spaces of Probability Measures

Weak convergence topology leads to important measurability properties, as we discuss in the following two theorems. The first one appears in [4] (see Theorem 15.13 in [4] or p. 215 in [53]).

Theorem D.2.1 *Let $\mathbb{S}$ be a Polish space and M be the set of all measurable and bounded functions $f : \mathbb{S} \to \mathbb{R}$. Then, for any $f \in M$, the integral*

$$\int \pi(dx) f(x)$$

defines a measurable function on $\mathcal{P}(\mathbb{S})$ under the topology of weak convergence.

This is a useful result since it allows us to define measurable functions in integral forms on the space of probability measures when we work with the topology of weak convergence. The second useful result follows from Theorem D.2.1, [110, Theorem 2.1] and [37, Proposition 7.25].

Theorem D.2.2 *Let $\mathbb{S}$ be a Polish space. A function $F : \mathcal{P}(\mathbb{S}) \to \mathcal{P}(\mathbb{S})$ is measurable on $\mathcal{B}(\mathcal{P}(\mathbb{S}))$ (under weak convergence), if for all $B \in \mathcal{B}(\mathbb{S})$ $(F(\cdot))(B) : \mathcal{P}(\mathbb{S}) \to \mathbb{R}$ is measurable under weak convergence on $\mathcal{P}(\mathbb{S})$, that is for every $B \in \mathcal{B}(\mathbb{S})$, $(F(\pi))(B)$ is a measurable function when viewed as a function from $\mathcal{P}(\mathbb{S})$ to $\mathbb{R}$.*

D.3 A Generalized Dominated Convergence Theorem

Under weak and setwise convergences, we can arrive at generalized forms of the dominated convergence theorem. In particular, from [212, Theorem 3.5] and [287, Theorem 3.5], we have the following:

Theorem D.3.1 *The following hold:*

(i) Suppose that $\{\mu_n\}_n \subset \mathcal{P}(\mathbb{X})$ converges weakly to some μ. For a bounded real valued sequence of functions $\{f_n\}_n$ such that $\|f_n\|_\infty < C$ for all $n > 0$ with $C < \infty$, if $\lim_{n\to\infty} f_n(x_n) = f(x)$ for all $x_n \to x$, i.e. f_n continuously converges to f, then

$$\lim_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu_n(dx) = \int_{\mathbb{X}} f(x)\mu(dx).$$

(ii) Suppose that $\{\mu_n\}_n \subset \mathcal{P}(\mathbb{X})$ converges setwise to some μ. For a bounded real valued sequence of functions $\{f_n\}_n$ such that $\|f_n\|_\infty < C$ for all $n > 0$ with $C < \infty$, if $\lim_{n\to\infty} f_n(x) = f(x)$ for all x, i.e. f_n pointwise converges to f, then

$$\lim_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu_n(dx) = \int_{\mathbb{X}} f(x)\mu(dx).$$

D.4 The w-s Topology

Let, as before, $\mathbb{X}$ and $\mathbb{Y}$ be Polish spaces.

Definition D.4.1 *The w-s topology on the set of probability measures $\mathcal{P}(\mathbb{X} \times \mathbb{Y})$ is the coarsest topology under which $\int f(x, y)\mu(dx, dy) : \mathcal{P}(\mathbb{X} \times \mathbb{Y}) \to \mathbb{R}$ is continuous for every measurable and bounded $f(x, y)$ which is continuous in y for every x (but unlike the weak topology, f does not need to be continuous in x).*

Theorem D.4.1 *[284, Theorem 3.10] [27, Theorem 2.5] Let $\mu_n \in \mathcal{P}(\mathbb{X} \times \mathbb{Y})$. If $\mu_n \to \mu$ weakly where the marginals $\mu_n(dx \times \mathbb{Y}) \to \mu(dx \times \mathbb{Y})$ setwise, then the convergence $\mu_n \to \mu$ is also in the w-s sense.*

D.5 Lusin's Theorem

Lusin's theorem is a very consequential result in mathematical analysis.

Theorem D.5.1 *[111, Theorem 7.5.2] Let $(\mathbb{X}, T)$ be any topological space and μ a finite, closed regular Borel measure on $\mathbb{X}$. Let $(\mathbb{S}, d)$ be a separable metric space and let f be a Borel-measurable function from $\mathbb{X}$ into $\mathbb{S}$. Then for any $\epsilon > 0$ there is a closed set $F \subset \mathbb{X}$ such that $\mu(\mathbb{X} \setminus F) < \epsilon$ and the restriction of f to F is continuous.*

We also recall Tietze's extension theorem, which is often used in conjunction with Lusin's theorem to construct a continuous extension of the continuous function defined on F in Theorem D.5.1 to $\mathbb{X}$.

Theorem D.5.2 *[114, Theorem 4.1][Tietze's extension theorem] Let $\mathbb{X}$ be an arbitrary metric space, A a closed subset of $\mathbb{X}$, L a locally convex linear space, and $f : A \to L$ a continuous map. Then there exists a continuous function $f_C : \mathbb{X} \to L$ such that $f_C(a) = f(a)\ \forall a \in A$. Furthermore, the image of f_C satisfies $f_C(\mathbb{X}) \subset$ [convex hull of $f(A)$].*

E

Relaxed Control Topologies

In deterministic as well as stochastic control theory, relaxed or randomized control policies allow for versatility in mathematical analysis, leading to continuity, compactness, convexity and approximation properties, in a variety of system models, cost criteria, and information structures.

Within the relaxed/randomized control framework, with $\mathbb{X}$ a state space, $\mathbb{U}$ a control space and with an $\mathbb{X}$-valued random variable $X \sim \mu$, instead of considering the set of deterministic admissible policies:

$$\mathit{\Gamma} = \Big\{ \gamma : \gamma \text{ is a measurable function from } \mathbb{X} \text{ to } \mathbb{U} \Big\}, \tag{E.1}$$

one considers

$$\mathit{\Gamma}_R = \Big\{ \gamma : \gamma \text{ is a measurable function from } \mathbb{X} \text{ to } \mathcal{P}(\mathbb{U}) \Big\}, \tag{E.2}$$

where $\mathcal{P}(\mathbb{U})$ is endowed with the Borel σ-algebra generated by the weak convergence topology.

On $\mathit{\Gamma}_R$, two commonly studied topologies are the following.

E.1 Young Topology on Control Policies

A prominent approach since Young's seminal paper [340] has been via the study of topologies on *Young measures* defined by randomized/relaxed controls, where one views policies to be identified with probability measures defined on a product space with a fixed marginal at an input/state space (typically taken to be the Lebesgue measure in optimal deterministic control) [228, 340], [75, Section 2.1], [322, p. 254], [224], [28, Theorem 2.2]. Thus, under the Young topology, one associates with $\mathit{\Gamma}_R$ in (E.2) the probability measure induced on the product space $\mathbb{X} \times \mathbb{U}$ with a fixed marginal μ on $\mathbb{X}$.

The generalization to stochastic control problems by considering more general input measures has been commonplace, with applications also to partially observed stochastic control and decentralized stochastic control.

To appreciate the Young topology on control policies, we first present a relevant representation result (see Borkar [56]). Let $\mathbb{X}, \mathbb{M}$ be Borel spaces. Let $\mathcal{P}(\mathbb{X})$ denote the set of probability measures on $\mathbb{X}$. Consider the set of probability measures

$$\begin{aligned} \mathit{\Theta} := \Big\{ & \zeta \in \mathcal{P}(\mathbb{X} \times \mathbb{M}) : \\ & \zeta(dx, dm) = P(dx)\, Q^f(dm|x), Q^f(\cdot|x) = 1_{\{f(x) \in \cdot\}}, f : \mathbb{X} \to \mathbb{M} \Big\} \end{aligned} \tag{E.3}$$

on $\mathbb{X} \times \mathbb{M}$ with fixed input marginal P on $\mathbb{X}$ and with the stochastic kernel from $\mathbb{X}$ to $\mathbb{M}$ realized by any measurable function $f : \mathbb{X} \to \mathbb{M}$. We equip this set with the weak convergence topology. This set is the (Borel measurable) set of the extreme

points of the set of probability measures on $\mathbb{X} \times \mathbb{M}$ with a fixed marginal P on $\mathbb{X}$. For compact $\mathbb{M}$, the Borel measurability of Θ follows [251] since the set of probability measures on $\mathbb{X} \times \mathbb{M}$ with a fixed marginal P on $\mathbb{X}$ is a *convex and compact* set in a complete separable metric space, and therefore, the set of its extreme points is Borel measurable; measurability for the non-compact case follows from [56, Lemma 2.3]. Furthermore, given a fixed marginal P on $\mathbb{X}$, any stochastic kernel Q from $\mathbb{X}$ to $\mathbb{M}$ can almost surely be identified by a probability measure $\Xi \in \mathcal{P}(\Theta)$ such that

$$Q(\cdot|x) = \int_{\Theta} \Xi(dQ^f)\, Q^f(\cdot|x). \tag{E.4}$$

In particular, a randomized policy can thus be viewed as a mixture of deterministic policies.

Definition E.1.1 Convergence of Policies in Γ_S under Young topology at reference (input) measure μ. *Let μ be a σ-finite measure. A sequence of stationary policies $\gamma_n \to \gamma \in \Gamma_S$ at input μ if the joint measure $(\mu\gamma_n) \to (\mu\gamma)$ weakly at input P, i.e., for every continuous and bounded $g : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ with*

$$\int \mu(dx) \sup_{u \in \mathbb{U}} |g(x,u)| < \infty,$$

$$\int \mu(dx) \Big(\gamma_n(du|x) g(x,u) \Big) \to \int \mu(dx) \Big(\gamma(du|x) g(x,u) \Big) \tag{E.5}$$

With the above, we observe that the Young topology allows for a convex and compact formulation.

We also note that in the above, the reference measure does not need to be a probability measure (and we view weak convergence to be one on signed measures that defines a locally convex space with (E.5) defining the semi-norms).

We finally note that since the marginal of the joint measure $(\mu\gamma_n)$ on $\mathbb{X}$ is fixed, the convergence in (E.5) is also in the setwise-weak sense (with $g(x,u)$ bounded but only continuous in u for every fixed $x \in \mathbb{X}$, see Section D.4), following Lemma D.4.1.

E.2 Borkar (Weak*) Topology on Control Policies

In the stochastic setup, another topology is the one introduced by Borkar on relaxed controls [55] (see also [15, Section 2.4], and [45] which [55] notes to be building on), formulated as a weak* topology on randomized policies viewed as maps from states/measurements to the space of signed measures with bounded variation $\mathcal{M}(\mathbb{U})$ of which probability measures $\mathcal{P}(\mathbb{U})$ is a subset. We also refer the reader to [105, 122] for further references on such a weak* formulation on relaxed controls, in particular when instead of countably additive signed measures, finitely additive such measures are also considered.

Under the Borkar topology one studies Γ_R in (E.2), with a weak* topology formulation, as a bounded subset of the set of maps from $\mathbb{X}$ to the space of signed measures with finite variation viewed as the topological dual of continuous functions vanishing at infinity, leading to a compact metric space by the Banach-Alaoglu theorem [134, Theorem 5.18] (and thus, as the unit ball of $L_\infty(\mathbb{X}, \mathcal{M}(\mathbb{U})) = (L_1(\mathbb{X}, C_0(\mathbb{U})))^*$ is compact under the weak* topology, this leads to a compact metric topology on relaxed control policies). We note that the presentations in [55, Section 3] and [15, Section 2.4] are slightly different, though the induced topologies are identical. An equivalent representation of this topology is given in [15, Lemma 2.4.1] (see also [55, Lemma 3.1]).

See [12, 55, 256] for a detailed analysis on some implications in stochastic control theory in continuous time (such as continuity of expected cost in control policies [55], approximation results [256] under various cost criteria, and continuity of invariant measures of diffusions in control policies [12]).

Definition E.2.1 *[55] [15, Lemma 2.4.1]* **Convergence of Policies in Γ_S under Borkar topology.** *With $\mathbb{X} = \mathbb{R}^d$, a sequence of stationary policies $\gamma_n \to \gamma \in \Gamma_S$ in the Borkar topology if for every continuous and bounded $g : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ and every $f \in L_1(\mathbb{X}) \cap L_2(\mathbb{X})$*

$$\int f(x) \int \gamma_n(du|x) g(x,u) dx \to \int f(x) \int \gamma_n(du|x) \gamma(du|x) g(x,u) dx \tag{E.6}$$

Building on Lemma D.4.1, the functions g in Definition E.2.1 may be relaxed to be continuous only in u for every $x \in \mathbb{X}$.

While $\mathbb{X}$ was taken to be $\mathbb{R}^n$ in [55], Saldi [268] generalized this to setups where $\mathbb{X}$ is a general standard Borel space with a fixed input (probability) measure. The generalization by Saldi [268] is the following, where the input space $\mathbb{X}$ is arbitrary standard Borel, though with a fixed input measure μ: Let $C_0(\mathbb{U})$ be the Banach space of all continuous real functions on $\mathbb{U}$ vanishing at infinity, endowed with the norm $\|g\|_\infty = \sup_{u\in\mathbb{U}} |g(u)|$. [268] formulated this topology via noting that with $L_1\big(\mu, C_0(\mathbb{U})\big)$ denoting the set of all Bochner-integrable functions from $\mathbb{X}$ to $C_0(\mathbb{U})$ endowed with the norm

$$\|f\|_1 \coloneqq \int_{\mathbb{X}} \|f(x)\|_\infty \, \mu(dx),$$

using the fact that $C_0(\mathbb{U})^* = \mathcal{M}(\mathbb{U})$, and that the topological dual of $\big(L_1\big(\mu, C_0(\mathbb{U}\big), \|\cdot\|_1\big)$ can be identified with $\big(L_\infty\big(\mu, \mathcal{M}(\mathbb{U})\big), \|\cdot\|_\infty\big)$ [77, Theorem 1.5.5, p. 27] (see also [105, 122] for further context on such duality results, in particular when instead of countably additive signed measures, finitely additive such measures are considered); that is,

$$L_1\big(\mu, C_0(\mathbb{U})\big)^* = L_\infty\big(\mu, \mathcal{M}(\mathbb{U})\big).$$

E.3 Some Properties of Young and Borkar topologies

Lemma E.3.1 *[347] Let $\eta \ll \kappa$, where κ is a σ-finite and η is a finite measure. Then, $\gamma_n \to \gamma$ under Young topology at input κ implies $\gamma_n \to \gamma$ (under Young topology) at input η.*

Theorem E.3.1 *[347] Let $\mathbb{X} = \mathbb{R}^n$ and λ be the Lebesgue measure. Consider convergence in Young topology at some input probability measure ψ.*

(i) If $\psi \ll \lambda$ with $h(x) = \frac{d\psi}{d\lambda}(x)$ is positive everywhere, then convergence in Young topology at input measure ψ implies convergence in Borkar topology.

(ii) If $\psi \ll \lambda$, then convergence in Borkar topology implies convergence in Young topology at input ψ.

In [347], several results on the significance of these topologies on existence of optimal policies and approximations (on near optimality of continuous policies or quantized policies in both measurement and action) have been presented.

References

1.

2. J. Abounadi, D. Bertsekas, and V.S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

3. M. Aicardi, F. Davoli, and R. Minciardi. Decentralized optimal control of Markov chains with a common past information set. *IEEE Transactions on Automatic Control*, 32:1028–1031, November 1987.

4. C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis*. Berlin, Springer, 3rd ed., 2006.

5. A. Almudevar. A stochastic contraction mapping theorem. *Systems & Control Letters*, 174:105482, 2023.

6. A. Almudevar. and E. F. Arruda. Optimal approximation schedules for a class of iterative algorithms, with an application to multigrid value iteration. *IEEE Transactions on Automatic Control*, 57:3132–3146, 2012.

7. E. Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, Boca Raton, FL, 1999.

8. M. Andersland and D. Teneketzis. Information structures, causality, and non-sequential stochastic control, I: design-independent properties. *SIAM J. Control and Optimization*, 30:1447 – 1475, 1992.

9. M. Andersland and D. Teneketzis. Information structures, causality, and non-sequential stochastic control, II: design-dependent properties. *SIAM J. Control and Optimization*, 32:1726 – 1751, 1994.

10. B. D. O. Anderson and J. B. Moore. Optimal filtering. *Englewood Cliffs*, 21, 1979.

11. B. D. O. Anderson and J. B. Moore. *Optimal control: Linear quadratic methods*. Courier Corporation, 2007.

12. A. Arapostathis and V. S. Borkar. Uniform recurrence properties of controlled diffusions and applications to optimal control. *SIAM Journal on Control and Optimization*, 48(7):4181–4223, 2010.

13. A. Arapostathis and V. S. Borkar. Average cost optimal control under weak ergodicity hypotheses: Relative value iterations. *Arxiv preprints*, 1902.01048, 2019.

14. A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM J. Control and Optimization*, 31:282–344, 1993.

15. A. Arapostathis, V. S. Borkar, and M. K. Ghosh. *Ergodic Control of Diffusion Processes*, volume 143. Cambridge University Press, 2012.

16. A. Arapostathis and S. Yüksel. Convex analytic method revisited: Further optimality results and performance of deterministic policies in average cost stochastic control. *Journal of Mathematical Analysis and Applications*, 517(2):126567, 2023.

17. L. Arnold and W. Kliemann. On unique ergodicity for degenerate diffusions. *Stochastics: an international journal of probability and stochastic processes*, 21(1):41–61, 1987.

18. E. F. Arruda, F. Ourique, J. Lacombe, and A. Almudevar. Accelerating the convergence of value iteration by using partial transition functions. *European Journal of Operational Research*, 229:190–198, 2013.

19. M. Athans. Survey of decentralized control methods. Washington D.C, 1974. 3rd NBER/FRB Workshop on Stochastic Control.

20. K. B. Athreya and P. Ney. A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978.

21. R. J. Aumann. Mixed and behavior strategies in infinite extensive games. Technical report, Princeton University NJ, 1961.

22. R. J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4:1236 – 1239, 1976.

23. R. J. Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18, 1987.

24. J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 24(3):601–632, 2020.

25. A. Bain and D. Crisan. *Fundamentals of stochastic filtering*, volume 3. Springer, 2009.

26. W. L. Baker. *Learning via Stochastic Approximation in Function Space*. PhD Dissertation, Harvard University, Cambridge, MA, 1997.

27. E. J. Balder. On ws-convergence of product measures. *Mathematics of Operations Research*, 26(3):494–518, 2001.

28. E.J. Balder. Generalized equilibrium results for games with incomplete information. *Mathematics of Operations Research*, 13(2):265–276, 1988.

29. Y. Bar-Shalom and E. Tse. Dual effect, certainty equivalence, and separation in stochastic control. *IEEE Transactions on Automatic Control*, 19(5):494–500, October 1974.

30. Y. Bar-Shalom and E. Tse. Dual effect certainty equivalence and separation in stochastic control. *IEEE Transactions on Automatic Control*, 19:494–500, October 1974.

31. E. Bayraktar, Yan Y. Dolinsky, and J. Guo. Continuity of utility maximization under weak convergence. *Mathematics and Financial Economics*, pages 1–33, 2020.

32. M. Beiglböck and D. Lacker. Denseness of adapted processes among causal couplings. *arXiv*, pages arXiv–1805, 2018.

33. V. E. Beneš. Existence of optimal stochastic control laws. *SIAM Journal on Control*, 9(3):446–472, 1971.

34. A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

35. D. Bertsekas. *Dynamic Programming and Optimal Control Vol. 1*. Athena Scientific, 2000.

36. D. P. Bertsekas. *Dynamic Programming and Stochastic Optimal Control*. Academic Press, New York, New York, 1976.

37. D. P. Bertsekas and S. Shreve. *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, New York, 1978.

38. D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

39. D.P. Bertsekas. Convergence of discretization procedures in dynamic programming. *IEEE Trans. Autom. Control*, 20(3):415–419, Jun. 1975.

40. D.P Bertsekas. A new value iteration method for the average cost dynamic programming problem. *SIAM journal on control and optimization*, 36(2):742–759, 1998.

41. D.P. Bertsekas and J.N. Tsitsiklis. Neuro-dynamic programming. Athena Scientific, 1996.

42. P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1968.

43. P. Billingsley. *Probability and Measure*. Wiley, New York, 1995.

44. P. Billingsley. *Probability and Measure*. Wiley, 3rd edition, 1995.

45. J.-M. Bismut. *Théorie Probabiliste du Contrôle des Diffusions*, volume 181. 1973.

46. J.-M. Bismut. Partially observed diffusions and their control. *SIAM Journal on Control and Optimization*, 20(2):302–309, 1982.

47. D. Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, pages 719–726, 1962.

48. D. Blackwell. Memoryless strategies in finite-stage dynamic programming. *Annals of Mathematical Statistics*, 35:863–865, 1964.

49. D. Blackwell. The stochastic processes of Borel gambling and dynamic programming. *Annals of Statistics*, pages 370–374, 1976.

50. D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.

51. D. Blackwell and C. Ryll-Nadrzewski. Non-existence of everywhere proper conditional distributions. *Annals of Mathematical Statistics*, 34:223–225, 1963.

52. R. K. Boel, M. R. James, and I. R. Petersen. Robustness and risk-sensitive filtering. *IEEE Transactions on Automatic Control*, 47(3):451–461, 2002.

53. V. I. Bogachev. *Measure Theory*. Springer-Verlag, Berlin, 2007.

54. V. S. Borkar. The probabilistic structure of controlled diffusion processes. *Acta Applicandae Mathematica*, 11(1):19–48, 1988.

55. V. S. Borkar. A topology for Markov controls. *Applied Mathematics and Optimization*, 20(1):55–62, 1989.

56. V. S. Borkar. White-noise representations in stochastic realization theory. *SIAM J. on Control and Optimization*, 31:1093–1102, 1993.

57. V. S. Borkar. *Probability Theory: An Advanced Course*. Springer, New York, 1995.

58. V. S. Borkar. Average cost dynamic programming equations for controlled Markov chains with partial observations. *SIAM J. Control Optim.*, 39(3):673–681, 2000.

59. V. S. Borkar. Convex analytic methods in Markov decision processes. In *Handbook of Markov Decision Processes, E. A. Feinberg, A. Shwartz (Eds.)*, pages 347–375. Kluwer, Boston, MA, 2001.

60. V. S. Borkar. Dynamic programming for ergodic control with partial observations. *Stochastic Processes and their Applications*, 103:293–310, 2003.

61. V. S. Borkar. Ergodic control of diffusion processes. In *Proceedings ICM*, 2006.

62. V. S. Borkar. Dynamic programming for ergodic control of Markov chains under partial observations: A correction. *SIAM J. Control Optim.*, 45(6):2299–2304, 2007.

63. V. S. Borkar and A. Budhiraja. A further remark on dynamic programming for partially observed Markov processes. *Stochastic Processes and their Applications*, 112:79–93, 2004.

64. V. S. Borkar and S. P. Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control and Optimization*, pages 447–469, December 2000.

65. V. S. Borkar and P. Varaiya. Asymptotic agreement in distributed estimation. *IEEE Transactions Automatic Cotrol*, 27:650–655, June 1982.

66. Vivek S Borkar and Mrinal K Ghosh. Ergodic control of multidimensional diffusions i: The existence results. *SIAM Journal on Control and Optimization*, 26(1):112–126, 1988.

67. V.S. Borkar. Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, volume 5, 2010.

68. V.S. Borkar and S.P. Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

69. P. Bougerol. Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization*, 31(4):942–959, 1993.

70. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

71. A. Brandenburger and E. Dekel. Common knowledge with probability 1. *J. Mathematical Economics*, 16:237–245, 1987.

72. R. W. Brockett. Asymptotic stability and feedback stabilization. *Differential Geometric Control Theory*, 27(1):181–191, 1983.

73. A. Budhiraja. On invariant measures of discrete time filters in the correlated signal-noise case. *The Annals of Applied Probability*, 12(3):1096–1113, 2002.

74. P. E. Caines. *Linear Stochastic Systems*. John Wiley & Sons, New York, NY, 1988.

75. C. Castaing, P. R. De Fitte, and M. Valadier. *Young measures on topological spaces: with applications in control theory and probability theory*, volume 571. Springer Science & Business Media, 2004.

76. C.D.Charalambous and N.U. Ahmed. Dynamic team theory of stochastic differential decision systems with decentralized noisy information structures via girsanov's measure transformation. *arXiv*, abs/1309.1913, 2013.

77. P. Cembranos and J. Mendoza. *Banach Spaces of Vector-Valued Functions*. Springer-Verlag, 1997.

78. S. Chandak, V.S. Borkar, and P. Dodhia. Reinforcement learning in non-markovian environments. *Systems & Control Letters*, 185:105751, 2024.

79. H. S. Chang and S. I. Marcus. Approximate receding horizon approach for markov decision processes: average reward case. *Journal of Mathematical Analysis and Applications*, 286(2):636–651, 2003.

80. C. D. Charalambous. Decentralized optimality conditions of stochastic differential decision problems via Girsanov's measure transformation. *Mathematics of Control, Signals, and Systems*, 28(3):1–55, 2016.

81. C. D. Charalambous and N. U. Ahmed. Equivalence of decentralized stochastic dynamic decision systems via Girsanov's measure transformation. In *IEEE Conference on Decision and Control (CDC)*, pages 439–444. IEEE, 2014.

82. C. D. Charalambous and N. U. Ahmed. Maximum principle for decentralized stochastic differential decision systems. In *IEEE Conference on Decision and Control (CDC)*, pages 1846–1851. IEEE, 2014.

83. C. T. Chen. *Linear Systems Theory and Design*. Oxford University Press, Oxford, 1999.

84. T.S. Chiang, C.R. Hwang, and S.J. Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

85. P. Chigansky and R. Liptser. On a role of predictor in the filtering stability. *Electronic Communications in Probability*, 11:129–140, 2006.

86. P. Chigansky, R. Liptser, and R. van Handel. Intrinsic methods in filter stability. *Handbook of Nonlinear Filtering*, 2009.

87. C. Y. Chong and M. Athans. On the periodic coordination of linear stochastic systems. *Automatica*, 12:321–335, 1976.

88. C.S. Chow and J. N. Tsitsiklis. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE transactions on automatic control*, 36(8):898–914, 1991.

89. S. B. Connor and G. Fort. State-dependent Foster-Lyapunov criteria for subgeometric convergence of Markov chains. *Stoch. Process Appl.*, 119:176–4193, 2009.

90. O. Costa and F. Dufour. A sufficient condition for the existence of an invariant probability measure for markov processes. *Journal of Applied probability*, 42(3):873–878, 2005.

91. O. Costa and F. Dufour. Average control of Markov decision processes with Feller transition probabilities and general action spaces. *Journal of Mathematical Analysis and Applications*, 396(1):58–69, 2012.

92. L. Cregg, T. Linder, and S. Yüksel. Reinforcement learning for near-optimal design of zero-delay codes for markov sources. *IEEE Transactions on Information Theory, arXiv:2311.12609*, 2024.

93. D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746, 2002.

94. I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

95. M. H. A Davis and P. Varaiya. Information states for linear stochastic systems. *Journal of Mathematical Analysis and Applications*, 37(2):384–402, 1972.

96. M. H. A Davis and P. Varaiya. Dynamic programming conditions for partially observable stochastic systems. *SIAM Journal on Control*, 11(2):226–261, 1973.

97. Eugenio Della Vecchia, Silvia Di Marco, and Alain Jean-Marie. Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost mdps. *Annals of Operations Research*, 199(1):193–214, 2012.

98. A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38. Springer, 2010.

99. Y.E. Demirci, A. D. Kara, and S. Yüksel. Wasserstein regularity of nonlinear filters as belief-mdps, and implications on ergodicity, optimality and learning for pomdps. In *2025 American Control Conference (ACC)*. IEEE, 2025.

100. Y.E. Demirci, A.D. Kara, and S. Yüksel. Average cost optimality of partially observed mdps: Contraction of non-linear filters and existence of optimal solutions. *SIAM Journal on Control and Optimization*, 62:2859–2883, 2004.

101. Y.E. Demirci, A.D. Kara, and S. Yüksel. Refined bounds on near optimality finite window policies in pomdps and their reinforcement learning. *arXiv*, 2024.

102. C. Derman. *Finite state Markovian decision processes*. Academic Press, Inc., 1970.

103. M.S. Derpich and S. Yüksel. Dual effect, certainty equivalence, and separation revisited: A counterexample and a relaxed characterization for optimality. *IEEE Transactions on Automatic Control*, 68(2):1259–1266, 2023.

104. L. Devroye and L. Györfi. *Non-parametric Density Estimation: The L_1 View*. John Wiley, New York, 1985.

105. J. Dieudonné. Sur le théorème de lebesgue-nikodym (v). *Canadian Journal of Mathematics*, 3:129–139, 1951.

106. R.L. Dobrushin. Central limit theorem for nonstationary Markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

107. S. Dong, B. van Roy, and Z. Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *The Journal of Machine Learning Research*, 23(1):11627–11680, 2022.

108. R. Douc, G. Fort, E. Moulines, and P. Soulier. Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Probab*, 14:1353–1377, 2004.

109. R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer, 2018.

110. L. Dubins and D. Freedman. Measurable sets of measures. *Pacific J. Math.*, 14:1211–1222, 1964.

111. R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 2nd edition, 2002.

112. F. Dufour and T. Prieto-Rumeau. Approximation of Markov decision processes with general state space. *J. Math. Anal. Appl.*, 388:1254–1267, 2012.

113. F. Dufour and T. Prieto-Rumeau. Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities. *Stochastics*, pages 1–35, 2014.

114. J. Dugundji. An extension of Tietze's theorem. *Pacific Journal of Mathematics*, 1(3):353–367, 1951.

115. P. G. Dupuis, M. R. James, and I. Petersen. Robust properties of risk-sensitive control. *Mathematics of Control, Signals and Systems*, 13(4):318–332, 2000.

116. R. Durrett. *Probability: Theory and Examples*, volume 3. Cambridge university press, 2010.

117. E. B. Dynkin and A. A. Yushkevich. *Controlled Markov Processes*, volume 235. Springer, 1979.

118. E. Erdoğan and G. N. Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2005.

119. P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, pages 1–52, 2017.

120. S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*, volume 282. John Wiley & Sons, 2009.

121. E. Even-Dar and Y. Mansour. Learning rates for q-learning. *Journal of Machine Learning Research*, 5:1–25, 2004.

122. H.O. Fattorini. Existence theory and the maximum principle for relaxed infinite-dimensional optimal control problems. *SIAM Journal on Control and Optimization*, 32(2):311–331, 1994.

123. E. A. Feinberg. Non-randomized Markov and semi-Markov strategies in dynamic programming. *Theory of Probability & Its Applications*, pages 116–126, 1982.

124. E. A. Feinberg. On measurability and representation of strategic measures in Markov decision processes. *Institute of Mathematical Statistics Lecture Notes. Eds. T. S. Ferguson, L. S. Shapley, J. B. MacQueen*, pages 29–43, 1996.

125. E. A. Feinberg, P. O. Kasyanov, and N. V. Zadoianchuk. Berge's theorem for noncompact image sets. *J. Math. Anal. Appl.*, 397(1):255–259, 2013.

126. E.A. Feinberg and P.O. Kasyanov. Mdps with setwise continuous transition probabilities. *Operations Research Letters*, 49(5):734–740, 2021.

127. E.A. Feinberg and P.O. Kasyanov. Equivalent conditions for weak continuity of nonlinear filters. *Systems & Control Letters*, 173:105458, 2023.

128. E.A. Feinberg, P.O. Kasyanov, and M.Z. Zgurovsky. Partially observable total-cost Markov decision process with weakly continuous transition probabilities. *Mathematics of Operations Research*, 41(2):656–681, 2016.

129. E.A. Feinberg, P.O. Kasyanov, and M.Z. Zgurovsky. Markov decision processes with incomplete information and semiuniform feller transition probabilities. *SIAM Journal on Control and Optimization*, 60(4):2488–2513, 2022.

130. W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, New York, 1971.

131. W. H. Fleming and H. M. Soner. *Controlled Markov processes and viscosity solutions*, volume 25. Springer Science & Business Media, 2006.

132. W.H. Fleming and E. Pardoux. Optimal control for partially observed diffusions. *SIAM J. Control Optim.*, 20(2):261–285, 1982.

133. S. R. Foguel. Positive operators on $c(x)$. *Proceedings of the American Mathematical Society*, pages 295–297, 1969.

134. G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley and Sons, 1999.

135. F. Forges. An approach to communication equilibria. *Econometrica: Journal of the Econometric Society*, pages 1375–1385, 1986.

136. P. K. Friz and M. Hairer. *A course on rough paths*. Springer, 2020.

137. Peter K. Friz and Martin Hairer. *A course on rough paths*. Universitext. Springer, Cham, [2020] ©2020. With an introduction to regularity structures, Second edition of [3289027].

138. C. Gaskett and A. Zelinsky D. Wettergreen. Q-learning in continuous state and action spaces. In *Australasian joint conference on artificial intelligence*, pages 417–428. Springer, 1999.

139. A. Gattami, B. M. Bernhardsson, and A. Rantzer. Robust team decision theory. *IEEE Transactions on Automatic Control*, 57:794–798, March 2012.

140. J. Geanakoplos and H. M. Polemarchakis. We can't disagree forever. *J. Economic Theory*, pages 192–200, 1982.

141. T. T. Georgiou and A. Lindquist. The separation principle in stochastic control, redux. *IEEE Transactions on Automatic Control*, 58(10):2481–2494, 2013.

142. A. Gersho. Stochastic stability of delta modulation. *Bell Syst. Tech. J*, 51(4):821–841, 1972.

143. J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer, New York, 2003.

144. I. I. Gihman and A. V. Skorohod. *Controlled Stochastic Processes*. Springer Science & Business Media, 2012.

145. I. V. Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability & Its Applications*, 5(3):285–301, 1960.

146. F. Le Gland and N. Oudjane. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *The Annals of Applied Probability*, 14(1):144–187, 2004.

147. E. Gordienko and O. Hernández-Lerma. Average cost Markov control processes with weighted norms: Existence of canonical policies. *Appl. Math.*, 23(2):199–218, 1995.

148. E. Gordienko, E. Lemus-Rodríguez, and R. Montes de Oca. Discounted cost optimality problem: stability with respect to weak metrics. *Mathematical Methods of Operations Research*, 68(1):77–96, 2008.

149. E. Gordienko, E. Lemus-Rodríguez, and R. Montes de Oca. Average cost markov control processes: stability with respect to the kantorovich metric. *Mathematical Methods of Operations Research*, 70:13–33, 2009.

150. A. Gosavi. Reinforcement learning for long-run average cost. *European journal of operational research*, 155(3):654–674, 2004.

151. G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford university press, 2020.

152. A. Gupta, S. Yüksel, T. Başar, and C. Langbort. On the existence of optimal policies for a class of static and sequential dynamic teams. *SIAM Journal on Control and Optimization*, 53:1681–1712, 2015.

153. M. Hairer. Ergodic properties of Markov processes. *Lecture Notes, University of Warwick*, 2006.

154. M. Hairer. Convergence of markov processes. Lecture Notes, 2010.

155. M. Hairer. On malliavins proof of hörmanders theorem. *Bulletin des sciences mathematiques*, 135(6-7):650–666, 2011.

156. B. Hajek. Optimal control of two interacting service stations. *IEEE transactions on automatic control*, 29(6):491–499, 1984.

157. B. Hajek. Lecture notes: Communication network analysis. *University of Illinois at Urbana-Champaign*, 2006.

158. B. Hajek. *Random Processes for Engineers*. Cambridge University Press, 2015.

159. R. Van Handel. Discrete time nonlinear filters with informative observations are stable. *Electron. Commun. Probab*, 13:562–575, 2008.

160. O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer-Verlag, 1989.

161. O. Hernández-Lerma. Existence of average optimal policies in markov control processes with strictly unbounded costs. *Kybernetika*, 29(1):1–17, 1993.

162. O. Hernández-Lerma. *Adaptive Markov control processes*, volume 79. Springer Science & Business Media, 2012.

163. O. Hernández-Lerma, R. Montes de Oca, and R. Cavazos-Cadena. Recurrence conditions for markov decision processes with borel state space: a survey. *Annals of Operations Research*, 28(1):29–46, 1991.

164. O. Hernández-Lerma and J. B. Lasserre. Error bounds for rolling horizon policies in discrete-time markov control processes. *IEEE Transactions on Automatic Control*, 35(10):1118–1124, 1990.

165. O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

166. O. Hernández-Lerma and J. B. Lasserre. *Further topics on discrete-time Markov control processes*. Springer, 1999.

167. O. Hernández-Lerma and J. B. Lasserre. *Markov Chains and Invariant Probabilities*. Birkhäuser, Basel, 2003.

168. O. Hernández-Lerma and J. B. Lasserre. *Markov Chains and Invariant Probabilities*. Birkhäuser, Basel, 2003.

169. C. J. Himmelberg, T. Parthasarathy, and F. S. Van Vleck. Optimal plans for dynamic programming problems. *Mathematics of Operations Research*, 1(4):390–394, 1976.

170. K. Hinderer. Lipschitz continuity of value functions in markovian decision processes. *Mathematical Methods of Operations Research*, 62:3–22, 2005.

171. Y. C. Ho and K. C. Chu. Team decision theory and information structures in optimal control problems - part I. *IEEE Transactions on Automatic Control*, 17:15–22, February 1972.

172. Y. C. Ho and K. C. Chu. On the equivalence of information structures in static and dynamic teams. *IEEE Transactions on Automatic Control*, 18(2):187–188, 1973.

173. I. Hogeboom-Burr and S. Yüksel. Sequential stochastic control (single or multi-agent) problems nearly admit change of measures with independent measurements. *Applied Mathematics and Optimization*, 2023.

174. J. Hunter and B. Nachtergaele. *Applied Analysis*. World Scientific, Singapore, 2005.

175. N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014.

176. O. C. Imer, S. Yüksel, and T. Başar. Optimal control of LTI systems over unreliable communication links. *Automatica*, 42(9):1429–1440, 2006.

177. G.N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

178. T. Jaakkola, M.I. Jordan, and S.P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201, 1994.

179. D. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic control*, 18(2):124–131, 1973.

180. S. Jafarpour and A.D Lewis. Locally convex topologies and control theory. *Mathematics of Control, Signals, and Systems*, 28(4):29, 2016.

181. O. Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, New York, 1997.

182. A. D. Kara, M. Raginsky, and S. Yüksel. Robustness to incorrect models and data-driven learning in average-cost optimal stochastic control. *Automatica*, 139:110179, 2022.

183. A. D. Kara and S. Yüksel. Robustness to approximations and model learning in MDPs and POMDPs. In A. B. Piunovskiy and Y. Zhang, editors, *Modern Trends in Controlled Stochastic Processes: Theory and Applications, Volume III*. Luniver Press, 2021.

184. A.D Kara, N. Saldi, and S. Yüksel. Weak Feller property of non-linear filters. *Systems & Control Letters*, 134:104–512, 2019.

185. A.D Kara, N. Saldi, and S. Yüksel. Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity. *Journal of Machine Learning Research*, pages 1–34, 2023.

186. A.D Kara and S. Yüksel. Robustness to incorrect priors in partially observed stochastic control. *SIAM Journal on Control and Optimization*, 57(3):1929–1964, 2019.

187. A.D Kara and S. Yüksel. Robustness to incorrect system models in stochastic control. *SIAM Journal on Control and Optimization*, 58(2):1144–1182, 2020.

188. A.D Kara and S. Yüksel. Near optimality of finite memory feedback policies in partially observed markov decision processes. *Journal of Machine Learning Research*, 23(11):1–46, 2022.

189. A.D Kara and S. Yüksel. Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability. *Mathematics of Operations Research*, 48(4):2066–2093, 2023.

190. A.D. Kara and S. Yüksel. Q-learning for continuous state and action mdps under average cost criteria. *arXiv preprint arXiv:2308.07591*, 2023.

191. A.D. Kara and S. Yüksel. Q-learning for stochastic control under general information structures and non-markovian environments. *Transactions on Machine Learning Research*, 2024. Featured Certification.

192. A.D Kara and S. Yüksel. Robustness to incorrect system models in stochastic control and application to data-driven learning. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2753–2758, Dec 2018.

193. R. Khasminskii. *Stochastic stability of differential equations*. Springer, 2011.

194. W. Kliemann. Recurrence and invariant measures for degenerate diffusions. *The Annals of Probability*, 15(2):690–707, 1987.

195. T. Komorowski, S. Peszat, and T. Szarek. On ergodicity of some markov processes. 2010.

196. J.C. Krainak, J.L. Speyer, and S.I. Marcus. Static team problems – part I: Sufficient conditions and the exponential cost criterion. *IEEE Transactions on Automatic Control*, 27:839–848, April 1982.

197. V. Krishnamurthy. *Partially Observed Markov Decision Processes: Filtering, Learning and Controlled Sensing*. Cambridge University Press, 2 edition, 2025.

198. N.V. Krylov. *Controlled Diffusion Processes*, volume 14. Springer, 2008.

199. H. Kuhn. Extensive games and the problem of information. *In Contributions to the Theory of Games, (H. Kuhn and A. Tucker, editors)*, pages 193–216, 1953.

200. P. R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

201. M. Kurano. The existence of a minimum pair of state and policy for markov decision processes under the hypothesis of doeblin. *SIAM journal on control and optimization*, 27(2):296–307, 1989.

202. K. Kuratowski and C. Ryll-Nardzewski. A general theorem on selectors. *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys*, 13(1):397–403, 1965.

203. H. J. Kushner. Numerical methods for stochastic control problems in continuous time. *SIAM Journal on Control and Optimization*, 28(5):999–1048, 1990.

204. H. J. Kushner. A partial history of the early development of continuous-time nonlinear stochastic systems theory. *Automatica*, 50(2):303–334, 2014.

205. H. J. Kushner and P. G. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*, volume 24. Springer Science & Business Media, 2001.

206. H.J. Kushner. *Stochastic stability and control*. Academic Press, New York, 1967.

207. H.J. Kushner. *Introduction to Stochastic Control Theory*. Holt, Rinehart and Winston, New York, 1972.

208. H.J. Kushner. *Weak convergence methods and singularly perturbed stochastic control and filtering problems*. Springer Science & Business Media, 2012.

209. H.J. Kushner and G. Yin. Stochastic approximation and recursive algorithms and applications, 2003.

210. D. Lacker. Probabilistic compactification methods for stochastic optimal control and mean field games. 2018.

211. H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

212. H.J. Langen. Convergence of dynamic programming models. *Mathematics of Operations Research*, 6(4):493–512, Nov. 1981.

213. J. B. Lasserre. Invariant probabilities for Markov chains on a metric space. *Statistics and Probability Letters*, 34:259–265, 1997.

214. J. B. Lasserre. Sample-path average optimality for markov control processes. *IEEE Transactions on Automatic Control*, 44(10):1966–1971, 1999.

215. H. Lee and Y. Lim. Invariant metrics, contractions and nonlinear matrix equations. *Nonlinearity*, 21(4):857, 2008.

216. E. Lehrer, D. Rosenberg, and E. Shmaya. Signaling and mediation in games with common interests. *Games and Economic Behavior*, 68:670–682, 2010.

217. B. C. Levy and M. Zorzi. A contraction analysis of the convergence of risk-sensitive filters. *SIAM Journal on Control and Optimization*, 54(4):2154–2173, 2016.

218. D. Liberzon. *Calculus of variations and optimal control theory: a concise introduction*. Princeton university press, 2011.

219. A. Lindquist. On feedback control of linear stochastic systems. *SIAM Journal on Control*, 11(2):323–343, 1973.

220. D.G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, NY, 1969.

221. T. J. Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998.

222. A. S. Manne. Linear programming and sequential decision. *Management Science*, 6:259–267, April 1960.

223. X. Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

224. E. Mascolo and L. Migliaccio. Relaxation methods in control theory. *Applied Mathematics and Optimization*, 20(1):97–103, 1989.

225. C. McDonald and S. Yüksel. Exponential filter stability via Dobrushin's coefficient. *Electronic Communications in Probability*, 25, 2020.

226. C. McDonald and S. Yüksel. Robustness to incorrect priors and controlled filter stability in partially observed stochastic control. *SIAM Journal on Control and Optimization*, 60(2):842–870, 2022.

227. C. McDonald and S. Yüksel. Stochastic observability and filter stability under several criteria. *IEEE Transactions on Automatic Control*, 69(5):2931–2946, 2024.

228. E. J. McShane. Relaxed controls and variational problems. *SIAM Journal on Control*, 5(3):438–485, 1967.

229. F. C. Melo, S. P. Meyn, and I. M. Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671, 2008.

230. J.-F. Mertens, S. Sorin, and S. Zamir. *Repeated games*, volume 55. Cambridge University Press, 2015.

231. S. Meyn. *Control systems and reinforcement learning*. Cambridge University Press, 2022.

232. S. P. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, 2007.

233. S. P. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

234. S. P. Meyn and R. Tweedie. State-dependent criteria for convergence of Markov chains. *Ann. Appl. Prob*, 4:149–168, 1994.

235. S. P. Meyn and R. L. Tweedie. Stability of Markovian processes ii: Continuous-time processes and sampled chains. *Advances in Applied Probability*, 25(3):487–517, 1993.

236. S. P. Meyn and R. L. Tweedie. Stability of Markovian processes iii: Foster-lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, pages 518–548, 1993.

237. S.P. Meyn and R.L. Tweedie. State-dependent criteria for convergence of markov chains. *Annals Appl. Prob.*, 4(1):149–168, 1994.

238. P. R. Milgrom and R. J. Weber. Distributional strategies for games with incomplete information. *Mathematics of operations research*, 10(4):619–632, 1985.

239. S. K. Mitter. Nonlinear filtering of diffusion processes a guided tour. In *Advances in Filtering and Optimal Stochastic Control*, pages 256–266. Springer, 1982.

240. A. Müller. How does the value function of a markov decision process depend on the transition probabilities? *Mathematics of Operations Research*, 22(4):872–885, 1997.

241. A. Nayyar, A. Mahajan, and D. Teneketzis. Optimal control strategies in delayed sharing information structures. *IEEE Transactions on Automatic Control*, 56:1606–1620, 2011.

242. A. Nayyar, A. Mahajan, and D. Teneketzis. The common-information approach to decentralized stochastic control. In Information and Control in Networks*, Editors: G. Como, B. Bernhardsson, A. Rantzer*. Springer, 2013.

243. J. Neveu. Discrete-parameter martingales. revised edition, 1975.

244. L. Nielsen. Common knowledge, communication and convergence of beliefs. *Mathematical Social Sciences*, 8:1–14, 1984.

245. A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

246. E. Nummelin. A splitting technique for harris recurrent markov chains. *Z. Wahrscheinlichkeitstheoric verw. Gebiete*, 43:309–318, 1978.

247. E. Nummelin. A splitting technique for harris recurrent markov chains. *Z. Wahrscheinlichkeitstheoric verw. Gebiete*, 43:309–318, 1978.

248. B. Øksendal. *Stochastic Differential Equations*. Springer, Berlin, 2003.

249. K.R. Parthasarathy. *Probability Measures on Metric Spaces*. AMS Bookstore, 1967.

250. I. Petersen, M. R. James, and P. Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 45(3):398–412, 2000.

251. R.P. Phelps. *Lectures on Choquet's theorem*. Van Nostrand, New York:, 1966.

252. J. Pitman and M. Yor. A guide to brownian motion and related stochastic processes. *arXiv preprint arXiv:1802.09679*, 2018.

253. A. B. Piunovskiy. Controlled random sequences: methods of convex analysis and problems with functional constraints. *Russian Mathematical Surveys*, 53(6):1233–1293, 1998.

254. P. Dai Pra, L. Meneghini, and W. J. Runggaldier. Connections between stochastic control and dynamic games. *Mathematics of Control, Signals and Systems*, 9(4):303–326, 1996.

255. S. Pradhan, Z. Selk, and S. Yüksel. Robustness of optimal controlled diffusions with near-brownian noise via rough paths theory. *arXiv:2310.09967*, 2023.

256. S. Pradhan and S. Yüksel. Continuity of cost in Borkar control topology and implications on discrete space and time approximations for controlled diffusions under several criteria. *Electronic Journal of Probability*, 29:1–32, 2024.

257. S. Pradhan and S. Yüksel. Near optimality of discrete-time approximations for controlled mckean-vlasov diffusions and interacting particle systems. *arXiv preprint arXiv:2510.21208*, 2025.

258. Somnath Pradhan and Serdar Yüksel. Controlled diffusions under full, partial, and decentralized information: Existence of optimal policies and discrete-time approximations. *SIAM Journal on Control and Optimization*, 63(5):3674–3702, 2025.

259. G. Da Prato and J. Zabczyk. *Ergodicity for infinite dimensional systems*, volume 229. Cambridge University Press, 1996.

260. R. Radner. Team decision problems. *Ann. Math. Statist.*, 33:857–881, 1962.

261. R. Radner. Team decision problems. *Annals of Mathematical Statistics*, 33:857–881, 1962.

262. D. Rhenius. Incomplete information in Markovian decision models. *Ann. Statist.*, 2:1327–1334, 1974.

263. H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

264. G.O. Roberts and J.S. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Survery*, 1:20–71, 2004.

265. K. W. Ross. Randomized and past-dependent policies for Markov decision processes with multiple constraints. *Operations Research*, 37:474–477, May 1989.

266. S. M. Ross. On the nonexistence of ϵ-optimal randomized stationary policies in average cost markov decision models. *The Annals of Mathematical Statistics*, 42(5):1767–1768, 1971.

267. E. P. Ryan. On brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback. *SIAM Journal on Control and Optimization*, 32(6):1597–1604, 1994.

268. N. Saldi. A topology for team policies and existence of optimal team policies in stochastic team theory. *IEEE Transactions on Automatic Control*, 65(1):310–317, 2020.

269. N. Saldi, T. Linder, and S. Yüksel. Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control. *IEEE Trans. Automatic Control*, 60:553 –558, 2015.

270. N. Saldi, T. Linder, and S. Yüksel. *Finite Approximations in Discrete-Time Stochastic Control: Quantized Models and Asymptotic Optimality*. Springer, Cham, 2018.

271. N. Saldi and S. Yüksel. Geometry of information structures, strategic measures and associated control topologies. *Probability Surveys*, 19:450–532, 2022.

272. N. Saldi, S. Yüksel, and T. Linder. Finite-state approximation of Markov decision processes with unbounded costs and Borel spaces. In *IEEE Conf. Decision Control*, Osaka. Japan, December 2015.

273. N. Saldi, S. Yüksel, and T. Linder. Near optimality of quantized policies in stochastic control under weak continuity conditions. *Journal of Mathematical Analysis and Applications*, 435(1):321–337, 2016.

274. N. Saldi, S. Yüksel, and T. Linder. Finite model approximations and asymptotic optimality of quantized policies in decentralized stochastic control. *IEEE Transactions on Automatic Control*, 62:2360 – 2373, 2017.

275. N. Saldi, S. Yüksel, and T. Linder. On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. *Mathematics of Operations Research*, 42(4):945–978, 2017.

276. N. Saldi, S. Yüksel, and T. Linder. Finite model approximations for partially observed Markov decision processes with discounted cost. *IEEE Transactions on Automatic Control*, 65, 2020.

277. S. Sanjari, T. Başar, and S. Yüksel. Isomorphism properties of optimality and equilibrium solutions under equivalent information structure transformations: Stochastic dynamic games and teams. *SIAM Journal on Control and Optimization*, 61(5):3102–3130, 2023.

278. S. Sanjari, N. Saldi, and S. Yüksel. Optimality of independently randomized symmetric policies for exchangeable stochastic teams with infinitely many decision makers. *Mathematics of Operations Research*, 48(3):1254–1285, 2023.

279. S. Sanjari and S. Yüksel. Optimal solutions to infinite-player stochastic teams and mean-field teams. *IEEE Transactions on Automatic Control*, 66(3):1071–1086, 2020.

280. S. Sanjari and S. Yüksel. Optimal policies for convex symmetric stochastic dynamic teams and their mean-field limit. *SIAM Journal on Control and Optimization*, 59(2):777–804, 2021.

281. A. V. Savkin and I. R. Petersen. Robust control of uncertain systems with structured uncertainty. *Journal of Mathematical Systems, Estimation, and Control*, 6(3):1–14, 1996.

282. M. Schäl. A selection theorem for optimization problems. *Archiv der Mathematik*, 25(1):219–224, 1974.

283. M. Schäl. Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrscheinlichkeitsth*, 32:179–296, 1975.

284. M. Schäl. On dynamic programming: compactness of the space of policies. *Stochastic Processes and their Applications*, 3(4):345–364, 1975.

285. M. Schäl. Average optimality in dynamic programming with general state space. *Mathematics of operations Research*, 18(1):163–172, 1993.

286. L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry. Foundations of control and estimation over lossy networks. *Proceedings of the IEEE*, 95(1):163–187, 2007.

287. R. Serfozo. Convergence of Lebesgue integrals with varying measures. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 380–402, 1982.

288. S. Shreve and D. P. Bertsekas. Universally measurable policies in dynamic programming. *Mathematics of Operations Research*, 4(1):15–30, 1979.

289. S. P. Singh, T. Jaakkola, and M. I. Jordan. Reinforcement learning with soft state aggregation. *Advances in neural information processing systems*, pages 361–368, 1995.

290. S.P. Singh, T. Jaakkola, and M.I. Jordan. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*, pages 284–292. Elsevier, 1994.

291. E. D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, volume 6. Springer Science & Business Media, 2013.

292. R.B Sowers and A.M. Makowski. Discrete-time filtering for linear systems with non-Gaussian initial conditions: asymptotic behavior of the difference between the MMSE and LMSE estimates. *IEEE Transactions on Automatic Control*, 37(1):114–120, 1992.

293. S. M. Srivastava. *A course on Borel sets*, volume 180. Springer Science & Business Media, 2008.

294. L. Stettner. On the existence and uniqueness of invariant measure for continuous time markov processes. Technical report, BROWN UNIV PROVIDENCE RI LEFSCHETZ CENTER FOR DYNAMICAL SYSTEMS, 1986.

295. Richard H Stockbridge. Time-average control of martingale problems: Existence of a stationary solution. *Annals of Probability*, pages 190–205, 1990.

296. C. Striebel. Sufficient statistics in the optimum control of stochastic systems. *J. Math. Anal. Appl*, 12:576–592, 1965.

297. D. Stroock and S. R. S. Varadhan. On the support of diffusion processes with applications to the strong maximum principle. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971)*, volume 3, pages 333–359, 1972.

298. H. Sun and H. Xu. Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*, 41:377–401, 07 2015.

299. C. Szepesvári. The asymptotic convergence-rate of q-learning. *Advances in neural information processing systems*, 10, 1997.

300. C. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan and Claypool, 2010.

301. C. Szepesvári. Algorithms for reinforcement learning. volume 4, pages 1–103, 2010.

302. C. Szepesvàri and W.D. Smart. Interpolation-based q-learning. 2004.

303. C. Szepesvári and M.L. Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural computation*, 11(8):2017–2060, 1999.

304. D. Teneketzis. On information structures and nonsequential stochastic control. *CWI Quarterly*, 9:241–260, 1996.

305. D. Teneketzis. On the structure of optimal real-time encoders and decoders in noisy communication. *IEEE Transactions on Information Theory*, 52:4017–4035, September 2006.

306. J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

307. J. N. Tsitsiklis and M. Athans. Convergence and asymptotic agreement in distributed decision problems. *IEEE Transactions on Automatic Control*, pages 42–50, January 1984.

308. P. Tuominen and R.L. Tweedie. Subgeometric rates of convergence of f-ergodic markov chains. *Adv. Annals Appl. Prob.*, 26(3):775–798, September 1994.

309. R. L. Tweedie. Topological conditions enabling use of Harris methods in discrete and continuous time. *Acta Appl. Math.*, 34(1-2):175–188, 1994.

310. R. L. Tweedie. Drift conditions and invariant measures for Markov chains. *Stochastic Processes and Their Applications*, 92:345–354, 2001.

311. V. A. Ugrinovskii. Robust H-infinity control in the presence of stochastic uncertainty. *International Journal of Control*, 71(2):219–237, 1998.

312. R. van Handel. Stochastic calculus, filtering, and stochastic control. *Course notes., URL http://www. princeton. edu/˜ rvan/acm217/ACM217. pdf*, 2007.

313. R. van Handel. Observability and nonlinear filtering. *Probability theory and related fields*, 145(1-2):35–74, 2009.

314. R. van Handel. The stability of conditional Markov processes and Markov chains in random environments. *Ann. Probab*, 37:1876–1925, 2009.

315. R. van Handel. Nonlinear filtering and systems theory. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS semi-plenary paper)*, 2010.

316. S.R.S. Varadhan. *Probability theory, volume 7 of Courant Lecture Notes in Mathematics*, volume 1. New York University Courant Institute of Mathematical Sciences, New York.

317. O. Vega-Amaya. The average cost optimality equation: a fixed point approach. *Bol. Soc. Mat. Mexicana*, 9(3):185–195, 2003.

318. Oscar Vega-Amaya. The average cost optimality equation: a fixed point approach. *Bol. Soc. Mat. Mexicana*, 9(1):185–195, 2003.

319. M. Vidyasagar. Convergence of stochastic approximation via martingale and converse lyapunov methods. *Mathematics of Control, Signals, and Systems*, pages 1–24, 2023.

320. C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

321. J. C. Walrand and P. Varaiya. Optimal causal coding-decoding problems. *IEEE Transactions on Information Theory*, 19:814–820, November 1983.

322. J. Warga. *Optimal Control of Differential and Functional Equations*. Academic press, 2014.

323. C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

324. G. L. Wise. A note on a common misconception in estimation. *Systems and Cont. Letters*, 5:355–356, April 1985.

325. H. S. Witsenhausen. A counterexample in stochastic optimal control. *SIAM J. Contr.*, 6:131–147, 1968.

326. H. S. Witsenhausen. On information structures, feedback and causality. *SIAM J. Control*, 9:149–160, May 1971.

327. H. S. Witsenhausen. A standard form for sequential stochastic control. *Mathematical Systems Theory*, 7:5–11, 1973.

328. H. S. Witsenhausen. The intrinsic model for discrete stochastic control: Some open problems. *Lecture Notes in Econ. and Math. Syst., Springer-Verlag*, 107:322–335, 1975.

329. H. S. Witsenhausen. The intrinsic model for discrete stochastic control: Some open problems. In *Control Theory, Numerical Methods and Computer System Modelling*, pages 322–335, A. Bensoussan and J. L. Lions Springer-Verlag 107, 1975. Lecture Notes in Economics and Mathematical Systems.

330. H. S. Witsenhausen. On the structure of real-time source coders. *Bell Syst. Tech. J*, 58:1437–1451, July/August 1979.

331. H. S. Witsenhausen. Equivalent stochastic control problems. *Math. Control, Signals and Systems*, 1:3–11, 1988.

332. E. Wong and B.E. Hajek. *Stochastic Processes in Engineering Systems*. Springer-Verlag, New York, 1985.

333. E. Wong and M. Zakai. On the convergence of ordinary integrals to stochastic integrals. *The Annals of Mathematical Statistics*, 36(5):1560–1564, 1965.

334. W. M. Wonham. On the separation theorem of stochastic control. *SIAM Journal on Control*, 6(2):312–326, 1968.

335. R.G. Wood, T. Linder, and S. Yüksel. Optimal zero delay coding of Markov sources: Stationary and finite memory codes. *IEEE Transactions on Information Theory*, 63:5968–5980, 2017.

336. D. T. H. Worm and S. C. Hille. Ergodic decompositions associated with regular markov operators on polish spaces. *Ergodic Theory and Dynamical Systems*, 31(2):571 – 597, 2010.

337. H. Xu and S. Mannor. Distributionally robust Markov decision processes. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2505–2513. Curran Associates, Inc., 2010.

338. T. Yoshikawa. Dynamic programming approach to decentralized control problems. *IEEE Transactions on Automatic Control*, 20:796–797, 1975.

339. L. C. Young. An inequality of the hölder type, connected with stieltjes integration. *Acta Mathematica*, 67:251–282, 1936.

340. L.C. Young. Generalized curves and the existence of an attained absolute minimum in the calculus of variations. *Comptes Rendus de la Societe des Sci. et des Lettres de Varsovie*, 30:212–234, 1937.

341. H. Yu. Some proof details for asynchronous stochastic approximation algorithms. 2012.

342. S. Yüksel. Stochastic nestedness and the belief sharing information pattern. *IEEE Transactions on Automatic Control*, 54:2773–2786, December 2009.

343. S. Yüksel. On optimal causal coding of partially observed Markov sources in single and multi-terminal settings. *IEEE Transactions on Information Theory*, 59:424–437, January 2013.

344. S. Yüksel. On stochastic stability of a class of non-Markovian processes and applications in quantization. *SIAM J. on Control and Optimization*, 55:1241–1260, 2017.

345. S. Yüksel. A note on the separation of optimal quantization and control policies in networked control. *SIAM Journal on Control and Optimization*, 57(1):773–782, 2019.

346. S. Yüksel. A universal dynamic program and refined existence results for decentralized stochastic control. *SIAM Journal on Control and Optimization*, 58(5):2711–2739, 2020.

347. S. Yüksel. On Borkar and Young relaxed control topologies and continuous dependence of invariant measures on control policy. *SIAM Journal on Control and Optimization*, 62(4):2367–2386, 2024.

348. S. Yüksel. Another look at partially observed optimal stochastic control: Existence, ergodicity, and approximations without belief-reduction. *Applied Mathematics & Optimization*, 91(1):16, 2025.

349. S. Yüksel and T. Başar. *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. Springer, New York, 2013.

350. S. Yüksel and S. P. Meyn. Random-time, state-dependent stochastic drift for Markov chains and application to stochastic stabilization over erasure channels. *IEEE Transactions on Automatic Control*, 58:47 – 59, January 2013.

351. S. Yüksel and N. Saldi. Convex analysis in decentralized stochastic control, strategic measures and optimal solutions. *SIAM J. on Control and Optimization*, 55:1–28, 2017.

352. A.A. Yushkevich. Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces. *Theory Prob. Appl.*, 21:153–158, 1976.

353. Yichen Zhou, Yanglei Song, and Serdar Yüksel. Robustness to model approximation, learning, and sample complexity in wasserstein regular mdps. *arXiv preprint arXiv:2410.14116*, 2024.

354. R. Zurkowski, S. Yüksel, and T. Linder. On rates of convergence for Markov chains under random time state-dependent stochastic drift criteria. *IEEE Transactions on Automatic Control*, 61(1):145–155, 2015.