University of Oxford

Mathematical Institute

C8.7 – Optimal Control

(Draft notes, March 26, 2025)

Samuel Cohen

Hilary Term 2025

Contents

Introduction

This version of the lecture notes contains the main results (with proofs) which we will cover in the course, but does not contain many examples – these will be added as we go through the course. The appendix is still being completed, with the addition of material from other courses which may useful depending on background.

Thanks to Lingyi Yang and Wojtek Anyszka for comments and pointing out errors in early versions of these notes.

Notation

We will try and be somewhat consistent with notation throughout this course.

- For the avoidance of doubt, $0 \notin \mathbb{N}$.
- A process (whether random or deterministic, in either discrete or continuous time) will be denoted with a capital letter (say X), and the value it takes at time t will be either X_t or $X(t)$ as convenient. The space it takes values in is the calligraphic $\mathcal{X}$, and a typical value in the set is denoted x.
- The set of times which we are considering in our problem will be $\mathbb{T}$, and may be $\{0, 1, ..., T\}$, $\{0, 1, ...\}$, $[0, T]$ or $[0, \infty)$ as context requires. We will use s and t as time variables, and typically take $s \leq t$.
- The size of a set A (that is, the number of elements it contains), will be written $|A|$ or $\#A$ if there might be confusion.
- The indicator function will be written 1_A, where A is some event or condition (so $1_A = 1$ if A occurs, and $1_A = 0$ otherwise).
- The expectation operator will be written $\mathbb{E}$, and the variance $\mathbb{V}$. These can be augmented with various superscripts, which specify (in some way) how the probabilities are chosen, for notational convenience.
- Partial derivatives will be written using the shorthand $\partial_t = \frac{\partial}{\partial t}$, and when there is a clear spatial variable x we write ∇ or D_x for the column vector with components ∂_{x_i}, so $\nabla v = D_x v$ is the gradient of v. Similarly, we write $D^2_{xx} v$ for the Hessian of v.
- The Euclidean norm will be denoted $\|x\|$, and the ℓ^∞ norm denoted $\|x\|_\infty = \max_i\{|x_i|\}$. (Euclidean) inner products will be denoted either $\langle x, y\rangle$ or $x^\top y$.

- The minimum of two quantities will be written $\min\{x, y\} = x \wedge y$, and the maximum $\max\{x, y\} = x \vee y$.

References

While these lectures are aiming to be self-contained (and the proofs may differ from those which are 'standard'), this is an area with many good books. However, you will find that there is a range of styles, with varying levels of rigour and applicability. A few books (in a roughly increasing level of complexity/rigour) are:

- Sutton and Barto *Reinforcement Learning: An Introduction*, MIT 1998
- Whittle, P. *Optimal Control: Basics and Beyond*, Wiley, 1996
- Bertsekas, Dimitri P. *A Course in Reinforcement Learning (2nd Edition)*, Athena Scientific, 2024
- Bensoussan, *Estimation and Control of Dynamical Systems*, Springer 2018
- Pham, *Continuous-time Stochastic Control and Optimization with Financial Applications*, Springer 2009
- Bertsekas and Shreve, *Stochastic Optimal Control: The Discrete-time case*, Athena Scientific, 1996
- Yong and Zhou *Stochastic Controls: Hamiltonian Systems and HJB equations*, Springer 1999
- Touzi, *Optimal Stochastic Control, Stochastic Target Problems and Backward SDE*, Fields Lecture Notes 2010
- Fleming and Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer 2006
- Krylov, *Controlled Diffusion Processes*, Springer 1980

Chapter 1

Discrete-time control and Markov Decision Processes

In this first part of the course, we will look at discrete time optimal control problems. We will begin by considering deterministic problems, and then introduce some randomness in our system.

1.1 State variables and a first control problem

In optimal control, we wish to make decisions about actions which modify the state of the world. To make a mathematical model of this, we first need to describe what we mean by 'the state of the world', and how this is affected by our actions. We will begin with a simple discrete time, deterministic setting, which avoids technicalities, while showing us some of the basic properties of these problems.

We suppose we have a *state process* $X = \{X_t\}_{t\in\mathbb{T}}$, which describes all (relevant) properties of the world, with $\mathbb{T} = \{0, 1, ..., T\}$. We will assume that X takes values in $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \geq 1$. This process will be affected by a control process, which we denote $\{U_t\}_{t\in\mathbb{T}}$, and takes values in some set $\mathcal{U}$.

We will assume that X can be described through its one-step dynamics, which we write as

$$X_{t+1} = f(t, X_t, U_t),$$

where $f : \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ is a function (which we will assume known, for now). We will make continuity assumptions about f as we go. This is known as the *state dynamics* or *plant equation*.

An agent wishes to optimize their rewards and costs. There are two conventions – in the mathematical control and optimization community, we usually think about minimizing some cost; in the reinforcement learning community, we usually think about maximizing rewards. For the sake of consistency, we will follow the convention of minimizing costs (even for when presenting reinforcement learning algorithms), but the only difference is a change of sign.

We describe the agent's costs by a function $g : \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, which represents the cost which the agent must pay at time t, in state X_t, if they choose control U_t. We seek to find an optimal control, that is, a

control which minimizes

$$J(x,U) = \sum_{s\geq 0} g(s, X_s^U, U_s) \tag{1.1.1}$$

with respect to $U = \{U_s\}_{s\geq 0}$, where $X_0 = x$ is the initial value of X (which is where the system begins) and where X^U is the solution of the plant equation (for $s \geq 0$) with control U.

At this point, before we try and solve this problem, we make the following observations:

- We have said that X should contain all relevant information. What do we mean by relevant? Clearly X should be enough to allow us to determine our costs/rewards at every time (we will define these later), as this allows us to describe our preferences about the world. Furthermore, it's important that X is enough to determine the future dynamics of the world, without needing to know any additional information. In particular, we will assume that the *current* state is enough to build a model of the future – we gain nothing by remembering more information (for example the past values of X). In a stochastic setting, this is closely related to a Markov assumption (but this is complicated by the control, as we will see later). If we want to include more memory, we can expand X to include its past values, at the cost of increasing the dimension of X.

We now have a couple of ways to proceed:

- We could try and find the cheapest U_t for each pair (X_t, X_{t+1}), and so define

  $$c(t, X_t, X_{t+1}) = \min_{U_t} \Big\{ g(t, X_t, U_t) : X_{t+1} = f(t, X_t, U_t) \Big\}.$$

  and $c(T, X_T, X_{T+1}) = \min_{U_T} g(T, X_T, U_T)$. This would convert our problem into minimizing a new functional $(\sum_{t\leq T} c(s, X_t, X_{t+1}))$, which is the problem of calculus of variations. Doing this conversion is not always simple, and it doesn't allow us to include randomness.

- We could consider minimizing J with respect to $\{X_s, U_s\}_{s\geq 0}$, by treating $X_{t+1} = g(t, X_t, U_t)$ as a constraint, which we can handle with Lagrange multipliers. This is a very high dimensional problem though, so can be tricky to solve (but we will return to this approach later!).

- The option we will pursue is to embed our optimization within a family of optimization problems, which exploits the dynamic nature of our problem. This will lend itself to stochastic problems as well.

1.1.1 Building a dynamic programming problem

Instead of just optimizing J in (1.1.1), we will consider the family of problems of minimizing the *remaining-cost function* (or *cost-to-go*), which we abuse notation and write as

$$J(t, x, U) = \sum_{s\geq t} g(s, X_s^{t,x,U}, U_s)$$

where $X^{t,x,U}$ solves the plant equation with control U and initial value $X_t = x$. With this notation $J = \mathbb{E}[J(0, x, U)]$ is our original optimization objective.

The basic principle of dynamic programming is then fairly simple. We observe that $J(t,x,U)$ depends on U only through the values of U_s for $s \geq t$. We then write

$$J(t,x,U) = g(t,x,U_t) + J(t+1, X_{t+1}^{t,x,U}, U).$$

So, with a further abuse of notation

$$J(t,x,U) = g(t,x,U_t) + J\Big(t+1, f(t,x,U_t), \{U_s\}_{s\geq t+1}\Big). \tag{1.1.2}$$

We can then optimize with respect to U_t and $\{U_s\}_{s\geq t+1}$ independently, to get the *Bellman equation*

$$\begin{aligned} V(t,x) &:= \inf_U J(t,x,U) \\ &= \inf_{U_t}\left\{ g(t,x,U_t) + \inf_{\{U_s\}_{s\geq t+1}} J\Big(t+1, f(t,x,U_t), \{U_s\}_{s\geq t+1}\Big)\right\} \\ &= \inf_{U_t}\Big\{ g(t,x,U_t) + V\Big(t+1, f(t,x,U_t)\Big)\Big\}. \end{aligned}$$

This allows us to compute the optimal cost function (or *value function*) V sequentially. Using V, we can then identify U_t as the arg min in the Bellman equation, which describes the (set of) optimal controls.

Example 1.1.1. Consider the one-dimensional Linear-Quadratic problem, where, for $t \leq T$,

$$\begin{aligned} X_{t+1} = a + bX_t + U_t \qquad &\Rightarrow \qquad f(t,x,u) = a + bx + u \\ g(t,x,u) = \alpha + \beta(x-\mu_x)^2 &+ \gamma(u-\mu_u)^2 \end{aligned}$$

We make an ansatz (i.e. an educated guess) that the value function is quadratic, so can be written in the form $V(t,x) = \pi_t + \rho_t(x-\xi_t)^2$, for some values of π_t, ρ_t, ξ_t. We have the trivial value $V(T+1,x) \equiv 0$, so can write $\pi_{T+1} = \rho_{T+1} = \xi_{T+1} = 0$. The Bellman equation then is

$$V(t,x) = \inf_u \Big\{ \alpha + \beta(x-\mu_x)^2 + \gamma(u-\mu_u)^2 + \pi_{t+1} + \rho_{t+1}\big(a+bx+u-\xi_{t+1}\big)^2 \Big\}$$

Basic calculus shows that the optimal strategy is of the form

$$u_t^* = \frac{\gamma\mu_u + \rho_{t+1}(\xi_{t+1} - a - bx)}{\gamma + \rho_{t+1}} = \frac{\gamma\mu_u + \rho_{t+1}(\xi_{t+1}-a)}{\gamma+\rho_{t+1}} - \frac{b\rho_{t+1}}{\gamma+\rho_{t+1}}x =: h_t + k_t x$$

and hence, by substitution,

$$\begin{aligned} &V(t,x) \\ &= \alpha + \beta(x-\mu_x)^2 + \gamma(h_t + k_t x - \mu_u)^2 + \pi_{t+1} + \rho_{t+1}\big(a + bx + h_t + k_t x - \xi_{t+1}\big)^2 \\ &= \alpha + \pi_{t+1} + \beta(x-\mu_x)^2 + k_t^2\gamma\Big(x - \frac{\mu_u - h_t}{k_t}\Big)^2 + \rho_{t+1}(b+k_t)^2\Big(x - \frac{\xi_{t+1} - a - h_t}{b+k_t}\Big)^2 \\ &= \Big[\beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2\Big]\Big(x - \frac{\beta\mu_x + k_t\gamma(\mu_u - h_t) + \rho_{t+1}(b+k_t)(\xi_{t+1}-a-h_t)}{\beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2}\Big)^2 \\ &\quad + \frac{\beta\gamma(k_t\mu_x - \mu_u + h_t)^2 + \beta\rho_{t+1}((b+k_t)\mu_x - \xi_{t+1} + a + h_t)^2 + \gamma\rho_{t+1}(b(\mu_u - h_t) + k_t(\mu_u - \xi_{t+1} - a))^2}{\beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2} \\ &\quad + \alpha + \pi_{t+1} \end{aligned}$$

From which, together with our ansatz, we can write the backward recursion

$$\begin{aligned} \rho_t &= \beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2, \\ \xi_t &= \frac{\beta\mu_x + k_t\gamma(\mu_u - h_t) + \rho_{t+1}(b+k_t)(\xi_{t+1}-a-h_t)}{\beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2}, \end{aligned}$$

and similarly for π_t (but note that this is not needed to compute the optimal strategy). Various algebraic simplifications of this are possible, as is making the parameters $\alpha, \beta, \gamma, \mu_x, \mu_u$ time dependent.

Remark 1.1.2. Even in this simple setting, there are some interesting things to say about dynamic programming.

- One way of looking at dynamic programming is as a computational tool. Instead of having to solve the high-dimensional constrained optimization problem where we find the optimal U subject to X being constrained to satisfy the specified dynamics, we solve a family of low dimensional, unconstrained optimization problems given by the dynamic programming equation. This may be computationally much easier, depending on our context.
- Another, more modelling-driven perspective, is that we might have an agent who is allowed to change their mind at any time. The dynamic programming equation tells us that our agent is dynamically-consistent, in that if we find an optimal strategy at time zero, then that strategy remains optimal at all future times (with the remaining-cost being used at time t) and, furthermore, if the agent changes to a different strategy, which at time t they might consider optimal, then at time $t = 0$ we are indifferent about such a change - the resulting changed policy will also be optimal.

The key fact that ensures dynamic programming holds here is the additive structure of J in (1.1.2), which ensures that J is monotone with respect to the future cost-to-go.

1.2 Discrete stochastic control

We now want to expand our class of problems to include randomness. This is a bit tricky, as our agent will be allowed to use their past observations when determining the control, which means their controls will also be random. We will now move in a somewhat abstract direction, and try and get some understanding of the basic structure of the control problem. Our aim is to describe what optimal strategies look like, in a fairly generic way, so that we can then use this description to solve explicit problems.

Even though we are considering a discrete-time setting, we will try and do things carefully. This means that we will use the tools and terminology of measure theory, as this gives us an efficient and precise way to study random processes. These tools also will allow us to more easily take our results into more difficult (continuous time) settings. Some introductory comments about this theory are in the appendix, and you can also consult the lecture notes for B8.1, or the textbook [2].

One somewhat uncommon object which we will need is the measurable essential supremum. This ensures we can take a maximum or minimum, of random objects, without worrying about measurability.

Definition 1.2.1. *Let $\mathcal{G}$ be a family of functions (measurable, on a σ-finite measure space $(\Omega, \mathcal{F}, \mu)$). Then there exists a measurable function g^* such that $g^* \geq g$ almost everywhere for all $g \in \mathcal{G}$, and if h is another function such that $h \geq g$ a.e. for all $g \in \mathcal{G}$, then $g^* \leq h$ almost everywhere. We say $g^* = \operatorname{ess\,sup} \mathcal{G}$.*

If $\mathcal{G}$ is upward directed (so for $g, g' \in \mathcal{G}$, there is $g'' \in \mathcal{G}$ with $g'' \geq g$ and $g'' \geq g'$), then the essential supremum can be approached by a sequence $g_n \subset \mathcal{G}$, that is $\operatorname{ess\,sup} \mathcal{G} = \lim_{n\to\infty} g_n$.

The essential infimum is defined analogously.

A proof that the essential supremum exists (and the limiting statement) can be found in the appendix (Theorem A.1.20). Essentially, given a family of random variables g_n (which is just a family of functions $g_n : \Omega \to \mathbb{R}$), this essential supremum is the pointwise supremum with respect to n for each $\omega \in \Omega$, but done in a way that makes sure we still have measurability.

As before, we suppose we have an agent who is going to choose a control U, which is a (random) process taking values in a set $\mathcal{U}$. We will assume that $\mathcal{U}$ is a topological space (so we can talk about open sets and hence (Borel) measurability).

1.2.1 Notation

We motivate our setup with an example:

Example 1.2.2. A company has to decide how to price and advertise a subscription product in the market – the number of subscribers changes randomly each day depending on the price and the number already subscribed (and possibly other factors, for example they might see momentum in the number of subscribers due to word-of-mouth). The reward (negative cost) is the number of subscribers at the start of the period multiplied by the price (which is the control), minus the cost of advertising. The control is the price and advertising level. We assume that the company knows the impact of their price and advertising strategy (although this can be relaxed), and has a fixed horizon over which they are trying to sell the product. The company is allowed to vary their control depending on how many subscribers they have had in the past, on seasonal effects, and based on other random events which may occur (for example, a competitor entering the market).

We start with a model of our world described by a filtered space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{T}})$. Here $\omega \in \Omega$ is the state of the world (you can think of it as the random seed, which determines the outcomes of every random event), while $\mathcal{F}_t$ determines the information available at time t (formally, it is the set of events whose outcome is known at time t).

While earlier we assumed that our controls changed the dynamics of a deterministic state variable, we will now suppose that our control modifies the probabilities of different outcomes occurring. We see this in our motivating example, as the choice of price and advertising will affect the probability that different numbers of subscribers will join in each period.

Example 1.2.3. A technical setting we will focus on later is where X_t represents the state of a Markov chain at each time, and $\mathcal{F}_t$ describes what we know from observing $\{X_0, X_1, ..., X_t\}$. In this world, it is natural that a control might change the probabilities of transitions from one state to another, which motivates the following construction.

We assume an agent is choosing a control (process) U (which we also call a policy or a strategy, as convenient), which changes the probabilities of different outcomes in the world. This control takes values in $\mathcal{U}$, which we assume is a topological space (in particular, it has a Borel σ-algebra, and measurability is defined using this σ-algebra.). In our example, $U \in \mathcal{U} = \{\text{advertising levels}, \text{price levels}\}$. We write the expectation when using control U as $\mathbb{E}^U$.

Let's suppose that our agent is trying to minimize their expected cost

$$J(U) = \mathbb{E}^U\Big[\Big(\sum_{s=0}^{T-1} g(\omega, s, U_s)\Big) + \Phi(\omega)\Big]$$

where $g : \Omega \times \mathbb{T} \times \mathcal{U} \to \mathbb{R}$, $\Phi : \Omega \to \mathbb{R}$. We separate out the cost at the final time, to emphasize that this does not depend (directly) on the control chosen. The running cost g is assumed to be *adapted*, that is, $g(s, u)$ is $\mathcal{F}_s$-measurable for every s and every $u \in \mathcal{U}$, which we interpret as stating that $g(s, u)$ is known at time s.

We say a strategy U is optimal if it minimizes $J(U)$. As usual, we drop the ω arguments of our terms to minimize notation.

As before, we define the cost-to-go function

$$J(t, U) := \mathbb{E}^U\Big[\Big(\sum_{s=t}^{T-1} g(s, U_s)\Big) + \Phi\Big|\mathcal{F}_t\Big]$$

We note that $J(t, U)$ is a random process (there's hidden dependence on the random state of the world ω), as we've used a *conditional* expectation.

Building on the deterministic theory we saw before, our job is to characterize optimal strategies in a dynamic way, so that we can avoid solving the very high dimensional optimization problem we have just written down. This is made worse by the fact that we are now allowed to use strategies which depend on the (random) state of the world ω, which is not something which made sense in the deterministic setting. Therefore, we will need to do a little more work in order to carefully specify what we mean by an optimal strategy.

Before we attempt this, we will give a more careful description of different classes of strategies which we might wish to consider.

Definition 1.2.4. *We say a strategy U is* admissible *if U is an $\{\mathcal{F}_t\}_{t\in\mathbb{T}}$-adapted process taking values in $\mathcal{U}$ (i.e. U_t is $\mathcal{F}_t$-measurable for all t). We write $\mathbb{U}$ for the space of admissible strategies.*

If our filtration is generated by a state process X, this can be reexpressed as saying U_t is a (Borel measurable) function of $\{X_s\}_{s\le t}$ (by the Doob–Dynkin lemma).

Assumption 1.2.5. *To avoid measure-theoretic issues*[1]*, we make the following two assumptions:*

- *There is some measure $\mathbb{P}^{\text{ref}}$ such that, for all $U \in \mathbb{U}$, we know $\mathbb{P}^U(A) = 0$ only if $\mathbb{P}^{\text{ref}}(A) = 0$, that is $\mathbb{P}^U$ is absolutely continuous with respect to $\mathbb{P}^{\text{ref}}$.*
- *The map $(\{u_t\}_{t\in\mathbb{T}}, \omega) \mapsto \frac{\mathrm{d}\mathbb{P}^U}{\mathrm{d}\mathbb{P}^{\text{ref}}}(\omega)$ (where $U_t = u_t$) is measurable as a map $(\mathcal{U}^{\mathbb{T}}, \Omega) \to \mathbb{R}$.*

Given this definition, we can now see that it's possible to combine different strategies, and not leave the class of admissible strategies. The point here is that we're allowed to switch strategies depending on the random outcomes we have observed so far.

[1]For fixed U, the Radon–Nikodym derivative of $\mathbb{P}^U$ with respect to $\mathbb{P}^{\text{ref}}$ is a random variable, denoted $\frac{\mathrm{d}\mathbb{P}^U}{\mathrm{d}\mathbb{P}^{\text{ref}}}$, such that for all $\mathbb{P}^U$-integrable Z,

$$\mathbb{E}^U[Z] = \mathbb{E}^{\text{ref}}\big[\frac{\mathrm{d}\mathbb{P}^U}{\mathrm{d}\mathbb{P}^{\text{ref}}} Z\big].$$

With this, we get a version of Bayes' theorem for conditioning on a general σ-algebra

$$\mathbb{E}^U[Z|\mathcal{F}_t] = \mathbb{E}^{\text{ref}}\big[\frac{\mathrm{d}\mathbb{P}^U}{\mathrm{d}\mathbb{P}^{\text{ref}}} Z\Big|\mathcal{F}_t\big]\Big/\mathbb{E}^{\text{ref}}\big[\frac{\mathrm{d}\mathbb{P}^U}{\mathrm{d}\mathbb{P}^{\text{ref}}}\Big|\mathcal{F}_t\big].$$

The second assumption guarantees that $\mathbb{E}^U[Z|\mathcal{F}_t]$ is a random variable (in particular, is measurable) for all integrable Z and $U \in \mathbb{U}$. Taken together, these assumptions ensure that the conditional expectation is simultaneously defined for all U (as it is defined $\mathbb{P}^{\text{ref}}$-a.e.).

Definition 1.2.6. *Let $t \in \mathbb{T}$ and $A \in \mathcal{F}_t$. Consider any two admissible strategies U and U'. Then the* pasting *of U and U' (at t, on the set A) is given by*

$$U''_s = (1 - 1_{s\geq t}1_A)U_s + 1_{s\geq t}1_A U'_s = \begin{cases} U_s & \text{if } s < t, \\ U_s & \text{if } s \geq t, \omega \notin A, \\ U'_s & \text{if } s \geq t, \omega \in A \end{cases}$$

In other words, after time t, we switch to U' if event A occurs.

Proposition 1.2.7. *The pasting U'' (of two admissible strategies) is also an admissible strategy.*

Proof. This is simply stating that pasting preserves measurability of a process. We can check this, by observing that, for any (Borel) measurable set $B \subset \mathcal{U}$ and any $s \geq t$, we have

$$\{\omega : U''_s \in B\} = \big(\{\omega : U'_s \in B\} \cap A\big) \cup \big(\{\omega : U_s \in B\} \cap A^c\big)$$

and the right hand term is $\mathcal{F}_s$-measurable, by admissibility of U_s and U'_s. Alternatively, we see that $1_{s\leq t}$, 1_A, U_t and U'_t are all $\mathcal{F}_t$-measurable, and products and sums preserve measurability, so U''_t is also $\mathcal{F}_t$-measurable. □

The final thing we need to specify is how U changes the probabilities. We do this in terms of pasting.

Definition 1.2.8. *We say our control problem is* dynamic *if, for any random variable ξ, with U'' as defined above, for $s \geq t$ we have*

$$\mathbb{E}^{U''}[\xi|\mathcal{F}_s] = 1_A\mathbb{E}^{U'}[\xi|\mathcal{F}_s] + 1_{A^c}\mathbb{E}^{U}[\xi|\mathcal{F}_s],$$

and for $s < t$,

$$\mathbb{E}^{U''}[\xi|\mathcal{F}_s] = \mathbb{E}^{U}\Big[\mathbb{E}^{U''}[\xi|\mathcal{F}_t]\Big|\mathcal{F}_s\Big].$$

Essentially, the above definition tells us that switching from U to U' at time t, if A occurs, only changes the probabilities of events which are not already known at time t, doesn't change the probabilities of events which only occur if A doesn't occur, and, if we switch, the conditional probabilities (given our information at time t) are completely determined by U' (not U). We will assume without further comment that our problem is dynamic.

We need one assumption on the integrability of our costs, in order to avoid trivialities.

Assumption 1.2.9. *The costs g and Φ are such that J satisfies the bounds*

- *For all $t \in \mathbb{T}$ and $U \in \mathbb{U}$, we know $\mathbb{E}^U[J(t, U)] > -\infty$; (i.e. there are no infinitely desirable controls);*
- *There exists at least one control $\bar{U} \in \mathbb{U}$ such that $\mathbb{E}^U[|J(t, \bar{U})|] < \infty$ for all $U \in \mathbb{U}$ and $t \in \mathbb{T}$; (i.e. there is at least one control which is always acceptable).*

It is convenient to allow, at least in principle, that the cost can be infinite – this encodes the idea that there can be situations and controls which must be avoided, and so are assigned infinite costs.

1.2.2 The martingale principle and dynamic programming

We are now ready to prove the dynamic programming principle for our problem. As we want to, at least in principle, allow any admissible control, this is a bit tricky, as we can't assume any Markov properties for our processes.

Definition 1.2.10. *We define the* value process

$$V_t = \operatorname*{ess\,inf}_{U \in \mathbb{U}} J(t, U)$$

where the ($\mathcal{F}_t$-measurable) essential infimum is taken over all admissible controls.

Lemma 1.2.11. *For each $t \in \mathbb{T}$, the set of random variables $\{J(t,U)\}_{U\in\mathbb{U}}$ is up/downward directed, that is, for any admissible $U, U' \in \mathbb{U}$ there exists $U'' \in \mathbb{U}$ such that $J(t,U'') \leq \min\{J(t,U), J(t,U')\}$. In particular, this implies that, for any $t' < t$, and any $U \in \mathbb{U}$,*

$$\operatorname*{ess\,inf}_{U'} \mathbb{E}^U[J(t',U')|\mathcal{F}_t] = \mathbb{E}^U\big[\operatorname*{ess\,inf}_{U'} J(t',U')|\mathcal{F}_t\big] = \mathbb{E}^U\big[V_{t'}|\mathcal{F}_t\big].$$

Observe that these two essential infima are quite different – the first optimizes the expected value, and is $\mathcal{F}_t$-measurable, while the second optimizes $J(t',U')$, and is $\mathcal{F}_{t'}$-measurable.

Proof. The upward/downward directed property follows by pasting, with the set $A = \{J(t,U) > J(t,U')\}$. We know immediately that, for any U, U',

$$\mathbb{E}^U[J(t,U')] \geq \mathbb{E}^U\big[V_t\big]$$

Therefore

$$\operatorname*{ess\,inf}_{U'} \mathbb{E}^U[J(t,U')] \geq \mathbb{E}^U\big[V_t\big].$$

Conversely, from Theorem A.1.20, we know that there is a sequence U^n such that $J(t,U^n) \to V_t$ almost surely. Without loss of generality, we can assume $J(t,U^n) \leq J(t,\bar{U})$, where $\bar{U}$ is the reference control in Assumption 1.2.9, for which we know $\mathbb{E}^U[|J(t,\bar{U})|] \leq \infty$. Therefore, $J(t,\bar{U}) - J(t,U_n)$ is an increasing nonnegative sequence, and monotone convergence shows that

$$\begin{aligned}\mathbb{E}^U[J(t,\bar{U})] - \lim_{n\to\infty} \mathbb{E}^U[J(t,U_n)] &= \lim_{n\to\infty} \mathbb{E}^U[J(t,\bar{U}) - J(t,U_n)] = \mathbb{E}^U[J(t,\bar{U}) - \lim_{n\to\infty} J(t,U_n)] \\ &= \mathbb{E}^U[J(t,\bar{U})] - \mathbb{E}^U[\lim_{n\to\infty} J(t,U_n)].\end{aligned}$$

It follows that

$$\operatorname*{ess\,inf}_{U'} \mathbb{E}^U[J(t,U')|\mathcal{F}_s] \leq \lim_{n\to\infty} \mathbb{E}^U[J(t,U^n)|\mathcal{F}_s] = \mathbb{E}^U\big[\lim_{n\to\infty} J(t,U^n)\big|\mathcal{F}_s\big] = \mathbb{E}^U[\operatorname*{ess\,inf}_{U'} J(t,U')|\mathcal{F}_s].$$

□

Theorem 1.2.12 (Bellman equation / Martingale optimality principle)**.** *The value process V satisfies the Bellman equation*

$$V_t = \operatorname*{ess\,inf}_{U} \mathbb{E}^U\Big[\Big(\sum_{s=t}^{t'-1} g(s,U_s)\Big) + V_{t'}\Big|\mathcal{F}_t\Big].$$

Furthermore, for any admissible control U with integrable costs, the process

$$M_t^U = \Big(\sum_{s=0}^{t-1} g(s, U_s)\Big) + V_t$$

is a submartingale ($M_t^U \leq \mathbb{E}^U[M_{t+1}^U|\mathcal{F}_t]$ for all t), and is a martingale ($M_t^U = \mathbb{E}^U[M_{t+1}^U|\mathcal{F}_t]$ for all t) if and only if U is optimal, that is, it minimizes $J(U)$.

Proof. For a given policy U, using the tower property of conditional expectation, for $t \leq t'$,

$$\begin{aligned} J(t,U) &= \mathbb{E}^U\Big[\Big(\sum_{s=t}^{T-1} g(s,U_s)\Big) + \Phi\Big|\mathcal{F}_t\Big] \\ &= \mathbb{E}^U\Big[\Big(\sum_{s=t}^{t'-1} g(s,U_s)\Big) + \mathbb{E}^U\Big[\Big(\sum_{s=t'}^{T-1} g(s,U_s)\Big) + \Phi\Big|\mathcal{F}_{t'}\Big]\Big|\mathcal{F}_t\Big] \\ &= \mathbb{E}^U\Big[\Big(\sum_{s=t}^{t'-1} g(s,U_s)\Big) + J(t',U)\Big|\mathcal{F}_t\Big]. \end{aligned}$$

We now separate the behaviour of U before and after t' , by considering the pasted strategy

$$\tilde{U}_s = 1_{\{s<t'\}}U_s + 1_{\{s\geq t'\}}U'_s,$$

for U' an arbitrary admissible control. Then we see that, by pasting and Lemma 1.2.11,

$$\begin{aligned} V_t &= \operatorname*{ess\,inf}_{\tilde{U}} J(t,\tilde{U}) = \operatorname*{ess\,inf}_{U,U'} J\Big(t, 1_{\{t'<t\}}U_s + 1_{\{t'\geq t\}}U'_s\Big) \\ &= \operatorname*{ess\,inf}_{U,U'} \mathbb{E}^U\Big[\Big(\sum_{s=t}^{t'-1} g(s,U_s)\Big) + J(t',U')\Big|\mathcal{F}_t\Big] \\ &= \operatorname*{ess\,inf}_{U} \mathbb{E}^U\Big[\Big(\sum_{s=t}^{t'-1} g(s,U_s)\Big) + \operatorname*{ess\,inf}_{U'} J(t',U')\Big|\mathcal{F}_t\Big] \\ &= \operatorname*{ess\,inf}_{U} \mathbb{E}^U\Big[\Big(\sum_{s=t}^{t'-1} g(s,U_s)\Big) + V_{t'}\Big|\mathcal{F}_t\Big]. \end{aligned}$$

This establishes the first statement.

It follows that, for any admissible U,

$$V_t \leq \mathbb{E}^U\Big[\Big(\sum_{s=t}^{t'-1} g(s,U_s)\Big) + V_{t'}\Big|\mathcal{F}_t\Big].$$

To conclude, we add the costs before time t, to get

$$M_t^U = \Big(\sum_{s=0}^{t-1} g(s,U_s)\Big) + V_t \leq \mathbb{E}^U\Big[\Big(\sum_{s=0}^{t'-1} g(s,U_s)\Big) + V_{t'}\Big|\mathcal{F}_t\Big] = \mathbb{E}^U\Big[M_{t'}^U\Big|\mathcal{F}_t\Big],$$

which proves the submartingale property.

Finally, suppose U^* is optimal. Then we know, for any stopping time $\tau \leq T$

$$\begin{aligned} V_0 = M_0^{U^*} &\leq \mathbb{E}^{U^*}\left[M_\tau^{U^*}\middle|\mathcal{F}_0\right] \\ &\leq \mathbb{E}^{U^*}\left[M_T^{U^*}\middle|\mathcal{F}_0\right] = \mathbb{E}^{U^*}\left[\Big(\sum_{s=0}^{T-1} g(s,U_s^*)\Big) + \Phi\middle|\mathcal{F}_0\right] = J(0,U^*) = V_0. \end{aligned}$$

Hence all these terms are equal, and by Lemma A.1.19 we know that M^{U^*} is a martingale. The converse statement follows from direct calculation. □

Using the martingale optimality principle, we can give a version of Bellman's dynamic programming principle. These two statements encapsulate the key ideas of dynamic programming – if we have a policy which is optimal at time t, then it will continue to be optimal after that time, and we don't mind, at time t, which optimal policy we choose at time t'. Because of this, we can simply say a control U is optimal, without specifying the time at which we are evaluating it!

Theorem 1.2.13 (Dynamic Programming Principle)**.** *We say an admissible control U is optimal at time t if $J(t,U) \leq J(t,U')$ a.s. for all admissible controls U', or equivalently, $J(t,U) = \operatorname{ess\,inf}_{U'} J(t,U') = V_t$. Then it holds that, for all $t' \geq t$,*

- *if U is optimal at t, then U is optimal at t';*
- *if $U^{(t)}$ is optimal at t, and $U^{(t')}$ is optimal at t', then the pasted strategy*

$$\tilde{U}_s = 1_{\{s<t'\}} U_s^{(t)} + 1_{\{s\geq t'\}} U_s^{(t')}$$

is optimal at t.

Proof. We know M^U is a submartingale. For stopping times $\tau \geq t$, this implies

$$\begin{aligned} \Big(\sum_{s=0}^{t-1} g(s,U_s)\Big) + V_t = M_t^U &\leq \mathbb{E}^U\left[M_\tau^U\middle|\mathcal{F}_t\right] \\ &\leq \mathbb{E}^U\left[M_T^U\middle|\mathcal{F}_t\right] = \mathbb{E}^U\left[\Big(\sum_{s=0}^{T-1} g(s,U_s)\Big) + \Phi\middle|\mathcal{F}_t\right] \\ &= \Big(\sum_{s=0}^{t-1} g(s,U_s)\Big) + J(t,U), \end{aligned}$$

and we see that U is optimal at t if and only if this is an equality, or equivalently, the process $M'_s := 1_{\{s\geq t\}}(M_s^U - M_t^U)$ is a martingale. However, this implies that $M_{t'}^U = \mathbb{E}^U[M_T^U|\mathcal{F}_{t'}]$, which shows that U is optimal at t'. The first statement follows.

To prove the second, we see that the strategy $\tilde{U}$ generates the process

$$M_s^{\tilde{U}} = \begin{cases} M_s^{U^{(t)}} & \text{if } s < t \\ M_s^{U^{(t')}} - M_t^{U^{(t')}} + M_t^{U^{(t)}} & \text{if } s \geq t \end{cases}$$

and calculating expectations shows that this is a martingale if M^U and $M^{U'}$ are martingales. □

1.3 Finite state Markov Decision Problems

The previous section gave a general approach, but this is not so convenient for computation. We now focus on a special case, where we are interested in controlling a Markov process (so we don't have any dependence on the past, except through our current state X). This will allow us to restrict our attention to *feedback* controls, and lead to nicer numerical approaches.

In our motivating example, we might say that the state of the system is the current number of subscribers, in particular, the number who joined yesterday has no impact on how many will join tomorrow (so there is no momentum). This is explicitly a modelling choice which we are making.

1.3.1 Controlled Markov chains

Formally, we describe our problem through the use of the transition law (aka kernel, density, generator),

$$p(x';t,x,u) = \mathbb{P}\Big(X_{t+1} = x' \Big| X_t = x, U_t = u\Big).$$

so $p : \mathcal{X} \times \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to [0,1]$, with $\mathcal{X}$ being the (discrete) state space in which X takes values, and $\mathbb{T} = \{0,1,...,T\}$. In order for X to have Markov-like properties, we need this to describe the probabilities conditional on $\mathcal{F}_t$, that is,

$$\mathbb{P}^U\big[X_{t+1} = x' \big| \mathcal{F}_t\big] = p(x';t,X_t,U_t)$$

(this is weaker than assuming our filtration only contains information about X, but is enough for us). We will focus here on discrete cases, but we can see that there is an extension to the case where X takes continuous values, in which case we modify p to describe a measure on the (infinite) state space $\mathcal{X}$, so

$$p(A;x,u,t) = \mathbb{P}\Big(X_{t+1} \in A \Big| X_t = x, U_t = u\Big)$$

for $A \subseteq \mathcal{X}$ a measurable set. Analysing this case would then require more integrability assumptions than we will need when $\mathcal{X}$ is finite.

1.3.2 Finite horizon MDPs

In the previous section, we considered the general control of a random system. We will now specialize this discussion to understand the optimal control of a Markov chain. We assume that our agent wants to minimize a quantity

$$J(U) = \mathbb{E}^U\Big[\sum_{s=0}^{T-1} g(s,X_s,U_s) + \Phi(X_T)\Big]$$

where we limit our costs to only depend on the current state X_s of the controlled system (whereas previously it could depend on the whole random seed ω). The Bellman equation states that our value process satisfies

$$V_t = \mathbb{E}^{U^*}\big[g(t,X_t,U_t^*) + V_{t+1}\big|\mathcal{F}_t\big],$$

with terminal value $V_T = \Phi(X_T)$, where U^* is an optimal control.

There are a range of problems of this type – see [9] for a classic selection. Two examples cited there are:

Example 1.3.1. Consider a hydroelectric power plant which needs to decide whether to release water from a dam and generate power. The excess power can be sold (leading to a negative cost), and the dam refils randomly each day, up to a maximum level. By discretizing the level in the dam, which we write X, we obtain an MDP.

Example 1.3.2. How much pest control should be used to manage weevils in an alfalfa crop? The state is the current condition of the crop and weevil levels, and the cost is made up of the level of production and the cost of pest control.

A classic financial example is:

Example 1.3.3. An insurance contract is written whose payoff depends on the weather in 6 months time. An investor needs to hedge their position by investing in weather sensitive assets (for example, in heating oil futures), in order to manage their risk. The state is the combination of weather forecasts (which are uncontrolled) and the investor's wealth, and the control is how much to invest in the oil market.

These examples show that there is a wide range of possible applications of this theory. Some examples will be considered in more detail on problem sheets.

A natural property, which we will now try and prove, is that the control doesn't depend on the past – clearly the dynamics don't incorporate the past, so the only relevant source of randomness is the current state of the controlled system.

Assumption 1.3.4. *The space of controls $\mathcal{U}$ is compact, and the function g and the transition density p depend continuously on the choice of control u.*

This assumption will be enough to guarantee that an optimal control exists, which simplifies our arguments. Without this assumption we can still do quite a lot, but the analysis is more fiddly (as we will see in continuous time).

Theorem 1.3.5. *Under Assumption 1.3.4,*

- *there is a function $v : \mathbb{T} \times \mathcal{X} \to \mathbb{R}$, known as the value function, such that the value process can be written $V_t = v(t, X_t)$;*
- *the value function satisfies the Bellman recursion*

$$v(t,x) = \min_{u \in \mathcal{U}} \Big\{ g(t,x,u) + \sum_{x' \in \mathcal{X}} p(x'; t, x, u) v(t+1, x') \Big\}$$

  with $v(T, x) = \Phi(x)$;
- *there exists at least one optimal control which is of* feedback *type, that is, $U_t^* = u^*(t, X_t)$ for some function $u^* : \mathbb{T} \times \mathcal{X} \to \mathcal{U}$, which achieves the minimum in the Bellman recursion, for every t, x.*

Proof. We proceed by backward induction. At the terminal time, the value is clearly given by $V_T = \Phi(X_T) = v(T, X_T)$, and we don't need to define the control here.

Now assume that $V_{t+1} = v(t+1, X_{t+1})$. We know that

$$\begin{aligned} V_t &= \operatorname*{ess\,inf}_U \left\{ \mathbb{E}^U \Big[g(t, X_t, U_t) + V_{t+1} \Big| \mathcal{F}_t \Big] \right\} \\ &= \operatorname*{ess\,inf}_U \left\{ g(t, X_t, U_t) + \mathbb{E}^U \Big[v(t+1, X_{t+1}) \Big| \mathcal{F}_t \Big] \right\} \\ &= \operatorname*{ess\,inf}_{U_t \in \mathcal{U}} \left\{ g(t, X_t, U_t) + \sum_{x' \in \mathcal{X}} p(x'; t, X_t, U_t) v(t+1, x') \right\} \end{aligned}$$

For every possible value of $X_t \in \mathcal{X}$, the term inside the infimum is a continuous function of U_t, which takes values in a compact set. Therefore the infimum will be realized at some value[2], which we denote $u^*(t, X_t)$. Substituting this value gives

$$V_t = g(t, X_t, u^*(t, X_t)) + \sum_{x' \in \mathcal{X}} p(x'; t, X_t, u^*(t, X_t)) v(t+1, x')$$

Now observe the right hand side is just a function of X_t, which we denote $v(t, X_t)$.

By induction, we hence construct $v(t, x)$ and $u^*(t, x)$ for all t, x, as required. □

In many approaches to control, we would have begun by only considering feedback controls, which simplifies the argument a little. Here we've done the hard work, and so have proven that there are optimal feedback controls *within the class of all admissible controls*. That is, when solving an MDP, there's no possible advantage to remembering past values of the state when determining your control.

Remark 1.3.6. One thing which is clear from this construction is that the value function and cost-to-go are not really intrinsic to the control problem – we can easily find alternative value functions, with slightly different properties, which work just as well. A common example is to define the discounted value, for some $\rho > 0$,

$$v^\rho(t, X_t) = \min_U \mathbb{E}^U \Big[\sum_{s=t}^{T-1} e^{-\rho(s-t)} g(s, X_s, U_s) + e^{-\rho(T-t)} \Phi(X_T) \Big| \mathcal{F}_t \Big]$$

which is related to our earlier value by $v^\rho = e^{\rho t} v$, when

$$v(t, X_t) = \min_U \mathbb{E}^U \Big[\sum_{s=t}^{T-1} e^{-\rho s} g(s, X_s, U_s) + e^{-\rho T} \Phi(X_T) \Big| \mathcal{F}_t \Big]$$

(which is within the class we have considered, by perturbing g and Φ). The Bellman equation is then modified to

$$v^\rho(t, x) = \min_{u \in \mathcal{U}} \Big\{ g(t, x, u) + e^{-\rho} \sum_{x' \in \mathcal{X}} p(x'; t, x, u) v^\rho(t+1, x') \Big\}.$$

We will return to this in the coming section.

1.4 Infinite-horizon Discounted MDPs

A further extension of our approach is to consider problems on an infinite horizon, but with a discounting term. Hence, we now have $\mathbb{T} = \{0, 1, ...\}$. Rather than repeat our general setup, we will give a version of these results for MDPs, based on our finite-time results.

[2] As $\mathcal{X}$ is discrete, we don't need to worry about measurability here, but otherwise there are various 'measurable selection theorems' which can help, as long as we also assume continuity of f and Φ in x, for every t, u. See Bertsekas and Shreve, Chapter 7, for details. We will come back to this point when we reach continuous stochastic control.

We consider an agent who seeks to minimize (at each time t),

$$J(t,U) = \mathbb{E}^U\Big[\sum_{s=t}^{\infty} e^{-\rho(s-t)} g(X_s,U_s)\Big|\mathcal{F}_t\Big]$$

where $\rho > 0$ is a fixed constant. Note that we've removed direct dependence on time in g. Under Assumption 1.3.4, we now notice that $\max_{x\in\mathcal{X},u\in\mathcal{U}} |g(x,u)| \le \bar{g}$ for some $\bar{g} < \infty$. Our control has impact through a transition law $p(x';x,u)$ *which does not depend on time* t. It's easy to check that the dynamic programming principle still holds for this problem.

Theorem 1.4.1. *There exists a function $v : \mathcal{X} \to \mathbb{R}$, such that*

$$v(X_t) = \min_U J(t,U)$$

for all $t \in \mathbb{T}$. This v (again called the value function) is a fixed point of the Bellman recursion

$$v(x) = \min_{u\in\mathcal{U}} \Big\{ g(x,u) + e^{-\rho} \sum_{x'\in\mathcal{X}} p(x';x,u) v(x') \Big\}$$

and there is an optimal strategy of the form $U_t = u^*(X_t)$, where $u^*(x)$ achieves the minimum in this Bellman recursion.

Proof. We will consider this problem by approximating with a finite horizon problem (this will also be relevant for understanding numerical methods). We observe that, for any $T \in \mathbb{T}$,

$$\Big|\sum_{t=T}^{\infty} e^{-\rho t} g(X_t,U_t)\Big| \le \sum_{t=T}^{\infty} e^{-\rho t}\bar{g} = \frac{e^{-\rho T}}{1-e^{-\rho}}\bar{g}$$

Therefore, for any $T > t$,

$$\begin{aligned}
\mathbb{E}^U\Big[\sum_{s=t}^{T-1} e^{-\rho(s-t)} g(X_s,U_s) - \frac{e^{-\rho(T-t)}}{1-e^{-\rho}}\bar{g}\Big|\mathcal{F}_t\Big] &=: J_T^-(t,U) \\
&\le J(t,U) \\
&\le J_T^+(t,U) := \mathbb{E}^U\Big[\sum_{s=t}^{T-1} e^{-\rho(s-t)} g(X_s,U_s) + \frac{e^{-\rho(T-t)}}{1-e^{-\rho}}\bar{g}\Big|\mathcal{F}_t\Big].
\end{aligned}$$

However, both $J_T^+(t,U)$ and $J_T^-(t,U)$ are then the costs for finite horizon problems up to time T (with discount rate ρ). By our earlier results, we see that there are value functions $v_T^+(t,x)$ and $v_T^-(t,x)$, satisfying Bellman recursions for $t < T$, such that

$$v_T^-(t,X_t) \le \operatorname*{ess\,inf}_U J(t,U) \le v_T^+(t,X_t).$$

Clearly, for any U,

$$0 \le J_T^+(t,U) - J_T^-(t,U) \le 2\frac{e^{-\rho(T-t)}}{1-e^{-\rho}}\bar{g},$$

so if U^- is the optimizer of J_T^-, we know

$$0 \le v_T^+(t,X_t) - v_T^-(t,X_t) \le J_T^+(t,U^-) - v_T^-(t,X_t) \le 2\frac{e^{-\rho(T-t)}}{1-e^{-\rho}}\bar{g} \to 0 \text{ as } T\to\infty.$$

Therefore, by the sandwich theorem we conclude that $\operatorname{ess\,inf}_U J(t,U) = \lim_{T\to\infty} v_T^+(t,X_t)$. It follows that (as the right hand side of this limit is just a function), we know $\operatorname{ess\,inf}_U J(t,U) = \tilde{v}(t,X_t)$ for some function $\tilde{v}$, and as $v_T^\pm$ both satisfy Bellman recursions, so does $\tilde{v}$:

$$\tilde{v}(t,x) = \min_{u\in\mathcal{U}} \Big\{ g(x,u) + e^{-\rho} \sum_{x'\in\mathcal{X}} p(x';x,u)\tilde{v}(t+1,x') \Big\},$$

and there is an optimal feedback policy $\tilde{u}^*(t,x)$ which achieves the minimum on the right hand side.

The next step is to show that we can eliminate time dependence. We write the expected value in terms of the transition law, this is most easily done by recursively defining the multistep transition probability, for a general feedback policy u, by $p_{t,t+1}(x';x,u) = p(x';x,u(t,x))$ and

$$p_{t,s+1}(x';x,u) = \sum_{x''\in\mathcal{X}} p(x';x'',u(s,x''))p_{t,s}(x'';x,u).$$

This gives us a deterministic way to represent $\mathbb{P}\big(X_s = x' \big| X_t = x, \{U_n = u(n,X_n)\}_{n\in\mathbb{T}}\big) = p_{t,s}(x',x,u)$. Using this, and knowing that there is an optimal (time-dependent) feedback policy, we can express our value function as a minimum

$$\begin{aligned}\tilde{v}(t,x) &= \min_{u:\mathbb{T}\times\mathcal{X}\to\mathcal{U}} \sum_{s=t}^{\infty} \sum_{x'\in\mathcal{X}} p_{t,s}(x';x,u)e^{-\rho(s-t)} g(x',u(s,x')) \\ &= \min_{u:\mathbb{T}\times\mathcal{X}\to\mathcal{U}} \sum_{s=0}^{\infty} \sum_{x'\in\mathcal{X}} p_{0,s}(x';x,u(\cdot+t,\cdot))e^{-\rho s} g(x',u(s+t,x')).\end{aligned}$$

Now see that this is the same optimization problem for every value of t, so we can take a single solution u, and know $u^t := u(\cdot - t,\cdot)$ is optimal at t for all t. However, the resulting action $U_t = u^t(t,x) = u(0,x)$ is then just the pasting together of u^t at each time, so is optimal by dynamic programming. That is, there exists a $u^* : \mathcal{X}\to\mathcal{U}$ such that

$$\begin{aligned}\tilde{v}(t,x) &= \sum_{s=t}^{\infty} \sum_{x'\in\mathcal{X}} p_{t,s}(x';x,u^*)e^{-\rho(s-t)} g(x',u^*(x')) \\ &= \sum_{s=0}^{\infty} \sum_{x'\in\mathcal{X}} p_{0,s}(x';x,u^*)e^{-\rho s} g(x',u^*(x')) =: v(x).\end{aligned}$$

Substituting v into the Bellman recursion for $\tilde{v}$ finishes the proof. □

Remark 1.4.2. If we do the work of constructing the infinite-horizon discounted problem for infinite-space processes then not much changes (apart from technicalities involving transition laws becoming measures), provided we assume that f is bounded. If this is not the case, some integrability constraints are needed, in order to take the limit correctly from the finite horizon problem to the infinite horizon problem.

In our motivating example, the infinite horizon case corresponds to the setting where the subscription campaign has no fixed end date. This may be particularly useful as a model if the campaign will last a long time, as it gives us a simplified value function and strategy (no time dependence), which makes understanding and implementation easier. By our construction, we can see that the infinite and finite-horizon problems become similar as $T\to\infty$.

This analysis gives us a fairly comprehensive description of these problems, but we still need to find ways of solving the relevant equations numerically, in order for this to be practically useful. This is the theme of our next topic.

Chapter 2

Numerical methods and Reinforcement Learning

In the previous part of the course we have shown how optimal control theory works in a discrete-time context. In particular, we have shown that the optimal control can be found, together with the value function (which represents the optimal cost-to-go), by solving a recursive equation.

In practice, this remains computationally difficult, particularly when the state space $\mathcal{X}$ is large (and especially if it is high-dimensional). This has lead to the study of various numerical approximations to the control problem, which we will now consider.

Many numerical approaches focus on the infinite-horizon discounted setting, which we will focus on in this chapter. This has the advantage of avoiding time dependence in our solution, while still displaying a wide range of technical challenges.

2.1 Value iteration

Value iteration is the most fundamental approximation for control problems, and focusses on approximating $v(x)$ directly. As $x \in \mathcal{X}$ and $|\mathcal{X}| < \infty$, we can identify functions $w : \mathcal{X} \to \mathbb{R}$ with vectors in $\mathbb{R}^{|\mathcal{X}|} \equiv \mathbb{R}^{\mathcal{X}}$, with components $w_i := w(x_i)$. We will use this identification liberally, to make notation simpler.

Assumption 2.1.1 (Assumptions for Value iteration)**.** *We know the transition law $p(x'; x, u)$, and costs $g(x, u)$ perfectly.*

We recall that v is our value function, defined[1] by

$$v(x) = \min_u J(x, u) = \min_u \mathbb{E}^u \Big[\sum_{t=0}^{\infty} e^{-\rho t} g(X_t, u(X_t)) \Big| X_0 = x \Big].$$

[1]Here, as we focus on time homogenous feedback controls (so our world is Markovian and independent of time), we can take J as being a function of the initial state x, rather than a random variable and dependent on t.

Definition 2.1.2. *We define the Bellman valuation operator $\mathcal{T}_u$ by*

$$\big(\mathcal{T}_u\hat{v}\big)(x) = g(x,u) + e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u)\hat{v}(x').$$

We also define the Bellman optimality operator $\mathcal{T}:\mathbb{R}^{\mathcal{X}}\to\mathbb{R}^{\mathcal{X}}$ by

$$\big(\mathcal{T}\hat{v}\big)(x) = \min_{u\in\mathcal{U}}\Big\{g(x,u) + e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u)\hat{v}(x')\Big\} = \min_u \big(\mathcal{T}_u\hat{v}\big)(x)$$

With this notation, the value associated with a specific control u is a fixed point of $\mathcal{T}_u$

$$J(x,u) = \big(\mathcal{T}_u J(\cdot,u)\big)(x).$$

The key idea of value iteration is to observe that v satisfies the Bellman equation

$$v(x) = \min_{u\in\mathcal{U}}\Big\{g(x,u) + e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u)v(x')\Big\} = (\mathcal{T}v)(x)$$

and so v is a fixed point of $\mathcal{T}$.

Definition 2.1.3. *The* value iteration *sequence is defined by an initial value $v_0\in\mathbb{R}^{\mathcal{X}}$, and the iteration $v_n = \mathcal{T}v_{n-1}$.*

Value iteration is fairly straightforward to implement. We start with some guess for the value function, then iteratively refine it by computing the Bellman operator, which corresponds to finding the best strategy *given the value function at the next time step.* If this converges to a unique fixed point, then that fixed point must be the value function.

Remark 2.1.4. We can see the Bellman operator also applies to finite-horizon problems (assuming no t dependence in g and p, or modifying appropriately), where we have $v_t = \mathcal{T}v_{t+1}$. Our proof that the finite-horizon problem converges to the infinite horizon problem then corresponds to saying $\mathcal{T}^n v_\pm \to v$, where $v_\pm$ was the trivial upper or lower bounds on the value function. This can be done more generally, as is shown by the following result.

Theorem 2.1.5. *With discount rate $\rho > 0$, the Bellman operators $\mathcal{T}, \mathcal{T}_u$ are strict contractions (with rate $e^{-\rho}$) under the $\|\cdot\|_\infty$ norm on $\mathbb{R}^{\mathcal{X}}$. Consequently, the Banach fixed point theorem shows that the value iteration sequence converges (exponentially quickly) to the value function v, that is,*

$$\|v_n - v\|_\infty \le e^{-\rho n}\|v_0 - v\|_\infty.$$

Proof. We consider $\mathcal{T}$ first. Consider $w, w'\in\mathbb{R}^{\mathcal{X}}$. By continuity and compactness, we know that there is u^* such that

$$(\mathcal{T}w')(x) = g(x,u^*) + e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u^*)w'(x') = (\mathcal{T}_{u^*}w')(x).$$

Therefore, with this u^*,

$$\begin{aligned}
&(\mathcal{T}w)(x) - (\mathcal{T}w')(x) \\
&= \min_{u\in\mathcal{U}} \big\{\big(\mathcal{T}_u w\big)(x)\big\} - \big(\mathcal{T}_{u^*} w'\big)(x) \\
&\le \big(\mathcal{T}_{u^*} w\big)(x) - \big(\mathcal{T}_{u^*} w'\big)(x) \\
&\le \Big\{g(x,u^*) + e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u^*)w(x')\Big\} - \Big\{g(x,u^*) + e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u^*)w'(x')\Big\} \\
&= e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u^*)\big(w(x') - w'(x')\big) \\
&\le e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u^*)\|w - w'\|_\infty \\
&= e^{-\rho}\|w - w'\|_\infty.
\end{aligned}$$

Interchanging w and w' we see that

$$\|\mathcal{T}w - \mathcal{T}w'\|_\infty \le e^{-\rho}\|w - w'\|_\infty$$

so $\mathcal{T}$ is an L^∞ contraction, with rate $e^{-\rho} < 1$. The argument for $\mathcal{T}_u$ is analogous (and simpler). The convergence result follows from Banach's fixed point theorem. □

There is another approach to showing convergence, which looks at monotonicity properties of the value iteration. While this is not so critical here (as we've shown we have a contraction), it is also interesting to see that the Bellman operator always improves the value, in a particular sense.

Theorem 2.1.6. *The Bellman operators are pointwise monotone, in the sense that if $w \le w'$, then $\mathcal{T}w \le \mathcal{T}w'$ (both inequalities being pointwise). In particular, if v_0 is such that $\mathcal{T}v_0 \le v_0$ (for example, if $v_0(x) = \frac{1}{1-e^{-\rho}}\max_{x,u} g(x,u)$), then value iteration decreases monotonically to v.*

Proof. Using exactly the same logic and notation as in the previous proof, we have

$$(\mathcal{T}w)(x) - (\mathcal{T}w')(x) \le e^{-\rho}\sum_{x'\in\mathcal{X}} p(x';x,u^*)\big(w(x') - w'(x')\big) \le 0$$

(and similarly for $\mathcal{T}_u$). This establishes the monotonicity of $\mathcal{T}$. The monotone convergence then follows by induction, as $v_1 = \mathcal{T}v_0 \le v_0$ implies $v_2 = \mathcal{T}v_1 \le \mathcal{T}v_0 = v_1$. □

Using this monotonicity, we can prove an elegant result on the *controls* which arise from value iteration. The key result is the following policy error bound:

Theorem 2.1.7. *For any $v' \in \mathbb{R}^{\mathcal{X}}$, let*

$$\tilde{u}(x) \in \arg\min_u \big\{(\mathcal{T}_u v')(x)\big\}.$$

Suppose $v \in \mathbb{R}^{\mathcal{X}}$ is the optimal value. Then $\tilde{u}$ has value

$$J(x,\tilde{u}) \le v(x) + \frac{2e^{-\rho}}{1-e^{-\rho}}\|v - v'\|_\infty$$

Proof. Write $\varepsilon = \|v - v'\|_\infty$, and $\tilde{v}(x) = J(x, \tilde{u})$ for the true value associated with the control $\tilde{u}$, so we know $\tilde{v} = \mathcal{T}_{\tilde{u}}\tilde{v}$. As v is optimal, we know $v \leq \tilde{v}$.

By definition, it's easy to check that $\mathcal{T}(w + a) = (\mathcal{T}w) + e^{-\rho}a$ for all $w \in \mathbb{R}^{\mathcal{X}}$ and $a \in \mathbb{R}$, and similarly for $\mathcal{T}_u$. As $\mathcal{T}$ is monotone, we know $v = \mathcal{T}v \geq \mathcal{T}(v' - \varepsilon) = (\mathcal{T}v') - e^{-\rho}\varepsilon$.

Therefore,

$$0 \leq \tilde{v} - v = \mathcal{T}_{\tilde{u}}\tilde{v} - \mathcal{T}v \leq \mathcal{T}_{\tilde{u}}\tilde{v} - \mathcal{T}v' + e^{-\rho}\varepsilon.$$

From the definition of $\tilde{u}$, we know $\mathcal{T}v' = \mathcal{T}_{\tilde{u}}v'$, and expanding the definition gives

$$\mathcal{T}_{\tilde{u}}\tilde{v} - \mathcal{T}_{\tilde{u}}v' = e^{-\rho}\sum_{x'} p(x'; x, \tilde{u})\big(\tilde{v}(x') - v'(x')\big) \leq e^{-\rho}\max_{x'}\big\{\tilde{v}(x') - v'(x')\big\} \leq e^{-\rho}\max_{x'}\big\{|\tilde{v}(x') - v(x')|\big\} + e^{-\rho}\varepsilon.$$

Combining these inequalities,

$$0 \leq \tilde{v} - v \leq e^{-\rho}\max_{x'}\big\{|\tilde{v}(x') - v(x')|\big\} + 2e^{-\rho}\varepsilon$$

and hence

$$\|v - \tilde{v}\|_\infty \leq e^{-\rho}\|v - \tilde{v}\|_\infty + 2e^{-\rho}\varepsilon.$$

The result follows by rearrangement. □

Corollary 2.1.8. *If v_n is the sequence obtained by value iteration[2], and we define*

$$u_n(x) \in \arg\min_u \big\{(\mathcal{T}_u v_n)(x)\big\},$$

then $J(x, u_n) \to v(x)$ (with geometric convergence), and for any subsequence of policies which converges, the limit u^* is an optimal policy.

Proof. From Theorem 2.1.7, we have the geometric convergence

$$0 \leq J(x, u_n) - v(x) \leq \frac{2e^{-\rho}}{1 - e^{-\rho}}\|v_n - v\|_\infty \leq \frac{2e^{-\rho}}{1 - e^{-\rho}}e^{-\rho n}\|v_0 - v\|_\infty \to 0.$$

As our policies lie in a compact set, there is a subsequence of u_n which is convergent. Writing

$$J(x, u) = \sum_{t=0}^{\infty}\sum_{x' \in \mathcal{X}} e^{-\rho t}p(x'; x, u)g(x, u(x)),$$

we use dominated convergence (in time) to show that along any convergent subsequence we have

$$v(x) = \lim_{n\to\infty} J(x, u_n) = J(x, \lim_{n\to\infty} u_n),$$

so $u^* = \lim u_n$ is an optimal policy. □

[2]The same result holds for any other scheme such that our value function approximations converge (geometrically) in $\|\cdot\|_\infty$ to the true value function.

2.2 Policy iteration

The next numerical method we will consider is related to value iteration, and works under the same assumptions. Essentially, the problem with value iteration is that it combines the estimation of the value of a policy and the optimization of the policy into a single step. Numerically, we often find that estimating the value accurately requires us to work over many steps, and requiring us to compute the minimum in every step becomes expensive. Policy iteration aims to separate these two operations, by allowing us to accurately compute the value using multiple (easy) steps, and only infrequently computing the minimum.

Definition 2.2.1. *Policy iteration consists of two alternating steps:*

- *Evaluation, where for a given policy u_n we compute its value function $v_n = J(\cdot, u_n)$.*
- *Improvement, where for a given value function v_n, we compute a new policy*

$$u_{n+1} \in \arg\min_u \mathcal{T}_u v_n.$$

This process can be initialized by specifying either u_0 or v_0, and proceeding iteratively.

The key difference between policy and value iteration is in the evaluation step, where we compute the true value associated with a given policy. Observe that this satisfies

$$v_n(x) = g(x, u_n(x)) + e^{-\rho} \sum_{x'} p(x'; x, u_n(x)) v_n(x'),$$

which is a finite dimensional linear system. Indeed, associating $v(x_i)$ with the component $\mathbf{v}^{(i)}$ of a vector in $\mathbb{R}^{|\mathcal{X}|}$, and similarly $\mathbf{g}(u)^{(i)} = g(x_i, u(x_i))$ and writing the matrix $P(u)_{ij} = p(x_j; x_i, u(x_i))$, we have the matrix-vector equation

$$\mathbf{v}_n = \mathbf{g}(u_n) + e^{-\rho} P(u_n) \mathbf{v}_n = \left(I - e^{-\rho} P(u_n)\right)^{-1} \mathbf{g}(u_n).$$

(With this notation, the fact $P(u_n)$ has all eigenvalues having real part below 1 is enough to prove this equation is well-posed (as we have previously argued), and this follows from the Perron–Frobenius theorem as P is a stochastic matrix). As this is a linear equation, it can be solved in at most $O(|\mathcal{X}|^3)$ operations.

Therefore, in classical policy iteration, we assume that we can solve the evaluation step perfectly in reasonable time, and hope that relatively few policy improvement steps will be sufficient to achieve convergence.

Lemma 2.2.2 (Improvement lemma)**.** *The sequence of values constructed through policy iteration satisifes $v_n \geq \mathcal{T} v_n \geq v_{n+1}$.*

Proof. By construction, we know that $\mathcal{T} v_n = \mathcal{T}_{u_{n+1}} v_n$ (improvement) and $v_n = \mathcal{T}_{u_n} v_n$ (evaluation). It follows that

$$v_n = \mathcal{T}_{u_n} v_n \geq \mathcal{T} v_n = \mathcal{T}_{u_{n+1}} v_n.$$

As $\mathcal{T}_{u_{n+1}}$ is monotone, we can apply $\mathcal{T}_{u_{n+1}}$ to both sides to see

$$\mathcal{T}_{u_{n+1}} v_n \geq \left(\mathcal{T}_{u_{n+1}}\right)^2 v_n.$$

and by induction,

$$v_n \geq \mathcal{T}_{u_{n+1}} v_n \geq \cdots \geq \left(\mathcal{T}_{u_{n+1}}\right)^k v_n$$

for all $k \geq 0$. As $\mathcal{T}_{u_{n+1}}$ is a contraction, the right hand side converges to the fixed point $v_{n+1} = \mathcal{T}_{u_{n+1}} v_{n+1}$, giving the desired result. □

Theorem 2.2.3. *The value functions constructed via policy iteration converge geometrically to the true value function (and hence the policies converge, in the sense that Corollary 2.1.8 applies).*

Proof. Using our lemma, with v the true value function,

$$v \leq v_n \leq \mathcal{T} v_{n-1} \leq \cdots \leq \mathcal{T}^n v_0.$$

Therefore

$$0 \leq v_n - v \leq \mathcal{T}^n v_0 - v = \mathcal{T}^n v_0 - \mathcal{T}^n v$$

and we see that, as $\mathcal{T}$ is a contraction with rate $e^{-\rho}$,

$$\|v_n - v\|_\infty \leq \|\mathcal{T}^n v_0 - \mathcal{T}^n v\|_\infty \leq e^{-\rho n} \|v_0 - v\|_\infty.$$

□

Remark 2.2.4. If our set $\mathcal{U}$ is finite, then the fact our system is improving at every step shows that policy iteration will terminate after finitely many steps, that is, the optimal policy will be found. This is because there are only finitely many policies better than u_0, and our choice of policy is improving at every step. In fact, one can show that this occurs after polynomially many steps, a result which is due to Yingyu Ye [10].

2.2.1 Approximate policy iteration

Various extensions of policy iteration are often considered. For example, we can replace the evaluation step with an approximation, for example taking $v_n = (\mathcal{T}_{u_n})^k v_{n-1}$ for some $k < \infty$. This effectively interpolates between value iteration (where $k = 1$) and policy iteration (where $k = \infty$). A small variation of our results above shows us that v_n converges to v geometrically fast, just as in the two extreme cases.

Another common idea is to replace the policy improvement step with a simpler calculation, where we simply try and find a u_{n+1} such that

$$\mathcal{T}_{u_{n+1}} v_n \leq v_n,$$

with the requirement that $\mathcal{T}_{u_{n+1}} v_n = v_n$ iff $v_n = v$. Looking back through our results, we can see that this is enough to guarantee that $v_n \geq v_{n+1}$, and we know that $v_n \geq v$, but proving $v_n \to v$ is more difficult.

2.3 Q-learning

Suppose we don't want to compute expected values, or don't know the probability law directly. Then value and policy iteration is still possible, if we have access to simulations or observations from the controlled

system. This is the fundamental trick that is sometimes called *model free* reinforcement learning, as we will avoid explicitly modelling the transition function p and cost g (but there is still a mathematical model behind everything we do, as we need to have state variables, repeated observations, conditional independence, etc...).

Remark 2.3.1. So far, our analysis has assumed that the cost g is a (known) function of (x_t, u). However, it is straightforward to include random costs in our setup, which can depend on both X_{t+1} and on some auxiliary randomness $\tilde{\omega}_{t+1}$. In this case, we can define $G(\tilde{\omega}_{t+1}, x_t, X_{t+1}, u_t)$ to be the (random) cost, and then set $g(x_t, u_t) = \mathbb{E}_u[G(\tilde{\omega}_{t+1}, x_t, X_{t+1}, u_t)|u = u_t, X_t = x_t]$. Using the tower law, we consider an agent who wishes to minimize

$$\mathbb{E}^U\Big[\sum_t e^{-\rho t} G(\tilde{\omega}_{t+1}, X_t, X_{t+1}, U_t)\Big] = \mathbb{E}^U\Big[\sum_t e^{-\rho t} g(X_t, U_t)\Big],$$

(or similarly for finite horizon problems), the first expectation being over both X and $\tilde{\omega}$. The second representation shows that this is formally the same problem as we have already considered. This is a minor change to the theory, but in practice allows us to consider situations where we do not know the distribution of the costs nor the probability transitions.

In our motivating example (Example 1.2.2), this tweak allows us to have a cost/reward depending on the number of subscribers who join in the next period, rather than only existing subscribers, and random advertising costs, which is a modest increase in modelling flexibility.

Q-learning attempts to build an approximation similar to value iteration, but using observations of the random transitions and costs. The fundamental object is the Q function, which is defined as the map $Q : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ which is a fixed point of the iteration

$$\begin{aligned} Q(x,u) &= g(x,u) + e^{-\rho} \sum_{x'} p(x'; x, u) \min_{u'} Q(x', u') \\ &= \mathcal{T}_u\{\min_{u'} Q(\cdot, u')\}(x) \\ &= \mathbb{E}^u\Big[G(x, X_{t+1}, u) + e^{-\rho} \min_{u'} Q(X_{t+1}, u')\Big| X_t = x\Big]. \end{aligned}$$

It is fairly clear that the Q function is closely related to the value function, in particular $v(\cdot) = \min_{u'} Q(\cdot, u')$ (which immediately implies the Q function exists). The advantage of knowing Q instead of v is that it also includes the expectation over the (random) costs G and the next transition, which may be unknown, and so allows us to find good policies based only on optimizing Q.

We will attempt to learn Q based on observing trajectories of X and G. For any $\alpha \in \mathbb{R}$, we can write

$$Q(x,u) = Q(x,u) + \alpha \mathbb{E}^u\Big[G(x, X_{t+1}, u) + e^{-\rho} \min_{u'} Q(X_{t+1}, u') - Q(x,u)\Big| X_t = x\Big].$$

Now suppose we have access to some collection of the variables $\{X_n, X_{n+}, U_n, G_n\}_{n\in\mathbb{N}}$, where

- $X_n \in \mathcal{X}$ and $U_n \in \mathcal{U}$ (sampled with an arbitrary distribution),
- $X_{n+} \sim p(\cdot; X_n, U_n)$ and $G_n \sim G(X_n, X_{n+}, U_n)$ (with an abuse of notation) and
- (X_{n+}, G_n) is independent of $\{(X_m, X_{m+}, U_m, G_m)\}_{m<n}$ given (X_n, U_n).

For notational convenience, we define the filtration $\mathcal{F}_n = \sigma(X_n, U_n, \{(X_m, X_{m+}, U_m, G_m)\}_{m<n})$.

We note that these variables could be from simulating or observing the controlled system, with some control rule U_n chosen randomly (dependent on X_n) and $X_{n+} = X_{n+1}$, or from another means. In particular, we could use different control rules U_n for each n.

Definition 2.3.2. *With $\alpha_n \in [0,1)$ an $\{\mathcal{F}_n\}_{n\geq 0}$-adapted learning rate process, the Q-learning iteration is defined by*

$$Q_{n+1}(x,u) = \begin{cases} Q_n(x,u) + \alpha_n\Big[G_n + e^{-\rho}\min_{u'} Q_n(X_{n+}, u') - Q_n(x,u)\Big] & \text{if } X_n = x, U_n = u, \\ Q_n(x,u) & \text{otherwise.} \end{cases}$$

We then have the following convergence result.

Theorem 2.3.3. *Consider a control problem with finitely many actions. Let $Q_0 : \mathcal{X}\times\mathcal{U} \to \mathbb{R}$ be arbitrary, and Q_n be the random sequence of functions defined by the Q-learning iteration. If*

$$\mathbb{E}\big[G(x, X_+, u)^2 \big| X_+ \sim p(\cdot; x, u)\big] < \infty \text{ for all } (x,u)$$

and α_n satisfies the Robbins–Monro-type condition

$$\sum_{n\in\mathbb{N}} \alpha_n 1_{\{(X_n,U_n)=(x,u)\}} = \infty, \qquad \sum_{n\in\mathbb{N}} \alpha_n^2 1_{\{(X_n,U_n)=(x,u)\}} < \infty$$

for all x, u, then $Q_n(x,u) \to Q(x,u)$ almost surely as $n \to \infty$.

In order to prove this, we will make use of the following stochastic approximation result, the (rather fiddly) proof of which is in Appendix A.1.2

Lemma 2.3.4. *Consider an adapted random process $\{Y_n\}_{n\in\mathbb{N}}$ with values in $\mathbb{R}^m$, with dynamics (for each i an index in $\{1, ..., m\}$)*

$$Y_{n+1}(i) = \big(1 - \alpha_n(i)\big)Y_n(i) + \alpha_n(i)Z_{n+1}(i)$$

where, for all i,

- $\alpha_n(i) \in [0,1]$, $\sum_{n\in\mathbb{N}} \alpha_n(i) = \infty$, $\sum_{n\in\mathbb{N}} \alpha_n^2(i) < \infty$,
- $|\mathbb{E}[Z_{n+1}(i)|\mathcal{F}_n]| \leq \gamma\|Y_n\|_\infty$, *with* $\gamma < 1$,
- $\mathbb{V}[Z_{n+1}(i)|\mathcal{F}_n] \leq c(1 + \|Y_n\|_\infty^2)$ *for* $c > 0$.

Then $\|Y_n\|_\infty \to 0$ a.s. as $n \to \infty$.

Proof of Theorem 2.3.3. Take Q_n to be the sequence generated by Q-learning, and Q to be the true Q-function. Define

$$\begin{aligned} Y_n(x,u) &= Q_n(x,u) - Q(x,u), \\ \alpha_n(x,u) &= \alpha_n 1_{\{(X_n,U_n)=(x,u)\}}, \\ Z_{n+1}(x,u) &= \Big[G_n + e^{-\rho}\min_{u'} Q_n(X_{n+}, u') - Q(X_n,u)\Big]1_{\{(X_n,U_n)=(x,u)\}}. \end{aligned}$$

With this notation, simple rearrangement shows that $Y_n(x,u)$ satisfies the dynamics

$$Y_{n+1}(x,u) = \big(1-\alpha_n(x,u)\big)Y_n(x,u) + \alpha_n(x,u)Z_{n+1}(x,u).$$

From the definition of Q, we know that

$$0 = \mathbb{E}^u\Big[G_n + e^{-\rho}\min_{u'} Q(X_{n+},u') - Q(x,u)\Big|\mathcal{F}_n\Big]1_{\{(X_n,U_n)=(x,u)\}},$$

and thus

$$\begin{aligned}\mathbb{E}[Z_{n+1}(x,u)|\mathcal{F}_n] &= \mathbb{E}\Big[G_n + e^{-\rho}\min_{u'} Q_n(X_{n+},u') - Q(X_n,u)\Big|\mathcal{F}_n\Big]1_{\{(X_n,U_n)=(x,u)\}}\\ &= e^{-\rho}\mathbb{E}\Big[\min_{u'} Q_n(X_{n+},u') - \min_{u'} Q(X_{n+},u')\Big|\mathcal{F}_n\Big]1_{\{(X_n,U_n)=(x,u)\}}.\end{aligned}$$

Taking an absolute value, we have the bound

$$\big|\mathbb{E}[Z_{n+1}(x,u)|\mathcal{F}_n]\big| \le e^{-\rho}\max_{(x,u)}\big|Q_n(x,u') - Q(x,u')\big| = e^{-\rho}\|Y_n\|_\infty.$$

We also know (using $(a+b)^2 \le 2a^2+2b^2$) that

$$(Z_{n+1}(x,u))^2 \le 2\Big(G_n - Q(x,u) + e^{-\rho}Q(X_{n+},u^*)\Big)^2 + 2e^{-\rho}\Big(Q_n(X_{n+},u^*) - Q(X_{n+},u^*)\Big)^2,$$

where $u^* \in \arg\min_u Q_n(X_{n+},u)$, and so

$$\mathbb{E}[(Z_{n+1}(x,u))^2|\mathcal{F}_n] \le c(1+\|Y_n\|_\infty^2)$$

for some $c>0$. Combining with our previous bound, we get the desired growth bound on the variance. As $e^{-\rho}<1$, applying Lemma 2.3.4 we conclude that $\|Y_n\|_\infty \to 0$ a.s., as desired. □

There are many variations of Q-learning, mainly depending on how the policies u_n are chosen, and whether $X_{n+1} = X_{n+}$. If $X_{n+1} = X_{n+}$, then we can see this as an *online* method, where we use our controlled system in order to learn the control.

When choosing u_n, we naturally want to focus our attention on policies which are likely to be optimal, but the Robbins–Monro condition shows that we need to try all state-policy pairs infinitely often (as well as tuning the learning rate accordingly). This trade-off is related to exploration-vs-exploitation (but this usually refers to trying to optimize the values we realize while running an online control method). In general the Q-learning algorithm is *off-policy*, in that the policy we use to learn Q does not need to be an approximation of the optimal policy.

We can also tweak the algorithm to prevent us having to perfectly compute the minimum in the Q-learning iteration (which is important for large action spaces), provided we are eventually sufficiently accurate.

Remark 2.3.5. What's described above is *classical* Q-learning. In recent years, there have been major advances in this space, mainly through using deep neural networks (or similar tools) as function approximators for the Q function. The basic Q learning iteration can then be rewritten as an iterative regression/function approximation problem, and hence an approximate method can be implemented.

Adding this approximation error makes the analysis significantly more complicated (but it's not too bad in this setting of finitely many states and actions), and is an active area of research. In infinite states, one of the key problems boils down to the choice of metric – we have used the $\|\cdot\|_\infty$ metric extensively above, but it is not so easy to prove approximation results in L^∞ on infinite spaces, where our usual approximation theory is in L^2. This distinction causes many headaches.

2.3.1 SARSA

A mild variation of the previous arguments can also be used to study the cost associated with a specific randomized control rule $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{U})$ (when $\mathcal{U}$ is finite, more on this later). This is commonly known as SARSA (as the iteration depends on $S_t, A_t, R_t, S_{t+}, A_{t+}$, with the notation S for the state, A for the action, and R for the reward). In our notation, the (optimized) Q-learning iteration is replaced by the sequence

$$Q^{\pi}_{n+1}(x,u) = Q^{\pi}_n(x,u) + \alpha_n \Big[G_n + e^{-\rho} Q^{\pi}_n(X_{n+}, U_{n+}) - Q^{\pi}_n(x,u) \Big] 1_{\{X_n = x, U_n = u\}}$$

where X_n, U_n are sampled arbitrarily, $X_{n+} \sim p(\cdot; X_n, U_n)$ and $U_{n+} \sim \pi(X_{n+})$.

It is then straightforward to use the same logic as we used for Q-learning to show (provided all state-action pairs are visited infinitely often, and the learning rate α satisfies the Robbins–Monro-type condition) that Q^{π}_n converges to the function

$$Q^{\pi}(x,u) = \mathbb{E}\Big[\sum_t e^{-\rho t} G(\tilde{\omega}_{t+1}, X_t, X_{t+1}, U_t) \Big| X_0 = x, U_0 = u, U_t \sim \pi(X_t), X_{t+1} \sim p(\cdot; X_t, U_t) \Big].$$

From this, we see that the cost function $J(x,\pi) = \mathbb{E}[Q^{\pi}(x,u) | u \sim \pi]$. This can be used as the basis for policy iteration methods, in particular by observing that a policy improvement step is given by $\tilde{u}(x) := \arg\min_u Q^{\pi}(x,u)$, which does not involve computing any additional expected values.

2.4 Aside: Entropy-regularized control

One common technique to improve the convergence of these methods, is to try and regularize the selection of the control. To do this, the usual technique is to think of having a finite number of controls $\{u_1, u_2, ..., u_m\} = \mathcal{U}_0$, but then working with randomized policies: we think of the agent as having access to an external source of randomness, which allows them to choose a control following a probability distribution on $\mathcal{U}_0$, independently at every point in time. Formally, this corresponds to $\mathcal{U} = \mathcal{P}(\mathcal{U}_0)$, and we extend our costs and transition law accordingly: for $\pi \in \mathcal{U} = \mathcal{P}(\mathcal{U}_0)$, we know $\pi = [\pi_1, \pi_2, ..., \pi_m]$ (where the probability our agent chooses u_i is π_i), and define

$$g_0(x,\pi) := \sum_{i=1}^m \pi_i g(x,u_i) = \mathbb{E}[G(\tilde{\omega}, x, X_{t+1}, u) | X_t = x, U_t \sim \pi]$$

and

$$p(x'; x, \pi) := \sum_{i=1}^m \pi_i p(x'; x, u_i).$$

If we simply solved the control problem with this g_0, we would achieve the same value as for the intial problem (as any control can be represented as a trivial distribution).

There are a couple of ways to make the probabilities nontrivial – the simplest is to use an ε-greedy method, which sets $\pi(u|x) = \frac{\varepsilon}{|\mathcal{U}|-1} + (1-\varepsilon) 1_{u = u^*(x)}$, where u^* is an optimal action, for some $\varepsilon > 0$. This is easy to implement, and we can easily choose ε to give desired properties (such as the Robbins–Monro condition and near-optimality of the strategy), but does not have good smoothness properties.

A common (and more interesting) alternative is to define a new cost which encourages randomization:

$$g_\lambda(x,\pi) := g_0(x,\pi) - \lambda H(\pi)$$

where $H(\pi) = -\sum_i \pi_i \log(\pi_i)$ is the Shannon entropy of the random control. For $\lambda > 0$, this encourages our control to randomize, and has the effect of smoothing out our problem.

We will use the following lemma.

Lemma 2.4.1. *Take any $c \in \mathbb{R}^m$ and consider minimizing $\{\langle c,\pi\rangle - \lambda H(\pi)\}$ for $\lambda > 0$, among probability vectors π. Then the (unique) minimum is achieved by the* Gibbs measure $\pi_i \propto e^{-c_i/\lambda}$*, and is given by the* log-sum-exp *function*

$$\min_\pi \{\langle c,\pi\rangle - \lambda H(\pi)\} = -\lambda \log\Big[\sum_j \exp\Big(\frac{-c_j}{\lambda}\Big)\Big] \geq \min_i c_i - \lambda\log(m).$$

Proof. As log has infinite slope near zero, and is convex, any local optimum will occur at an interior point. Observing that $\langle \pi, \mathbf{1}\rangle = \sum_i \pi_i = 1$ (for $\mathbf{1}$ a vector of 1s), we have the Lagrangian

$$\langle c,\pi\rangle + \lambda \sum_i \pi_i \log(\pi_i) + \eta(\langle \pi,\mathbf{1}\rangle - 1).$$

Differentiating with respect to π_i, we have the first order condition

$$0 = c_i + \lambda\log(\pi_i) + \lambda + \eta.$$

Rearranging gives $\pi_i = e^{-(c_i+\lambda+\eta)/\lambda)} \propto e^{-c_i/\lambda}$, with η chosen to guarantee $\sum_i \pi_i = 1$ (that is, $\lambda + \eta = \lambda\log(\sum_j e^{-c_j/\lambda})$). We can then compute

$$\lambda H(\pi) = -\lambda\sum_i \pi_i\log(\pi_i) = \sum_i \pi_i\Big(c_i + \lambda\log\Big(\sum_j e^{-c_j/\lambda}\Big)\Big) = \Big(\sum_i \pi_i c_i\Big) + \lambda\log\Big(\sum_j e^{-c_j/\lambda}\Big).$$

Substitution gives the explicit minimizer, which we see is unique.

We also know that $H(\pi) \leq \log(m)$ (as this is a general bound on the Shannon entropy, which is maximized by a uniform distribution), and hence the final inequality follows. □

Remark 2.4.2. This approach allows us to avoid explicitly optimizing our Q function, as is needed at each step of value iteration. It also stabilizes the choice of policies, which ensures that we have convergence to an optimizer, even if policies are not uniquely determined. For simplicity, in discrete time, we can also take the limiting π^*, sample a rule $u^*(x) \sim \pi^*(x)$ once for each x, and verify that this will be an optimal (deterministic) policy for our problem. In continuous time, however, this becomes much more delicate.

Proposition 2.4.3. *Let v be the value function of the unregularized problem, and v^λ the value of the regularized problem. Then*

$$v^\lambda \leq v \leq v^\lambda + \lambda\frac{\log(m)}{1-e^{-\rho}},$$

and the (unique) optimal policy for the regularized problem is given by

$$\pi(u|x) = \exp\Big(\frac{v^\lambda(x) - g_0(x,u) - e^{-\rho}\sum_{x'} p(x';x,u)v^\lambda(x')}{\lambda}\Big) = \exp\Big(\frac{v^\lambda(x) - Q^\lambda(x,u)}{\lambda}\Big) \propto e^{-Q^\lambda(x,u)/\lambda},$$

where

$$Q^\lambda(x,u) = g(x,u) + e^{-\rho}\sum_{x'} p(x';x,u)v^\lambda(x')$$

is a Q function associated with the value function v^λ.

Proof. As $g_\lambda = g_0 - \lambda H(\pi)$, we know that $g_\lambda \leq g_0 = g_\lambda + \lambda H(\pi) \leq g_\lambda + \lambda \log(m)$. Solving for the value function using each of these terms as the cost, we get

$$v^\lambda \leq v \leq v^\lambda + \lambda \frac{\log(m)}{1 - e^{-\rho}},$$

as desired.

The Bellman equation satisfied by v^λ is then

$$\begin{aligned} v^\lambda(x) &= \min_\pi \Big\{ g_0(x, \pi^\lambda(x,\cdot)) - \lambda H(\pi^\lambda(x,\cdot)) + e^{-\rho} \sum_{x'} p(x'; x, \pi) v^\lambda(x') \Big\} \\ &= \min_\pi \Big\{ \sum_{u \in \mathcal{U}} \pi^\lambda(x,u) \Big(g(x,u) + e^{-\rho} \sum_{x'} p(x'; x, u) v^\lambda(x') \Big) - \lambda H(\pi(x,\cdot)) \Big\} \\ &= -\lambda \log \sum_{u \in \mathcal{U}} \exp\Big(\frac{-g(x,u) - e^{-\rho} \sum_{x'} p(x'; x, u) v^\lambda(x')}{\lambda} \Big) \\ &= -\lambda \log \sum_{u \in \mathcal{U}} \exp\Big(\frac{-Q^\lambda(x,u)}{\lambda} \Big), \end{aligned}$$

where we use Lemma 2.4.1 in the penultimate line. The optimal strategy is given by

$$\pi^\lambda(u|x) \propto \exp\Big(\frac{-Q^\lambda(x,u)}{\lambda} \Big)$$

and checking the constant of proportionality gives us the stated form. □

Remark 2.4.4. Given this proposition, in a Q-learning context, we can guarantee that all actions are considered infinitely often, by counting $n_t(x) = \#\{s < t : X_s = x\}$ and taking a state-dependent regularizer $\lambda(x) \sim 1/n(x)$ and learning rate $\alpha(x) \sim 1/n(x)$. If we also know that all states are visited infinitely often (which is an assumption on the underlying controlled Markov chain), then online Q-learning is guaranteed to converge to the true Q function when actions are chosen according to the entropy-regularized rule.

2.5 Policy gradients

A common variation of policy iteration is to replace the optimization of the policy with a gradient-based method. This is a major area, and we won't seek to prove any convergence results (as they depend on a wide range of results from optimization theory and probability), but will give an introduction to the core method. We will assume that we have a finite action space, and work with randomized policies (which will allow us to differentiate easily).

We suppose that we have an agent who is using a randomized policy $\pi_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{U}_0) = \mathcal{U}$, parameterized by $\theta \in \mathbb{R}^n$ (for some n). We write $\pi_\theta(u; x)$ for the probability of taking action u when in state x. A classic example (sometimes called 'logits in tabular form'), inspired by the entropy-regularized controls above, is to take $\pi(u; x) \propto e^{\theta(x,u)}$, which is valid for all $\theta \in \mathbb{R}^{|\mathcal{U}_0| \times |\mathcal{X}|}$. In this case,

$$\partial_{\theta(u,x)} \log \pi(u'; x') = \Big[1_{\{u=u'\}} - \frac{e^{\theta(u,x)}}{\sum_{u''} e^{\theta(u'',x)}} \Big] 1_{\{x=x'\}}$$

which simplifies our calculations below. If we use this formulation for an entropy-regularized problem, then we know that $\theta^\lambda(u,x) := \frac{-1}{\lambda}\big(g_0(x,u) + e^{-\rho}\sum_{x'} p(x';x,u)v^\lambda(x')\big)$ represents the true optimal policy, so an interior solution (with finite θ) is optimal.

We are aiming to minimize the cost $J(\pi_\theta) = \mathbb{E}[J(X_0,\pi_\theta)]$ (averaged with respect to a distribution over X_0, which has no real impact, as the optimal strategy will minimize $J(x,\pi_\theta)$ for all x, by dynamic programming), using a gradient method. The idea is to calculate the gradient $\nabla_\theta J|_{\theta_n}$, and then use the gradient descent iteration $\theta_{n+1} = \theta_n - \alpha_n \Sigma \nabla_\theta J|_{\theta_n}$, where α_n is the step size/learning rate, and Σ is a (positive definite) preconditioning matrix. The challenge is to estimate $\nabla_\theta J|_{\theta_n}$ efficiently, based on samples of the controlled system.

In order to do this, we make a few observations: For simplicity, we abuse our previous notation and write $g(x_t, x_{t+1}, u_t)$ for the expected value of $G(\tilde{\omega}_{t+1}, X_t, X_{t+1}, U_t)$ given $X_t = x_t, X_{t+1} = x_{t+1}, U_t = u_t$. If we use the randomized control π_θ, then the probability of a sequence is given by

$$\mathbb{P}\Big(X_t = x_t, U_t = u_t \text{ for } t \le T \Big| X_0 = x_0\Big) = \prod_{t=0}^{T} p(x_{t+1}; x_t, u_t)\pi_\theta(u_t; x_t).$$

In particular, we can write

$$\begin{aligned} J(x_0,\pi_\theta) &= \sum_T \mathbb{E}\Big[e^{-\rho T} G_T \Big| X_0 = x_0, U_t \sim \pi_\theta(X_t)\Big] \\ &= \sum_T \sum_{\{x_t,u_t\}_{t\le T}} \Big[\prod_{t=0}^{T} p(x_{t+1};x_t,u_t)\pi_\theta(u_t;x_t) e^{-\rho T} g(x_T, x_{T+1}, u_T)\Big]. \end{aligned}$$

We now observe that for any positive function f_t,

$$\nabla_\theta\Big(f_t(\theta)\Big) = f_t \frac{\nabla_\theta f_t(\theta)}{f_t(\theta)} = f_t(\theta)\nabla_\theta \log f_t(\theta)$$

and hence, by the product rule,

$$\nabla_\theta\Big(\prod_{t\le T} f_t(\theta)\Big) = \Big(\prod_{t\le T} f_t(\theta)\Big)\Big(\sum_{s\le T}\nabla_\theta \log f_s(\theta)\Big) = \sum_{s\le T}\Big(\Big(\prod_{t\le T} f_t(\theta)\Big)\nabla_\theta \log f_s(\theta)\Big).$$

Therefore, by differentiating through the sum for $J(x_0,\pi_\theta)$, we obtain

$$\begin{aligned} \nabla_\theta J(x_0,\pi_\theta) &= \sum_T \sum_{\{x_t,u_t\}_{t\le T}} \sum_{s\le T} \Big[\Big(\prod_{t=0}^{T} p(x_{t+1};x_t,u_t)\pi_\theta(u_t|x_t)\Big)\Big(\nabla_\theta \log \pi_\theta(u_s;x_s)\Big) e^{-\rho T} g(x_T,x_{T+1},u_T)\Big] \\ &= \sum_T \sum_{s\le T} \mathbb{E}\Big[e^{-\rho T}\Big(\nabla_\theta \log \pi_\theta(U_s;X_s)\Big) G_T \Big| X_0 = x, U_t \sim \pi_\theta(X_t)\Big] \\ &= \sum_T \mathbb{E}\Big[e^{-\rho T}\Big(\sum_{s\le T}\nabla_\theta \log \pi_\theta(U_s;X_s)\Big) G_T \Big| X_0 = x, U_t \sim \pi_\theta(X_t)\Big]. \end{aligned}$$

This is sometimes called the 'fundamental lemma of policy gradients', as it allows us to estimate the gradient of J using simulations of trajectories, by multiplying our observed costs with the gradients of the log probabilities of actions, and taking a weighted average.

We can also rearrange our expression by changing the order of summation (between s and T), to give the representations (with all expectations conditional on $U_t \sim \pi_\theta(X_t)$)

$$\begin{aligned}
\nabla_\theta J(x_0, \pi_\theta) &= \sum_{T<\infty} \sum_{s \le T} e^{-\rho T} \mathbb{E}\Big[G_T \nabla_\theta \log \pi_\theta(U_s; X_s) \Big| X_0 = x \Big] \\
&= \sum_{s<\infty} \mathbb{E}\Big[\Big(\sum_{T \ge s} e^{-\rho T} G_T \Big) \nabla_\theta \log \pi_\theta(U_s | X_s) \Big| X_0 = x \Big] \\
&= \sum_{s<\infty} \mathbb{E}\Big[\mathbb{E}\Big[\sum_{T \ge s} e^{-\rho T} G_T | X_s = x_s \Big] \nabla_\theta \log \pi_\theta(U_s; X_s) \Big| X_0 = x \Big] \\
&= \sum_{s<\infty} \mathbb{E}\Big[e^{-\rho s} \mathbb{E}\Big[\sum_{T \ge s} e^{-\rho (T-s)} G_T | X_s = x_s \Big] \nabla_\theta \log \pi_\theta(U_s; X_s) \Big| X_0 = x \Big] \\
&= \sum_{s<\infty} \mathbb{E}\Big[e^{-\rho s} J(X_s, \pi_\theta) \nabla_\theta \log \pi_\theta(U_s; X_s) \Big| X_0 = x \Big].
\end{aligned}$$

Particularly when G is deterministic given X_t, U_t, this expression is useful for analysis, as it expresses the gradient in terms of the discounted occupation density under the stated control.

Given these expressions, we now have a fairly simple recipe for a policy gradient method:

- Simulate long trajectories $\{X^j, U^j, G^j\}_{j=1,\dots,N}$ of the controlled system using the control π_θ, with a variety of starting points x_0.
- Estimate the corresponding average gradients through

$$\widehat{\nabla_\theta J|_{\theta_n}} = \frac{1}{N} \sum_{j=1}^{N} \sum_{T} e^{-\rho T} G_T^j \Big(\sum_{s \le T} \nabla_\theta \log \pi_\theta(U_s^j; X_s^j)|_{\theta=\theta_n} \Big).$$

- Increment θ using the step $\theta_{n+1} = \theta_n - \alpha_n \Sigma \widehat{\nabla_\theta J|_{\theta_n}}$

If all goes well, then this should (at least with high probability), lead to an improvement in the values associated with the policy, and ultimately to convergence to an optimal policy (when our parameterization is rich enough). We can also simulate finite length paths, and replace the value after that point with an estimate of $e^{-\rho T} J(x_T, \pi_\theta)$, which corresponds to a version of value iteration.

The details under which this process converges to the optimal solution are somewhat subtle, see for example Bhandari and Russo [1] for recent results in this area.

There are many variations of this basic approach that are used in practice, by varying the exact calculation which is done in order to reduce variance, adjusting the step size dynamically, adding tweaks to the steps to improve the geometry (e.g. choosing a good preconditioning matrix) etc...

Chapter 3

Continuous Deterministic Control

We will now leave behind the discrete time-space theory that we have been considering, and move into a continuous time setting. In this chapter we will focus on understanding deterministic problems – these have some significant differences to stochastic problems, which we will consider in the next part of the course.

We will not focus on numerical methods for these problems. In practice, many of the equations we consider can be seen as 'standard' PDEs (for which numerical methods are well-known), or can be approximated by discretization (or function approximation), and hence the RL methods we have just seen can be applied. In fact, discretization of the state space is just a finite-difference approximation scheme for the PDE, so this gives a very close connection between our continuous problems and their discretized versions.

3.1 Notation and problem formulation

We suppose we have a state process X, taking values in $\mathcal{X} \subseteq \mathbb{R}^n$, which satisfies a controlled explicit inhomogeneous first-order ODE:

$$\dot{X}_t^U := \frac{\mathrm{d}X_t^U}{\mathrm{d}t} = f(t, X_t^U, U_t), \tag{3.1.1}$$

with initial condition $X_0^U = x_0$, where U_t is a control process to be determined, taking values in a topological space[1] $\mathcal{U}$, and $f : \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}^n$. We consider this equation over the time domain $\mathbb{T} = [0, T]$ (with $T < \infty$) or $\mathbb{T} = [0, \infty)$. We do not need to make the assumption that $\mathcal{U}$ is compact (but correspondingly won't show that optimizers will always exist).

Of course, as we allow multiple dimensions, the fact that our ODE is first-order is not a particular restriction as we can embed more derivatives in more dimensions in X; in this sense, requiring our equation to be first-order is analogous to requiring X to be a Markov process, as we want (X_t, U_t) to determine the value of $X_{t+\varepsilon}$, without needing further information (such as the derivatives of X at t).

We have an agent who chooses the control U within some class of admissible controls, which needs to be defined in such a way that the state dynamics admit a nice solution. In particular, we usually need

[1] This is needed only so that we can talk about Borel measurability (which requires a notion of an open set). Even this can be relaxed, by just using a measurable space.

the solution to be unique (otherwise we would again need more information in order to solve for X), and sufficiently smooth that we can make sense of the equation above.

Example 3.1.1. Consider the problem of minimizing the value of $\int_0^1 |X_t^U| \mathrm{d}t$, given controls in the set $\mathcal{U} = \{1, -1\}$, and the one-dimensional dynamics $\dot{X}_t = U_t$. If we require our state dynamics to have a C^1-smooth solution, there are no non-trivial admissible controls! Furthermore, even if we don't want smoothness, the intuitively optimal control $U_t = -\text{sign}(X_t)$ behaves really weirdly when $X_t = 0$, suggesting that we need to be cautious.

Because of this and similar examples, and because we don't want to put unnecessary restrictions on U, but want to guarantee our equations are meaningful, we often assume that (3.1.1) needs only hold in a weak sense, that is, X^U should satisfy (for all $t \in \mathbb{T}$)

$$X_t^U = x_0 + \int_0^t f(s, X_s^U, U_s) \mathrm{d}s.$$

To allow us to work with our problem more dynamically, we define the more general state process, started at time t in state x:

$$X_{t'}^{t,x,U} = x + \int_t^{t'} f(s, X_s^{t,x,U}, U_s) \mathrm{d}s. \tag{3.1.2}$$

The next ingredient we need in our problem is the cost, which we assume is of the form

$$J(U) = \int_0^T e^{-\rho s} g(s, X_s^U, U_s) \mathrm{d}s + e^{-\rho T} \Phi(X_T^U) \tag{3.1.3}$$

where $T = \infty$ if $\mathbb{T} = [0, \infty)$, in which case we assume $\rho > 0$. We assume $\Phi : \mathcal{X} \to \mathbb{R}$ and $g : \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, and that g is Borel measurable.

As before, we define the rescaled cost-to-go for a control U as

$$J(t, x, U) = \int_t^T e^{-\rho(s-t)} g(s, X_s^{t,x,U}, U_s) \mathrm{d}s + e^{-\rho(T-t)} \Phi(X_T^{t,x,U}).$$

Definition 3.1.2. *For our deterministic problem, we say a control $U : \mathbb{T} \to \mathcal{U}$ is* admissible*, and write $U \in \mathbb{U}$ if U is measurable and* (3.1.2) *admits a unique solution $X^{t,x,U}$, taking values in $\mathcal{X}$, for all $(t, x) \in \mathbb{T} \times \mathcal{X}$.*

Remark 3.1.3. This is a fairly vague definition of admissibility. In many problems we will know that $f(s, x, u)$ is (locally) Lipschitz continuous with respect to x, uniformly with respect to (s, u), is continuous with respect to (s, u). Together with some integrability assumptions, this is enough to guarantee X^U is well-defined for any (Lebesgue) measurable U.

To avoid trivial cases, we assume that there exists at least one admissible control U with $J(U) < \infty$. Furthermore, to avoid problems where the cost can be made infinitely negative, we assume that there is an integrable function $g_* : \mathbb{T} \to \mathbb{R}$ such that $g(t, x, u) \geq g_*(t)$ and $\Phi(x) \geq g_*(T)$, for all x, u, so we have the lower bounds $J(t, x, U) \geq \int_t^T g_*(s) \mathrm{d}s + g_*(T) > -\infty$.

We see that this defines the set of *open loop* controls – that is, general functions of time, rather than feedback functions of time and the current state X (sometimes called *closed loop* controls).

Lemma 3.1.4. *With the above definitions,*

- *the set of admissible controls is closed under pasting, that is, given two admissible controls U, U' the control defined by $U''_s = U_s 1_{s\le t} + U'_s 1_{s>t}$ is also admissible;*
- *for any admissible control, X and J satisfy the flow properties, for $t \le t' \le t''$,*

$$X^{t,x,U}_{t''} = X^{t,x,U}_{t'} + \int_{t'}^{t''} f(s, X^{t,x,U}_s, U_s)\mathrm{d}s = X^{t',X^{t,x,U}_{t'},U}_{t''},$$
$$J(t,x,U) = \int_t^{t'} e^{-\rho(s-t)} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s + e^{-\rho(t'-t)} J(t', X^{t,x,U}_{t'}, U);$$

- *the cost-to-go does not depend on past actions, that is*

$$J(t,x,U) = J\big(t,x,\{1_{s\le t}U'_s + 1_{s>t}U_s\}_{s\ge 0}\big)$$

for all admissible U'.

Proof. The uniqueness of $X^{t,x,U}$ guarantees that $X^{t,x,U}$ satisfies the flow property. By direct calculation, it follows that we have the flow property for J. Using these flow properties, the admissibility of the pasted control is almost immediate. □

3.2 Dynamic programming and the Hamilton–Jacobi equation

As we have seen in discrete time, a key to understanding these problems is the dynamic programming principle. Given the flow properties we have obtained, and the fact that the space of admissible controls is closed under pasting, it is not too difficult to obtain a result in this direction.

Theorem 3.2.1 (Dynamic programming)**.** *The value function $v(t,x) := \inf_{U\in\mathbb{U}} J(t,x,U)$ satisfies the dynamic programming equation*

$$v(t,x) = \inf_{U\in\mathbb{U}} \Big\{ \int_t^{t'} e^{-\rho(s-t)} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s + e^{-\rho(t'-t)} v(t', X^{t,x,U}_{t'}) \Big\}.$$

Proof. Fix t, x. We clearly see that for all $U \in \mathbb{U}$,

$$v(t,x) \le J(t,x,U) = \int_t^{t'} e^{-\rho(s-t)} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s + e^{-\rho(t'-t)} J(t', X^{t,x,U}_{t'}, U).$$

Fix $\varepsilon > 0$, and write $\tilde{v}(t,x) = \inf_{U\in\mathbb{U}} \big\{ \int_t^{t'} e^{-\rho(s-t)} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s + e^{-\rho(t'-t)} v(t', X^{t,x,U}_{t'}) \big\}$. Then there exists $U \in \mathbb{U}$ such that

$$v(t,x) \le J(t,x,U) \le v(t,x) + \varepsilon.$$

Therefore

$$\begin{aligned} v(t,x) + \varepsilon &\ge \int_t^{t'} e^{-\rho(s-t)} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s + e^{-\rho(t'-t)} J(t', X^{t,x,U}_{t'}, U) \\ &\ge \int_t^{t'} e^{-\rho(s-t)} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s + e^{-\rho(t'-t)} v(t', X^{t,x,U}_{t'}) \ge \tilde{v}(t,x). \end{aligned}$$

As ε is arbitrary, we conclude $v \ge \tilde{v}$.

For the converse inequality, choose a $U \in \mathbb{U}$ (which may depend on the fixed values t, x) such that

$$\int_t^{t'} e^{-\rho(s-t)} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + e^{-\rho(t'-t)} v(t', X_{t'}^{t,x,U}) \leq \tilde{v}(t,x) + \varepsilon,$$

and use this to fix the value of $X_{t'}^{t,x,U}$. As before, there exists $U' \in \mathbb{U}$ such that

$$J(t', X_{t'}^{t,x,U}, U') \leq v(t', X_{t'}^{t,x,U}) + \varepsilon$$

and by pasting, we build a control $\tilde{U} = 1_{s \leq t'} U + 1_{s>t} U'$ such that $X_s^{t,x,U} = X_s^{t,x,\tilde{U}}$ for all $s \leq t'$ and $J(t', X_{t'}^{t,x,\tilde{U}}, \tilde{U}) = J(t', X_{t'}^{t,x,U}, U')$. Therefore

$$\begin{aligned} J(t,x,\tilde{U}) &= \int_t^{t'} e^{-\rho(s-t)} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + e^{-\rho(t'-t)} J(t', X_{t'}^{t,x,U}, U') \\ &\leq \int_t^{t'} e^{-\rho(s-t)} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + e^{-\rho(t'-t)} (v(t', X_{t'}^{t,x,U}) + \varepsilon) \\ &\leq \tilde{v}(t,x) + \big(1 + e^{-\rho(t'-t)}\big)\varepsilon. \end{aligned}$$

This implies that

$$v(t,x) \leq J(t,x,\tilde{U}) \leq \tilde{v}(t,x) + \big(1 + e^{-\rho(t'-t)}\big)\varepsilon$$

and therefore, as ε is arbitrary, we conclude $v \leq \tilde{v}$. □

Remark 3.2.2. This definition of the value function guarantees very little regularity – as it involves an uncountable infimum, it's not even clear that v is measurable (but it usually will be!).

Now that we have a dynamic programming equation, the natural thing to do is to try and convert this into a differential form (i.e. a PDE for v), by taking $t' \to t$. The challenge, as we have seen in our example above, is that when we do this, we might run into some serious problems in defining the state variable – just because we have a sequence of admissible strategies with convergent costs does not mean that we can take the limit when defining the state variable.

For this reason, we begin by giving a rather heuristic derivation of the PDE, and then argue that this is the right equation provided the PDE admits sufficiently smooth solutions.

If we assume v is smooth, then we can do a Taylor expansion and write

$$\begin{aligned} v(t', X_{t'}^{t,x,U}) &= v(t,x) + \big(\partial_t v(t,x)\big)(t'-t) + \big\langle \nabla v(t,x), X_{t'}^{t,x,U} - x \big\rangle + o\big(|X_{t'}^{t,x,U} - x| + |t'-t|\big) \\ &= v(t,x) + \big(\partial_t v(t,x)\big)(t'-t) + \big\langle \nabla v(t,x), f(t,x,U_t) \big\rangle (t'-t) + o(t'-t). \end{aligned}$$

We can then rearrange our dynamic programming equation to read

$$\begin{aligned} 0 = \inf_{U \in \mathbb{U}} \Big\{ &\int_t^{t'} e^{-\rho(s-t)} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + e^{-\rho(t'-t)} \big(\partial_t v(t,x)\big)(t'-t) \\ &+ (e^{-\rho(t'-t)} - 1) v(t,x) + e^{-\rho(t'-t)} \big\langle \nabla v(t,x), f(t,x,U_t) \big\rangle (t'-t) + o(t'-t) \Big\}. \end{aligned}$$

As this should hold on every interval $[t, t']$, dividing through by $t'-t$ and taking a limit, we obtain

$$0 = \big(\partial_t v(t,x)\big) - \rho v(t,x) + \inf_{u \in \mathcal{U}} \Big\{ g(s,x,u) + \big\langle \nabla v(t,x), f(t,x,u) \big\rangle \Big\}. \tag{3.2.1}$$

(Note that $\mathbb{U}$ has become $\mathcal{U}$, as only the initial value $u = U_t$ is relevant.) Recalling that $\mathcal{X} \subseteq \mathbb{R}^n$, and defining the *Hamiltonian* to be

$$H : \mathbb{T} \times \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}; \big(t, x, q\big) \mapsto \inf_{u \in \mathcal{U}} \Big\{ g(t, x, u) + \big\langle q, f(t, x, u) \big\rangle \Big\}$$

we can express this PDE in the standard form

$$-\partial_t v = -\rho v + H\Big(t, x, \nabla v\Big).$$

This equation is a form of the classical Hamilton–Jacobi equation in physics[2]. The boundary condition for the PDE varies somewhat – if we have a fixed terminal time T, then we know $v(T, \cdot) = \Phi(\cdot)$. If our problem is on an infinite horizon and g is bounded, then these get replaced with the growth condition $|v| \leq \|g\|_\infty / \rho$ (and it often turns out that this is enough to determine a unique solution to the PDE on $[0, \infty) \times \mathcal{X}$). Similarly if our system stops when x hits some boundary values $x \in \mathcal{X}_b$, then we have $v(\cdot, x) = \Phi(x)$ for all $x \in \mathcal{X}_b$, together with boundedness of v.

Remark 3.2.3. It's helpful to point out that the (unoptimized) Hamiltonian $\tilde{H}(\cdot, \cdot, \nabla V)$ is, in some sense, playing a role similar to what the Q function played in discrete time. To see this, compare the Bellman iteration (for discrete deterministic systems, with $\rho = 0$) written in the form

$$-\Big(v(t+1, x) - v(t, x)\Big) = \min_{u \in \mathcal{U}} \Big\{ \underbrace{g(t, x, u) + v(t+1, f(t, x, u))}_{Q(t,x,u)} - v(t+1, x) \Big\}$$

with the Hamilton–Jacobi equation

$$-\partial_t v = \min_{u \in \mathcal{U}} \Big\{ \underbrace{g(t, x, u) + \big\langle \nabla v, f(t, x, u) \big\rangle}_{\tilde{H}(t,x,\nabla v)} \Big\}.$$

The same connection also holds true in stochastic problems.

We can now obtain the verification step – if we have a smooth solution to the PDE, then it must be the optimal control. We use somewhat restrictive assumptions in this theorem, mainly to allow us to ensure stability of the resulting ODEs, and to allow us to have an intuitively straightforward proof. The verification theorem we will give for stochastic problems supercedes this one, and has weaker assumptions.

Theorem 3.2.4 (Verification theorem)**.** *Consider a control problem with finite terminal time T. In this case, the Hamilton–Jacobi equation with boundary conditions is given, for $v : [0, T] \times \mathbb{R}^n \to \mathbb{R}$, by*

$$-\partial_t v = -\rho v + H\Big(t, x, \nabla v\Big); \qquad v(T, \cdot) = \Phi(\cdot).$$

Suppose, for some constants $K > 0, k \geq 1$, (for all $(t, x_t), (s, x_s) \in \mathbb{T} \times \mathcal{X}$ as appropriate)

- *the Hamilton–Jacobi equation admits a solution w in $C^1\big([0, T] \times \mathbb{R}^n\big)$ satisfying the bound[3],*

$$|\nabla w(t, x_t) - \nabla w(s, x_s)| \leq K\big(|t - s| + \|x_t - x_s\|\big)\big(1 + \|x_t\|^k + \|x_s\|^k\big);$$

  and $\|\nabla w(0, 0)\| \leq K$.

[2]Some authors call it the Hamilton–Jacobi–Bellman equation, but this is arguably the more general version which we will meet in the context of stochastic control.

[3]We impose a growth bound on the changes in ∇w for simplicity, and will see that this can be weakened to assuming a growth bound on w when we consider the Hamilton–Jacobi–Bellman equation.

- *f is K-Lipschitz continuous in (t,x) uniformly in u, that is, for any $u \in \mathcal{U}$,*

$$\|f(t,x_t,u) - f(s,x_s,u)\| \le K\big(|t-s| + \|x_t - x_s\|\big),$$

and $\|f(0,0,u)\| \le K$;

- *g satisfies the growth bound, for every u,*

$$|g(t,x_t,u) - g(s,x_s,u)| \le K\big(|t-s| + \|x_t - x_s\|\big)\big(1 + \|x_t\|^k + \|x_s\|^k\big).$$

Then

- *w is the unique solution to the Hamilton–Jacobi equation satisfying these continuity and growth bounds;*
- *w is the value function of the control problem;*
- *a control U is optimal if and only if*

$$U_t \in \arg\min_{u\in\mathcal{U}} \Big\{ g(t, X_t^U, u) + \Big\langle \nabla w(t, X_t^U), f(t, X_t^U, u) \Big\rangle \Big\}.$$

Proof. For notational convenience, we write

$$\tilde{H}(t,x,q,u) = g(t,x,u) + \big\langle q, f(t,x,u)\big\rangle \tag{3.2.2}$$

so that $H(t,x,q) = \inf_{u\in\mathcal{U}} \tilde{H}(t,x,q,u)$.

Step 1: Stability of $H, \tilde{H}$. We begin by showing some stability estimates for our problem[4]. Observe that, from Grönwall's inequality, as f is Lipschitz, given $X_0 = x_0$, there exists a constant K' depending on x_0 (and T, K) such that, if $\frac{\mathrm{d}}{\mathrm{d}t} X_t = f(t, X_t, U_t)$ for some U, then

$$\|X_t - X_s\| \le K'|t-s| \quad \text{and} \quad \|X_t\| \le K', \quad \text{for all } s,t \in [0,T].$$

By the triangle inequality, for any control $u \in \mathcal{U}$,

$$\begin{aligned}
&\Big|\tilde{H}\big(t, X_t, \nabla w(t,X_t), u\big) - \tilde{H}\big(s, X_s, \nabla w(s,X_s), u\big)\Big| \\
&\le \Big|g(t,X_t,u) - g(s,X_s,u)\Big| + \Big\|f(t,X_t,u) - f(s,X_s,u)\Big\| \cdot \|\nabla w(t,X_t)\| \\
&\qquad + \|f(s,X_s,u)\| \cdot \Big\|\nabla w(t,X_t) - \nabla w(s,X_s)\Big\| \\
&\le K\Big(|t-s| + \|X_t - X_s\|\Big)\Big(1 + \|X_t\|^k + \|X_s\|^k\Big) \\
&\qquad + K\Big(|t-s| + \|X_t - X_s\|\Big) K\Big(1 + t + \|X_t\|^k\Big)\Big(1 + \|X_t\|\Big) \\
&\qquad + K^2\Big(1 + s + \|X_s\|\Big)\Big(1 + \|X_s\|^k + \|X_t\|^k\Big)\Big(|t-s| + \|X_t - X_s\|\Big).
\end{aligned}$$

In particular, there exists a constant $c > 0$, depending on x_0, such that

$$\big|\tilde{H}\big(t, X_t, \nabla w(X_t), u\big) - \tilde{H}\big(s, X_s, \nabla w(X_s), u\big)\big| \le c|t-s|. \tag{3.2.3}$$

[4]We notice that this is the only point where we need the various continuity estimates on f, g, w, so if these vary, this is the only step that needs to be redone.

Taking $\varepsilon > 0$ and u^ε such that $\tilde{H}\big(s, X_s, \nabla w(s, X_s), u^\varepsilon\big) \le H\big(s, X_s, \nabla w(s, X_s)\big) + \varepsilon$ we have

$$\begin{aligned} H\big(t, X_t, \nabla w(t, X_t)\big) - H\big(s, X_s, \nabla w(s, X_s)\big) &\le \tilde{H}\big(t, X_t, \nabla w(t, X_t), u^\varepsilon\big) - H\big(s, X_s, \nabla w(s, X_s)\big) \\ &\le \tilde{H}\big(t, X_t, \nabla w(t, X_t), u^\varepsilon\big) - \tilde{H}\big(s, X_s, \nabla w(s, X_s), u^\varepsilon\big) + \varepsilon \\ &\le c|t - s| + \varepsilon \end{aligned}$$

and thus, exchanging (t, X_t) and (s, X_s), and taking $\varepsilon \to 0$,

$$\big|H\big(t, X_t, \nabla w(t, X_t)\big) - H\big(s, X_s, \nabla w(s, X_s)\big)\big| \le c|t - s|. \tag{3.2.4}$$

Step 2: Finding an optimizer. The next step is to construct our candidate near-optimal control. We do this by choosing U to be piecewise-constant, which has the advantage that it's easy to guarantee that U is admissible. Fix $\varepsilon > 0$, $x_0 \in \mathcal{X}$ and $\delta = \varepsilon/c$, where c is as in (3.2.3). Let U_0^ε be such that

$$\tilde{H}(t_0, x_0, \nabla w(x_0), U_0^\varepsilon) \le H(t_0, x_0, \nabla w(t_0, x_0)) + \varepsilon.$$

Using this U_0^ε, for $t \le \delta$, define $U_t^\varepsilon = U_0^\varepsilon$ and the ODE solution

$$X_t^\varepsilon = x_0 + \int_{t_0}^t f(s, X_s^\varepsilon, U_0^\varepsilon)\mathrm{d}s.$$

As $\delta = \varepsilon/c$, from (3.2.3) and (3.2.4) we know that

$$\begin{aligned} \tilde{H}(t, X_t^\varepsilon, \nabla w(t, X_t^\varepsilon), U_t^\varepsilon) &\le \tilde{H}(t_0, x_0, \nabla w(0, x_0), U_0^\varepsilon) + \varepsilon \le H(t_0, x_0, \nabla w(0, x_0)) + 2\varepsilon \\ &\le H(t, X_t^\varepsilon, \nabla w(t, X_t^\varepsilon)) + 3\varepsilon. \end{aligned} \tag{3.2.5}$$

We then repeat this construction started at $(t_0 + \delta, X_\delta^\varepsilon)$ instead of (t_0, x_0), which defines $U_t^\varepsilon, X_t^\varepsilon$ for $t \in (\delta, 2\delta]$. Iterating, we define $U^\varepsilon, X^\varepsilon$ for all t, and know that $\frac{\mathrm{d}}{\mathrm{d}t}X^\varepsilon = f(t, X_t^\varepsilon, U_t^\varepsilon)$ and $\tilde{H}(t, X_t^\varepsilon, \nabla w(t, X_t^\varepsilon), U_t^\varepsilon) \le H(t, X_t^\varepsilon, \nabla w(t, X_t^\varepsilon)) + 3\varepsilon$ for all t (and as c is a constant depending only on our first choice of x_0, we know ε, δ can be left fixed through the iteration). In particular, $U^\varepsilon \in \mathbb{U}$.

Step 3: Connecting the PDE to the cost-to-go. The next step is to show that the solution w to our PDE lower-bounds the cost-to-go $J(t, X_t^*, U)$ for all $U \in \mathbb{U}$. We know that

$$-\partial_t w = -\rho w + H(t, x, \nabla w).$$

For an arbitrary admissible control U, write $\tilde{w}(t) = e^{-\rho t}w(t, X_t^U)$, where $X_s^U = X_s^{U, t_0, x_0}$ for $s > t_0$. Then the chain rule tells us that (all derivatives of w being evaluated at (t, X_t^U), and the derivative of X^{U^*} interpreted in a weak sense)

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\tilde{w} &= -\rho e^{\rho t}w + e^{-\rho t}\Big[\Big\langle \nabla w(t, X_t^U), \frac{\mathrm{d}}{\mathrm{d}t}X_t^U \Big\rangle + \partial_t w(t, X_t^U)\Big] \\ &= e^{-\rho t}\Big[\Big\langle \nabla w(t, X_t^U), f(t, X_t^U, U_t)\Big\rangle - H(t, X_t^U, \nabla w(t, X_t^U))\Big]. \end{aligned}$$

Integration, along with the terminal value[5] of w, shows that

$$\begin{aligned}
&e^{-\rho t} w(t, X_t^U) = \tilde{w}(t) \\
&= \int_t^T e^{-\rho s} \Big[- \Big\langle \nabla w, f(s, X_s^U, U_s) \Big\rangle + H(s, X_s^U, \nabla w(s, X_s^U)) \Big] \mathrm{d}s + e^{-\rho T} \Phi(X_T^U) \\
&= \int_t^T e^{-\rho s} g(s, X_s^U, U_s) \mathrm{d}s + e^{-\rho T} \Phi(X_T^U) \\
&\quad + \int_t^T e^{-\rho s} \Big[H(s, X_s^U, \nabla w(s, X_s^U)) - \underbrace{\Big(g(s, X_s^U, U_s) + \Big\langle \nabla w(s, X_s^U), f(s, X_s^U, U_s) \Big\rangle \Big)}_{= \tilde{H}(s, X_s^U, \nabla w(s, X_s^U), U)} \Big] \mathrm{d}s.
\end{aligned}$$

In particular, by rearranging, we see that

$$w(t, X_t^U) = J(t, X_t^U, U) + \int_t^T e^{-\rho(s-t)} \underbrace{\Big[H(s, X_s^U, \nabla w(s, X_s^U)) - \tilde{H}(s, X_s^U, \nabla w(s, X_s^U), U) \Big]}_{\leq 0} \mathrm{d}s.$$

This immediately tells us that $w(t_0, x_0) \leq \inf_{U \in \mathbb{U}} J(t_0, x_0, U)$.

Step 4: Connecting w to the value function. Using this identity together with our candidate controls U^ε, we see that

$$w(t_0, x_0) \geq J(t_0, x_0, U^\varepsilon) - 3\varepsilon \int_t^T e^{-\rho(s-t)} \mathrm{d}s.$$

Taking $\varepsilon \to 0$, we conclude that $w(t_0, x_0) = \inf_{U \in \mathbb{U}} J(t_0, x_0, U)$. As t_0 and x_0 are arbitrary, we conclude that w is indeed the value function.

Now that we have the value function, we see that $w(0, x_0) = J(0, x_0, U)$ if and only if, for all t,

$$H(t, X_t^U, \nabla w(t, X_t^U)) = \tilde{H}(t, X_t^U, \nabla w(t, X_t^U), U_t),$$

that is, U is a minimizer in the Hamiltonian.

Step 5: Uniqueness of w. Finally, we know that the value function $v(t, x) = \inf_{U \in \mathbb{U}} J(t, x, U)$ is unique. However, we have shown that every solution of the PDE must be a value function for our control problem, hence the PDE must admit at most one solution $w = v$. □

Remark 3.2.5. As a corollary to this proof, we see that the infimal cost is approached by a sequence of piecewise constant controls.

Corollary 3.2.6. *Suppose the assumptions of Theorem 3.2.4 hold, and also the state dynamics f are Lipschitz with respect to u, and we can find a locally Lipschitz continuous map $u^* : \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ such that $u^*(t, x) \in \arg\min_{u \in \mathcal{U}} \tilde{H}\big(t, x, v(t, x), \nabla v(t, x), u\big)$. It follows that $\frac{\mathrm{d}}{\mathrm{d}s} X_s^{t,x} = f(s, X_s^{t,x}, u^*(s, X_s^{t,x}))$ admits a unique solution, and hence $U_s = u^*(s, X_s^{0,x_0})$ is admissible, and therefore is an optimal control.*

[5] This is where we use the boundary conditions, so if we have a different type of terminal condition for our control problem, this would need to change. For example, if we assume w is bounded and $\rho > 0$, we simply take the limit as $T \to \infty$ (after rescaling) and get $w(t, X^U) = \int_t^\infty e^{-\rho(s-t)} g(s, X_s^U, U_s) \mathrm{d}s + \int_t^\infty e^{-\rho(s-t)} (H(s, X_s^U, \nabla w(s, X_s^U)) - \tilde{H}(s, X_s^U, \nabla w(s, X_s^U), U)) \mathrm{d}s$. As (3.2.3) and (3.2.4) do depend on the terminal time T, we need to scale δ accordingly when defining the ε-optimal control, but we can do this sequentially without much difficulty. The argument then contines as before.

Remark 3.2.7. We might ask what happens when we do not have a smooth solution to the PDE. This naturally leads us into questions of what the right solution concept for PDEs is – we will discuss this further later.

Example 3.2.8. Consider the problem of minimizing $\Phi(X_T) = -X_T^2$ using controls in $\mathcal{U} = [-1, 1]$, and dynamics $\dot{X}_t = U_t$. Then the intuitively optimal strategy is to push as fast as you can away from the origin, leading to the value $v(t, x) = -(|x| + T - t)^2$. We observe that this is not C^1, and the optimal strategy is not unique when $x_0 = 0$. The Hamiltonian is simply $H(t, x, p) = \inf_{u\in[-1,1]}\{pu\} = -|p|$, while $\nabla v = 2\,\mathrm{sign}(x)(|x| + T - t)$, $\partial_t v = -2(|x| + T - t)$, so the Hamilton–Jacobi equation is satisfied, except at $x = 0$ (where we don't have enough derivatives to evaluate it).

Example 3.2.9. In the setting of Example 3.1.1, (minimize $\int_0^T |X_t|\mathrm{d}t$ with $\mathcal{U} = \{-1, 1\}$ and dynamics $\dot{X}_t = U_t$), the intuitively 'optimal' strategy is to push towards the origin as fast as you can and then oscillate close to zero arbitrarily quickly. The value function is given by

$$v(t, x) = \begin{cases} \big(|x| - \frac{T-t}{2}\big)(T - t) & \text{if } |x| > T - t, \\ \frac{x^2}{2} & \text{if } |x| \le T - t \end{cases}$$

which is C^1 but not C^2, and if $|x_0| < 1$ then the optimal strategy does not exist (as you can only approximate the optimal strategy near $X_t = 0$). We can check that

$$H(t, x, p) = \inf_{u\in\{\pm 1\}}\{|x| + pu\} = |x| - |p|$$

and then it's easy to check that, for $|x| \le T - t$, we have $H(t, x, \nabla v) = 0 = -\partial_t v$, and for $x > T - t$ we have $\nabla v = T - t$, $\partial_t v = T - t - x$, and so $H(t, x, \nabla v) = x - T + t = -(\partial_t v)$, and similarly for $x < -(T - t)$.

Example 3.2.10. Consider a linear-quadratic problem with state $X \in \mathcal{X} = \mathbb{R}^n$ and control space $U \in \mathcal{U} \subset \mathbb{R}^m$. We suppose X follows the the linear dynamics

$$\frac{\mathrm{d}X}{\mathrm{d}t} = f(t, X_t, u) = AX_t + Bu + C$$

and we face costs

$$g(t, x, u) = x^\top Qx + u^\top Ru + 2x^\top Su + 2Wx + 2Yu + Z$$

and

$$\Phi(x) = x^\top \Sigma_T x + 2\Psi_T x + \Gamma_T.$$

for matrices/vectors/scalars $A, B, C, Q, R, S, W, Y, Z$ and $\Sigma_T, \Psi_T, \Gamma_T$ of appropriate dimensions (one can make the parameters time dependent, and/or include a discount term, with a perturbation of notation). We assume Q, R are symmetric and R is strictly positive definite. The Hamilton–Jacobi equation becomes

$$-\partial_t v = H(t, x, \nabla v) = \min_u \Big\{x^\top Qx + u^\top Ru + 2x^\top Su + 2Wx + 2Yu + Z + (\nabla v)^\top (Ax + Bu + C)\Big\}.$$

We can then guess that the solution to the PDE should be a quadratic

$$v(t, x) = x^\top \Sigma_t x + 2\Psi_t x + \Gamma_t,$$

in which case

$$\nabla v = (\Sigma_t + \Sigma_t^\top)x + 2\Psi_t^\top.$$

Taking a first-order condition to optimize the Hamiltonian, we find (assuming $\mathcal{U}$ is sufficiently large), we have (with $\bar{\Sigma}_t = \frac{1}{2}(\Sigma_t + \Sigma_t^\top)$ the symmetric part of Σ_t)

$$0 = 2u^\top R + 2x^\top S + 2Y + 2(x^\top \bar{\Sigma}_t + \Psi_t)B$$

and hence the optimal strategy is

$$u_t^* = -R^{-1}\Big(S^\top x + Y^\top + B^\top(\bar{\Sigma}_t x + \Psi_t^\top)\Big) =: K_t x + H_t.$$

Substituting into the Hamilton–Jacobi equation, we have

$$\begin{aligned}
&-\partial_t(x^\top \Sigma_t x + 2\Psi_t x + \Gamma_t)\\
&= x^\top Q x + (K_t x + H_t)^\top R(K_t x + H_t) + 2x^\top S(K_t x + H_t) + 2Wx + 2Y(K_t x + H_t) + Z\\
&\quad + 2(x^\top \bar{\Sigma}_t + \Psi_t)(Ax + B(K_t x + H_t) + C).
\end{aligned}$$

Matching coefficients of x, we find the matrix Riccati system of equations

$$\begin{aligned}
-\partial_t \Sigma_t &= Q + K_t^\top R K_t + 2SK_t + 2\bar{\Sigma}_t(A + BK_t),\\
-\partial_t \Psi_t &= H_t^\top R K_t + H_t^\top S^\top + W + Y K_t + \Psi_t(A + BK_t) + (BH_t + C)^\top \bar{\Sigma}_t,\\
-\partial_t \Gamma_t &= H_t^\top R H_t + 2Y H_t + Z + 2\Psi_t(BH_t + C),
\end{aligned}$$

with terminal values Σ_T, Ψ_T and Γ_T specified. This can be simplified further if desired. While this equation is long, it is explicit, which makes it relatively easy to work with numerically. Solving this system of matrix ODEs, we get a smooth solution to the Hamilton–Jacobi equation, and hence the optimal control and value function. We can then check that the state process X does not get too large, so we can take $\mathcal{U}$ bounded by a large constant (ensuring the growth bounds in the derivation we have given are satisfied).

3.3 Pontryagin's maximum principle

For deterministic control, there is another elegant result that we can obtain. This is essentially the first-order condition for the minimization problem, and we can either obtain it through a constrained optimization argument, or from the Hamilton–Jacobi equation. This result is commonly known as Pontryagin's maximum principle (as Pontryagin derived it for control problems where we maximize rewards), but for the sake of consistency we will give a version for minimizing costs instead.

While it is only a necessary condition for optimality (like other first-order conditions), it turns out that in many cases this is enough – in particular if there is only one solution to the conditions, then that path must be optimal. More generally, the first order condition will give us *locally optimal* or *extremal* trajectories.

It turns out that it is then often possible to calculate this path by solving an ODE, rather than solving a PDE as we did when computing the Hamilton–Jacobi equation. We give only a derivation of this result (rather than stating a theorem), assuming f and g are smooth and the Hamilton–Jacobi equation admits a C^2 solution.

Fix $X_0 = x_0$ and suppose U^* is an optimal control, with the controlled process X^*. In order to derive a first order condition, assuming the Hamilton–Jacobi equation admits a sufficiently smooth solution, a

simple approach would be to say that, with the notation $\tilde{H}$ as in (3.2.2) $\partial_u \tilde{H}(t, X^*, \nabla v(t, X^*), U^*) = 0$. However, this is not immediately useful, as it still requires us to find ∇v, which involves solving the PDE. The trick is to find a representation of $\nabla v(t, X^*)$ which we can use directly.

We define the *adjoint* process $q : \mathbb{T} \times \mathbb{R}^n \to \mathbb{R}^n$ by

$$q_t = \nabla v(t, X_t^*).$$

Differentiating, we see that the ith component of q satisifes

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t^{(i)} = \frac{\mathrm{d}}{\mathrm{d}t}\Big(\partial_{x_i} v(t, X_t^*)\Big) = [\partial_t \partial_{x_i} v](t, X_t^*) + \Big\langle [\partial_{x_i} \nabla v](t, X_t^*), f(t, X_t^*, U_t^*) \Big\rangle.$$

As v satisfies the Hamilton–Jacobi equation, we know

$$-\partial_t v(t, x) = -\rho v + g(t, x, U_t^*) + f(t, x, U_t^*)^\top \nabla v(t, x)$$

and hence

$$\begin{aligned}\partial_t \partial_{x_i} v = \partial_{x_i} \partial_t v &= \rho \partial_{x_i} v_t - \partial_{x_i} g|_{U=U^*} - \partial_{x_i} \Big\langle \nabla v, f|_{U=U^*} \Big\rangle \\ &= \rho \partial_{x_i} v_t - \partial_{x_i} g|_{U=U^*} - \Big\langle \partial_{x_i} \nabla v, f|_{U=U^*} \Big\rangle - \Big\langle \nabla v, \partial_{x_i} f|_{U=U^*} \Big\rangle.\end{aligned}$$

Combining these, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t^{(i)} = \rho q_t^{(i)} - [\partial_{x_i} g](t, X_t^*, U_t^*) - \Big\langle q_t, [\partial_{x_i} f](t, X_t^*, U_t^*) \Big\rangle,$$

or as a vector equation

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t = \rho q_t - \nabla g(t, X_t^*, U_t^*) - \Big(D_x f(t, X_t^*, U_t^*)\Big)^\top q_t,$$

where $D_x f$ is the Jacobian of f, that is, the matrix with ith row given by the vector $\nabla[f^{(i)}]$. This ODE has terminal value $q(T) = \nabla v(T, X_T^*) = \nabla \Phi(X_T^*)$, and we observe that this gives q_t as the solution of a vector ODE whose dynamics do not involve v, assuming we already know the value of U^* and hence X^*.

Summarizing, if we just assume the Hamilton–Jacobi equation admits sufficiently smooth solutions, and an optimal control exists, we can try and identify the optimal control as a fixed point to the system of equations

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} X_t^* &= f(t, X_t^*, U_t^*), & X_0^* &= x_0; \\ \frac{\mathrm{d}}{\mathrm{d}t} q_t &= \rho q_t - \nabla g(t, X_t^*, U_t^*) - \Big(D_x f(t, X_t^*, U_t^*)\Big)^\top q_t, & q_T &= \nabla \Phi(X_T^*); \\ U^* &\in \operatorname*{arg\,min}_{u \in \mathcal{U}} \Big\{ g(t, X_t^*, u) + \Big\langle q_t, f(t, X_t^*, u) \Big\rangle \Big\}.\end{aligned}$$

Particularly when the dimension of X is high, it may be much more efficient to try and solve this forward-backward system of ODEs, rather than computing the solution to the PDE.

Remark 3.3.1. Numerous variations of this result exist, for different types of boundary conditions. In many cases, this involves computing an additional adjoint process, given by $\mu_t = [\partial_t H](t, X_t^*)$, which should satisfy certain additional boundary conditions (often called transversality conditions), related to what happens if the end point of our problem depends on the trajectory of X. If you look in the literature you will often see these presented as Pontryagin's maximum principle.

Remark 3.3.2. It is slightly odd that our derivation depends on having a C^2 solution to the Hamilton–Jacobi PDE, but we do not have the PDE appearing in the end result. This suggests that this result can be extended to situations where the PDE only has solutions in a weaker sense (e.g. in the viscosity sense). In fact this is completely true, but proving it requires more care.

Remark 3.3.3. We mentioned that this can also be seen in terms of the Lagrange multipliers/calculus of variations of a constrained optimization problem, without reference to the Hamilton–Jacobi equation or dynamic programming. For simplicity, set $\rho = 0$, and $n = 1$. We observe that we are trying to maximize $\int_0^T g(t, X_t, U_t)\mathrm{d}t + \Phi(X_T)$ subject to the constraint $\frac{\mathrm{d}}{\mathrm{d}t}X_t = f(t, X_t, U_t)$ for almost all t, over possible choices of U and X. A Lagrangian for this problem is given by

$$\int_0^T g(t, X_t, U_t)\mathrm{d}t + \Phi(X_T) - \int_0^T \lambda_t\Big(\frac{\mathrm{d}X_t}{\mathrm{d}t} - f(t, X_t, U_t)\Big)\mathrm{d}t.$$

Integrating by parts, we see that

$$\int_0^T \lambda_t \frac{\mathrm{d}X_t}{\mathrm{d}t}\mathrm{d}t = \lambda_T X_T - \lambda_0 x_0 - \int_0^T \frac{\mathrm{d}\lambda_t}{\mathrm{d}t}X_t\mathrm{d}t,$$

and so our Lagrangian becomes

$$\int_0^T \Big(g(t, X_t, U_t) + \frac{\mathrm{d}\lambda_t}{\mathrm{d}t}X_t + \lambda_t f(t, X_t, U_t)\Big)\mathrm{d}t + \Phi(X_T) - \lambda_T X_T + \lambda_0 x_0.$$

Differentiating with respect to X_t (or more formally, taking a variation $X_t + \varepsilon\eta$ for some smooth function η supported on a compact in $(0, T)$ and using the fundamental lemma of calculus of variations) we see that for almost all t we should have

$$\partial_x g(t, X_t, U_t) + \frac{\mathrm{d}\lambda_t}{\mathrm{d}t} + \lambda_t \partial_x f(t, X_t, U_t) = 0.$$

Differentiating with respect to X_T (or, formally, taking a variation with a smooth function with support in $[T - \varepsilon, T]$), we see that $\lambda_T = \partial_x \Phi(X_T)$. Therefore, λ satisfies the same differential equation as q did in our earlier derivation.

Finally, differentiating with respect to U_t (by taking a measurable variation) shows that for almost all t, with the notation of (3.2.2),

$$0 = \partial_u g(t, X_t, U_t) + \lambda_t \partial_u f(t, X_t, U_t) = \partial_u \tilde{H}(t, X_t, \lambda_t, U_t)$$

so U is an extreme point of $\tilde{H}$ (in particular, a minimum).

Chapter 4

Continuous Stochastic Control

In this final chapter of the course, we will consider the problem where our state process is stochastic, and we are in continuous time. We will work under somewhat restrictive assumptions on the class of problems that we consider – this will allow us to directly establish some continuity estimates for the value function, which will side-step the problem of proving measurability (which we really need, in order to be able to use probability, and becomes really tricky in many cases).

4.1 Notation and problem formulation

We assume that $\mathbb{T} = [0, T]$, and we have a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{T}}, \mathbb{P})$ (where $\{\mathcal{F}_t\}_{t\in\mathbb{T}}$ satisfies the usual conditions of completeness and right-continuity) and our state in $\mathcal{X} = \mathbb{R}^d$ follows the stochastic differential equation

$$\mathrm{d}X_t = f(t, X_t, U_t)\mathrm{d}t + \sigma(t, X_t, U_t)\mathrm{d}W_t \tag{4.1.1}$$

where W is an $\mathbb{R}^m$-dimensional Brownian motion, $f : \mathbb{T} \times \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$, $\sigma : \mathbb{T} \times \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^{d\times m}$. The process U is allowed to be a general $\{\mathcal{F}_t\}_{t\in\mathbb{T}}$-progressive[1] process in $\mathcal{U}$ (which we assume is a subset of some topological[2] space), that is, it can depend on all the information available at each point in time. We write $\mathbb{U}$ for the set of progressive processes.

We say U is of feedback form if there is a (Borel measurable) function u such that $U_t = u(t, X_t)$ at least $\mathrm{d}t \times \mathrm{d}\mathbb{P}$-a.e. (Sometimes, the terminology of U being *Markov* is used instead[3], but this is somewhat ambiguous, as $u(t, X_t)$ is not a Markov process, even if X is Markov.)

We interpret the state equation in an integral sense using Itô's integration theory. In order to work with dynamic programming more easily, we will define the family of controlled processes

$$X_{t'}^{t,x,U} = x + \int_t^{t'} f(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \int_t^{t'} \sigma(s, X_s^{t,x,U}, U_s)\mathrm{d}W_s$$

[1] Progressive measurability ensures that U is measurable in both time and in ω, in a nice way, see the appendix.

[2] This is again just so that we can talk about Borel maps to $\mathcal{U}$, which needs a notion of an open set.

[3] In fact, some authors say U is feedback if it is adapted to the filtration generated by X, and Markov if it is a function of (t, X_t). However, this is inconsistent with the common use of the term in deterministic control theory, and usage seems to vary in practice.

for all $t \in \mathbb{T}$, $x \in \mathcal{X}$, and $U \in \mathbb{U}$.

Our agent wishes to choose U to minimize their expected costs, which are given by $J(0, x_0, U)$, where J is the expected cost-to-go:

$$J(t,x,U) = \mathbb{E}\Big[\int_t^T g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \Phi(X_T^{t,x,U})\Big|\mathcal{F}_t\Big],$$

for cost functions $g : \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ and $\Phi : \mathcal{X} \to \mathbb{R}$. We can also include a discount term, but this simply adds notational complexity.

We will vary slightly from the approach we took in discrete time, and want to define

$$v(t,x) \stackrel{?}{=} \operatorname*{ess\,inf}_{U \in \mathbb{U}} J(t,x,U).$$

This definition allows v to be a random function of t, x, as we've simply not written the dependence on ω in v and J. However, we will see later that v is described by the solution (in some sense) to a PDE, and hence is deterministic. However, for now, we don't know this, and we just let $v : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathbb{R}$. In fact, we will be a little more careful in our definition (see later), as we need to make sure this is well defined simultaneously for all t, x. This is an issue because conditional expectations and essential infima are only defined almost everywhere, and we have uncountably many choices of t, x, so things can go wrong...

Example 4.1.1 (Merton portfolio problem)**.** A classic financial example is as follows: We have a financial asset described by the SDE

$$\mathrm{d}S_t = \mu S_t \mathrm{d}t + S_t \sigma \mathrm{d}W_t$$

for $\mu, \sigma > 0$. An investor has wealth x, and they choose an investment policy u^S determining the fraction of their wealth to invest. They also choose a consumption policy $u^C \geq 0$, determining how much of their wealth to consume at each time. Their wealth (assuming a zero interest rate) then is modelled by the SDE

$$\mathrm{d}X_t = \underbrace{-u_t^C X_t \mathrm{d}t}_{\text{consumption}} + \underbrace{u_t^S X_t(\mu \mathrm{d}t + \sigma dW_t)}_{\text{gains from trading}}.$$

and we model their costs as

$$-\int_0^T \mathfrak{u}(u_t^C X_t)\mathrm{d}t - \mathfrak{u}(X_T)$$

where $\mathfrak{u}$ is the utility they obtain from consumption, commonly of the form $\mathfrak{u}(c) = c^{1-\gamma}/(1-\gamma)$ for some $\gamma > 0$. We can assume u^C and u^S are bounded by some large constant (and with extensions of the results below, can show this is optimal).

Assumption 4.1.2. *In order for all of this to be well posed, we make the following assumptions:*

There exists a constant $K < \infty$ such that

- *f and σ are Lipschitz continuous with respect to x (uniformly in t, u), that is,*

$$\|f(t,x,u) - f(t,x',u)\| + \|\sigma(t,x,u) - \sigma(t,x',u)\| \leq K\|x - x'\|;$$

- *f and σ are continuous in t, Borel measurable in $\mathcal{U}$, and satisfy (for all t, u)*

$$\|f(t,0,u)\|^2 + \|\sigma(t,0,u)\|^2 \leq K;$$

- *g and Φ satisfy the bounds, for some $k \geq 1$ and all t, x, x', u,*

$$|g(t,x,u) - g(t,x',u)| + |\Phi(x) - \Phi(x')| \leq K\|x - x'\|(1 + \|x\|^k + \|x'\|^k),$$
$$|g(t,0,u)|^2 \leq K;$$

Remark 4.1.3. These bounds immediately imply that, for some $K' < \infty$,

$$\|f(t,x,u)\|^2 \leq 2\|f(t,x,u) - f(t,0,u)\|^2 + 2\|f(t,0,u)\|^2 \leq K'(1 + \|x\|^2),$$

and similarly for σ and g. We usually don't need to think about the exact value of K, so we are free to assume this also holds with $K = K'$, for notational simplicity.

Remark 4.1.4. The assumption that controls have only a bounded impact on costs[4] is somewhat restrictive, but it is possible to lift it in various ways. See Remark 4.2.6.

Remark 4.1.5. The key property of $\mathbb{U}$, the space of $\{\mathcal{F}_t\}_{t\geq 0}$-progressive processes in $\mathcal{U}$, that we will use is that it is closed under (countable) pastings. In particular, for any stopping time τ, if $\{A_i\}_{i\in\mathbb{N}} \subset \mathcal{F}_\tau$ is a countable partition of Ω, then for any $U_0, U_1, ... \in \mathbb{U}$, we know that $\tilde{U}_s := U_s^0 1_{s\leq\tau} + \sum_{i\in\mathbb{N}} 1_{s>\tau} 1_{A_i} U_s^i \in \mathbb{U}$.

4.1.1 Useful estimates

We want to define the value function by minimizing $J(t,x,U)$ with respect to $U \in \mathbb{U}$. The problem with this is that this involves taking an infimum over an uncountable set, which can lead to non-measurable functions. To avoid this, the *classic* method is to do a fairly careful analysis of how to do the selection of minimizers, as in Bertsekas and Shreve, Chapter 7.

We will present an alternative (somewhat non-standard) approach, where we first show that J has very strong continuity properties. This will allow us to obtain the standard results, but in a slightly different order to what is usual.

Lemma 4.1.6. *Given these assumptions on f and σ, we have the following standard properties:*

- *For every $(t,x,U) \in \mathbb{T} \times \mathcal{X} \times \mathbb{U}$ there exists a unique solution $X_s^{t,x,U}$ to the state equation, which is continuous in $s \geq t$.*

- *For each $p \geq 2$, there exists $K > 0$ such that, for all (t,x,U), the process $X_s^{t,x,U}$ satisfies the bound,*

$$\mathbb{E}\Big[\sup_{s\in[t,T]} \|X_s^{t,x,U}\|^p \Big| \mathcal{F}_t\Big] \leq K(1 + \|x\|^p).$$

- *There exists a constant $K < \infty$ such that, for all $(t,x,U), (t',x',U) \in \mathbb{T} \times \mathcal{X} \times \mathbb{U}$, with $t \leq t'$,*

$$\mathbb{E}\Big[\big\|X_T^{t,x,U} - X_T^{t',x',U}\big\|^2 \Big| \mathcal{F}_t\Big] \leq K(1 + \|x\|^2)\big(|t - t'| + \|x - x'\|^2\big).$$

[4]Some authors require that $\mathbb{E}[|g(t, X^{t,x,U}, U_t)|^2] < \infty$, or a similar condition, in the definition of admissibility, instead of assuming controls have bounded effects. However, such a requirement is a little difficult to work with, as it is not closed under countable pastings of strategies – just because U^i satisfies this bound for every i, we don't know that $\sum_i 1_{A_i} U_i$ also satisfies the bound for a countable partition of $\Omega \times \mathbb{T}$ by progressively measurable sets $\{A_i\}_{i\in\mathbb{N}}$. This makes understanding dynamic programming more difficult.

Proof. The core of these results is presented in Appendix A.2.1, which gives the more general case of SDEs with stochastic dynamics – this can be applied here by setting $\mu(\omega,t,x) = f(t,x,U_t(\omega))$, and similarly for σ, from which Theorem A.2.1 yields the existence of the unique solution.

To see the stated bounds, set $\mu(\omega,r,\xi) = f(r,\xi,U_r(\omega))1_{t\le r\le T}$ and $\tilde{\mu}(\omega,r,\xi) = f(r,\xi,U_r(\omega))1_{t'\le r\le T}$, and similarly for σ and $\tilde{\sigma}$. Then applying Lemma A.2.2 gives the growth bound on X. Applying Lemma A.2.3 with $\beta = 0$ implies that (for some constant C' depending on K), for $t \le t'$,

$$\begin{aligned}
&\mathbb{E}\Big[\|X_T^{t,x,U} - X_T^{t',x',U}\|^2\Big|\mathcal{F}_t\Big] \\
&\le C'\Big(\|x-x'\|^2 + \int_{[t,T]} \mathbb{E}[\|\mu_r(X_r^{t,x,U}) - \tilde{\mu}_r(X_r^{t,x,U})\|^2|\mathcal{F}_t] + \mathbb{E}[\|\sigma_r(X_r^{t,x,U}) - \tilde{\sigma}_r(X_r^{t,x,U})\|^2|\mathcal{F}_t]\mathrm{d}r\Big) \\
&\le C'\Big(\|x-x'\|^2 + \int_{[t,t']} \big(\mathbb{E}[\|\mu_r(X_r^{t,x,U})\|^2 + \|\sigma_r(X_r^{t,x,U})\|^2|\mathcal{F}_t]\big)\mathrm{d}r\Big) \\
&\le C'\Big(\|x-x'\|^2 + \sup_{r\in[t,T]}\Big(\mathbb{E}\Big[\|\mu_r(X_r^{t,x,U})\|^2 + \|\sigma_r(X_r^{t,x,U})\|^2\Big|\mathcal{F}_t\Big]\Big)|t-t'|\Big) \\
&\le C'\Big(\|x-x'\|^2 + \sup_{r\in[t,T]}\Big(\mathbb{E}\Big[\|f(r,X_r^{t,x,U},U_r)\|^2 + \|\sigma(r,X_r^{t,x,U},U_r)\|^2\Big|\mathcal{F}_t\Big]\Big)|t-t'|\Big) \\
&\le C'\Big(\|x-x'\|^2 + K\sup_{r\in[t,T]}\Big(1+\mathbb{E}\big[\|X_r^{t,x,U}\|^2\big|\mathcal{F}_t\big]\Big)|t-t'|\Big)
\end{aligned}$$

The result follows as $\mathbb{E}\big[\|X_r^{t,x,U}\|^2|\mathcal{F}_t] \le C(1+\|x\|^2)$. □

Lemma 4.1.7. *For any $K > 0$, the set of random variables*

$$\Big\{\Phi(X_T^U), \int_t^{t'} g(s,X_s^{t,x,U},U_s)\mathrm{d}s\Big\}_{U\in\mathbb{U},t\le t',\|x\|\le K}$$

is uniformly $\mathbb{P}$-integrable.

Proof. This follows directly from the bounds we have just established, together with Jensen's inequality (to deal with the integral), and the de la Vallée Poussin criterion for uniform integrability. □

Using these bounds, we can show that the following continuity estimate holds.

Theorem 4.1.8. *There exists a (deterministic) constant $K < \infty$ such that, with $k \ge 1$ from the growth bound on g and Φ, for every $U \in \mathbb{U}$ we have the almost sure (crude) inequality, for all $t,t' \in \mathbb{T}$ and $x,x' \in \mathcal{X}$,*

$$\mathbb{E}\Big[|J(t,x,U) - J(t',x',U)|\Big|\mathcal{F}_t\Big] \le K(1+\|x\|^{2k} + \|x'\|^{2k})\big(|t-t'|^{1/2} + \|x-x'\|\big).$$

Consequently, for each $U \in \mathbb{U}$, we can find a single function $\mathfrak{J}(\cdots,U) : \Omega\times\mathbb{T}\times\mathcal{X} \to \mathbb{R}$, which is continuous in t,x, and agrees with $J(t,x,U)$ almost surely for every (t,x).

Furthermore, there exists a function $v : \Omega\times\mathbb{T}\times\mathcal{X} \to \mathbb{R}$ such that $v(\omega,t,x) = \operatorname{ess\,inf}_{U\in\mathbb{U}} J(\omega,t,x,U)$ almost surely for each t,x (where the essential infimum is taken in the $\mathcal{F}_t$-measurable random variables), and for all $t,t' \in \mathbb{T}$ and $x,x' \in \mathcal{X}$, with $t \le t'$,

$$\mathbb{E}\Big[|v(t,x) - v(t',x')|\Big|\mathcal{F}_t\Big] \le K(1+\|x\|^{2k} + \|x'\|^{2k})\big(|t-t'|^{1/2} + \|x-x'\|\big).$$

Proof. Using our bounds and applying Cauchy–Schwarz, for K a constant which can vary from line to line,

$$
\begin{aligned}
&\mathbb{E}\Big[|J(t,x,U) - J(t',x',U)|\Big|\mathcal{F}_t\Big] \\
&\leq \mathbb{E}\Big[\int_t^{t'} |g(s, X_s^{t,x,U}, U_s)|\mathrm{d}s\Big|\mathcal{F}_t\Big] + \mathbb{E}\Big[|\Phi(X_T^{t,x,U}) - \Phi(X_T^{t',x',U})|\Big|\mathcal{F}_t\Big] \\
&\quad + \mathbb{E}\Big[\int_{t'}^{T} |g(s, X_s^{t,x,U}, U_s) - g(s, X_s^{t',x',U}, U_s)|\mathrm{d}s\Big|\mathcal{F}_t\Big] \\
&\leq \mathbb{E}\Big[\int_t^{t'} K(1 + \|X_s^{t,x,U}\|^2)\mathrm{d}s\Big|\mathcal{F}_t\Big] + \mathbb{E}\Big[K\|X_T^{t,x,U} - X_T^{t',x',U}\|(1 + \|X_T^{t,x,U}\|^k + \|X_T^{t',x',U}\|^k)\Big|\mathcal{F}_t\Big] \\
&\quad + \mathbb{E}\Big[\int_0^{T} K\|X_s^{t,x,U} - X_s^{t',x',U}\|(1 + \|X_s^{t,x,U}\|^k + \|X_s^{t',x',U}\|^k)\mathrm{d}s\Big|\mathcal{F}_t\Big] \\
&\leq K\int_t^{t'} (1 + \mathbb{E}\big[\|X_s^{t,x,U}\|^2\big|\mathcal{F}_t\big])\mathrm{d}s + K\mathbb{E}\Big[\|X_T^{t,x,U} - X_T^{t',x',U}\|^2\Big|\mathcal{F}_t\Big]^{1/2}\mathbb{E}\Big[\big(1 + \|X_T^{t,x,U}\|^k + \|X_T^{t',x',U}\|^k\big)^2\Big|\mathcal{F}_t\Big]^{1/2} \\
&\quad + K\int_0^{T} \mathbb{E}\Big[\|X_s^{t,x,U} - X_s^{t',x',U}\|^2\Big|\mathcal{F}_t\Big]^{1/2}\mathbb{E}\Big[(1 + \|X_s^{t,x,U}\|^k + \|X_s^{t',x',U}\|^k)^2\Big|\mathcal{F}_t\Big]^{1/2}\mathrm{d}s \\
&\leq K(1 + \|x\|^2)|t - t'| + K\Big((1 + \|x\|^2)\big(|t - t'| + \|x - x'\|^2\big)\Big)^{1/2}(1 + \|x\|^{2k} + \|x'\|^{2k})^{1/2} \\
&\quad + KT\Big((1 + \|x\|^2)(|t - t'| + \|x - x'\|^2)\Big)^{1/2}(1 + \|x\|^{2k} + \|x'\|^{2k})^{1/2} \\
&\leq K(1 + \|x\|^{2k} + \|x'\|^{2k})\Big(|t - t'|^{1/2} + \|x - x'\|\Big),
\end{aligned}
$$

where we have repeatedly used the inequality $(x + y)^{1/2} \leq x^{1/2} + y^{1/2}$ and the fact $(t - t')^{1/2} < T^{1/2}$.

These results are valid for each value of (t, x) and (t', x'), and we need to be careful, as the bound only holds almost surely, and we have uncountably many points to consider. However, by the Kolmogorov continuity criterion[5], we can find a single function $\mathfrak{J}(\cdots, U) : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathbb{R}$ which is continuous in (t, x), and agrees with these conditional expectations with probability one at every point. It's then easy to check (by inspecting a dense set and using continuity), that $\mathfrak{J}$ satisfies the continuity bounds we have just established for J.

We now seek to define v. For fixed values of t, x, define $\tilde{v}(t, x) = \operatorname{ess\,inf}_U J(t, x, U)$ (which we don't expect to have good properties in t, x). Observe that for each $(t, x), (t', x') \in \mathbb{T} \times \mathcal{X}$, there exists a sequence U^n such that $J(t', x', U^n) \to \tilde{v}(t', x')$. It follows that we have the almost sure inequality

$$
\begin{aligned}
\tilde{v}(t, x) - \tilde{v}(t', x') &\leq \lim_n J(t, x, U^n) - \tilde{v}(t', x') \\
&= \lim_n \Big(J(t, x, U^n) - J(t', x', U^n)\Big) \\
&\leq K(1 + \|x\|^{2k} + \|x'\|^{2k})(|t - t'|^{1/2} + \|x - x'\|).
\end{aligned}
$$

Exchanging the roles of (t, x) and (t', x') gives the lower bound. As for J, this inequality only holds almost surely for each choice of (t, x) and (t', x'), and we use Kolmogorov's continuity criterion to find a

[5]This is usually stated for random processes $X : \Omega \times \mathbb{T} \to \mathbb{R}$, but the proof extends reasonably easily to random fields $\mathfrak{J} : \Omega \times \mathbb{T} \times \mathbb{R}^n \to \mathbb{R}$, for any finite n. See [8, Theorem 25.2, p59] for a proof. Essentially, you fix $\mathfrak{J}$ to equal of $J(t, x, U)$ on the dyadic rationals, and then use Borel–Cantelli to show that taking limits in $\mathfrak{J}$ is valid, as there are not 'too many' dyadic rationals in a small set.

single function $v : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathbb{R}$ which is continuous in (t, x), agrees with $\tilde{v}(t, x)$ almost surely for each (t, x), and satisfies the above continuity bounds. □

Given this theorem, we will simply assume that $J = \mathfrak{J}$ in what follows, and always take this continuous version of the essential infimum v.

4.1.2 Dynamic programming

The following lemma may seem obvious from the definition, but the complexity is that on the left we are evaluating the random function J at the point $\tau(\omega), X_\tau^{t,x,U}(\omega)$, while on the right we are computing the conditional expectation of a random variable given the σ-algebra $\mathcal{F}_\tau$.

Lemma 4.1.9. *For any $(t, x, U) \in \mathbb{T} \times \mathcal{X} \times \mathbb{U}$ and any stopping time $\tau \geq t$, the cost-to-go function J satisfies*

$$J(\tau, X_\tau^{t,x,U}, U) = \mathbb{E}\Big[\int_\tau^T g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \Phi(X_T^{t,x,U})\Big|\mathcal{F}_\tau\Big] \qquad \mathbb{P}\text{-}a.s.$$

Proof. We begin by considering a deterministic time $t' > t$. For arbitrary $\varepsilon > 0$, and take a countable partition $\{A_n^\varepsilon\}_{n\in\mathbb{N}}$ of $\mathcal{X}$ such that $\max_{x,x'\in A_n^\varepsilon}\{\|x - x'\|^2\} \leq \varepsilon$. Choose[6] a point $x_n \in A_n^\varepsilon$ for each n. For $s \geq t'$, we consider the difference between $X_s^{t,x,U}$ and $\sum_n 1_{\{X_{t'}^{t,x,U}\in A_n^\varepsilon\}} X_s^{t',x_n,U}$; by a variation of the bounds we showed above, we know that there exists K such that

$$\begin{aligned}\sum_n 1_{\{X_{t'}^{t,x,U}\in A_n^\varepsilon\}}\mathbb{E}\Big[\big\|X_s^{t,x,U} - X_s^{t',x_n,U}\big\|^2\Big|\mathcal{F}_{t'}\Big] &\leq K(1 + \|X_{t'}^{t,x,U}\|^2)\Big(\sum_n 1_{\{X_{t'}^{t,x,U}\in A_n^\varepsilon\}}\Big\|X_{t'}^{t,x,U} - x_n\Big\|^2\Big)\\ &\leq K(1 + \|X_{t'}^{t,x,U}\|^2)\varepsilon.\end{aligned}$$

In particular, as L^2 convergence implies convergence of a subsequence almost everywhere, there is a sequence $\varepsilon_k \to 0$ such that

$$\sum_n 1_{\{X_s^{t,x,U}\in A_n^{\varepsilon_k}\}} X_s^{t',x_n,U} \to X_s^{t,x,U} \qquad (\mathrm{d}s + \delta_{s=T}) \times \mathrm{d}\mathbb{P}\text{-a.e.}$$

where $\delta_{s=T}$ is a point mass at T, and we look only on the interval $[t', T]$. From our assumed continuity of g and Φ, together with uniform integrability, it follows that

$$\begin{aligned}&\lim_{\varepsilon_k\to 0}\sum_n 1_{\{X_{t'}^{t,x,U}\in A_n^{\varepsilon_k}\}} J(t', x_n, U)\\ &= \lim_{\varepsilon_k\to 0}\sum_n 1_{\{X_{t'}^{t,x,U}\in A_n^{\varepsilon_k}\}}\mathbb{E}\Big[\int_{t'}^T g(s, X_s^{t',x_n,U}, U_s)\mathrm{d}s + \Phi(X_T^{t',x_n,U})\Big|\mathcal{F}_{t'}\Big]\\ &= \mathbb{E}\Big[\int_{t'}^T \lim_{\varepsilon_k\to 0}\sum_n 1_{\{X_{t'}^{t,x,U}\in A_n^{\varepsilon_k}\}} g(s, X_s^{t',x_n,U}, U_s)\mathrm{d}s + \lim_{\varepsilon_k\to 0}\sum_n 1_{\{X_{t'}^{t,x,U}\in A_n^{\varepsilon_k}\}}\Phi(X_T^{t',x_n,U})\Big|\mathcal{F}_{t'}\Big]\\ &= \mathbb{E}\Big[\int_{t'}^T g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \Phi(X_T^{t,x,U})\Big|\mathcal{F}_{t'}\Big].\end{aligned}$$

[6] For example, take the points with rational coordinates (which is a countable set) under your favourite ordering, and let x_i be the first point in A_i^ε. This can be done without using the axiom of choice!

On the other hand, from continuity of J we know that

$$J(t', X_{t'}^{t,x,U}, U) = \lim_{\varepsilon_k \to 0} \sum_n 1_{\{X_{t'}^{t,x,U} \in A_n^{\varepsilon_k}\}} J(t', x_n, U).$$

Combining these results, we see that, for each $t' \in [t, T]$,

$$J(t', X_{t'}^{t,x,U}, U) = \mathbb{E}\Big[\int_{t'}^{T} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \Phi(X_T^{t,x,U}) \Big| \mathcal{F}_{t'} \Big] \qquad \mathbb{P}\text{-a.e.}$$

Define a (right continuous, uniformly integrable) martingale M by

$$\begin{aligned} M_{t'} &= \mathbb{E}\Big[\int_{t}^{T} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \Phi(X_T^{t,x,U}) \Big| \mathcal{F}_{t'} \Big] \\ &= \int_{t}^{t'} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + J(t', X_{t'}^{t,x,U}, U). \end{aligned}$$

Observe that both sides of the outer equality are right continuous in t', so this equality holds up to a null set independent of time (and we have continuity, as the right hand side is continuous in t'). The martingale optional stopping theorem implies that

$$\int_{t}^{\tau} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + J(\tau, X_\tau^{t,x,U}, U) = M_\tau = \mathbb{E}\Big[\int_{t}^{T} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \Phi(X_T^{t,x,U}) \Big| \mathcal{F}_{\tau} \Big].$$

The result then follows by rearrangement and standard properties of the conditional expectation. □

Theorem 4.1.10. *The value function v satisfies the dynamic programming equation*

$$v(t,x) = \operatorname*{ess\,inf}_{U \in \mathbb{U}} \mathbb{E}\Big[\int_{t}^{\tau} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + v(\tau, X_\tau^{t,x,U}) \Big| \mathcal{F}_t \Big]$$

for any stopping time τ with $t \le \tau \le T$.

Proof. By construction, we know that $v(t,x) \le J(t,x,U)$ for all $U \in \mathbb{U}$. Fix t, x and τ. Using the previous lemma,

$$\begin{aligned} J(t,x,U) &= \mathbb{E}\Big[\int_{t}^{\tau} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \int_{\tau}^{T} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \Phi(X_T^{t,x,U}) \Big| \mathcal{F}_t \Big], \\ &= \mathbb{E}\Big[\int_{t}^{\tau} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + J(\tau, X_\tau^{t,x,U}, U) \Big| \mathcal{F}_t \Big] \\ &\ge \mathbb{E}\Big[\int_{t}^{\tau} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + v(\tau, X_\tau^{t,x,U}) \Big| \mathcal{F}_t \Big]. \end{aligned}$$

Taking the essential infimum with respect to U, we obtain

$$v(t,x) \ge \operatorname*{ess\,inf}_{U \in \mathbb{U}} \mathbb{E}\Big[\int_{t}^{\tau} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + v(\tau, X_\tau^{t,x,U}) \Big| \mathcal{F}_t \Big].$$

Conversely, fix $\varepsilon > 0$ and observe that we can find a countable partition of $\mathbb{T} \times \mathcal{X}$ into rectangles of the form $(t_i, t_{i+1}] \times A_i$ (with $A_i \in \mathcal{B}(\mathbb{R}^n)$ having nonempty interior), such that

$$\max_{(t,x),(t',x') \in A_i} \Big\{ (1 + \|x\|^{2k} + \|x'\|^{2k}) \big(|t - t'|^{1/2} + \|x - x'\| \big) \Big\} < \varepsilon.$$

Associated with each A_i we again choose a point $x_i \in A_i$ For each i, we take a sequence $U^{i,n} \subset \mathbb{U}$ such that $J(t_i, x_i, U^{i,n}) \to v(t_i, x_i)$ as $n \to \infty$. Define the pasted strategy

$$\tilde{U}^n = 1_{t<\tau} U + 1_{\{t \geq \tau\}} \sum_{i \in \mathbb{N}} 1_{\{(\tau, X_\tau^{t,x,U}) \in (t_i, t_{i+1}] \times A_i\}} U^{i,n}.$$

As this is based on a countable pasting, it is still admissible. Furthermore, from the continuity estimates above, writing $\tilde{K}_i = K(1 + 2\|X_\tau^{t,x,U}\|^{2k} + \|x_i\|^{2k} + \varepsilon)$ for K as in Theorem 4.1.8,

$$\begin{aligned}
&\limsup_n J(\tau, X_\tau^{t,x,U}, \tilde{U}^n) \\
&\leq \limsup_n \sum_{i \in \mathbb{N}} 1_{(\tau, X_\tau^{t,x,U}) \in (t_i, t_{i+1}] \times A_i} \Big(J(t_i, x_i, \tilde{U}) + \tilde{K}_i (|\tau - t_i|^{1/2} + \|X_\tau^{t,x,U} - x_i\|) \Big) \\
&\leq \sum_{i \in \mathbb{N}} 1_{(\tau, X_\tau^{t,x,U}) \in (t_i, t_{i+1}] \times A_i} \Big(\limsup_n J(t_i, x_i, U^{i,n}) + \tilde{K}_i \varepsilon \Big) \\
&= \sum_{i \in \mathbb{N}} 1_{(\tau, X_\tau^{t,x,U}) \in (t_i, t_{i+1}] \times A_i} \Big(v(t_i, x_i) + \tilde{K}_i \varepsilon \Big) \\
&\leq v(\tau, X_\tau^{t,x,U}) + 2\varepsilon \sum_{i \in \mathbb{N}} 1_{(\tau, X_\tau^{t,x,U}) \in (t_i, t_{i+1}] \times A_i} \tilde{K}_i
\end{aligned}$$

By definition, we know that

$$v(t,x) \leq \mathbb{E}\Big[\int_t^\tau g(s, X_s^{t,x,U}, U_s) \mathrm{d}s + J(\tau, X_\tau^{t,x,U}, \tilde{U}^n) \Big| \mathcal{F}_t \Big].$$

Therefore, taking the lim sup as $n \to \infty$ (by uniform integrability), we see that, for some K' (depending on the moment bounds on $X^{t,x,U}$, and hence on (t,x)),

$$\begin{aligned}
v(t,x) &\leq \limsup_n \mathbb{E}\Big[\int_t^\tau g(s, X_s^{t,x,U}, U_s) \mathrm{d}s + J(\tau, X_\tau^{t,x,U}, \tilde{U}^n) \Big| \mathcal{F}_t \Big] \\
&\leq \mathbb{E}\Big[\int_t^\tau g(s, X_s^{t,x,U}, U_s) \mathrm{d}s + v(\tau, X_\tau^{t,x,U}) \Big| \mathcal{F}_t \Big] + K' \varepsilon.
\end{aligned}$$

Finally, taking $\varepsilon \to 0$, we see that for any $U \in \mathbb{U}$, we have

$$v(t,x) \leq \mathbb{E}\Big[\int_t^\tau g(s, X_s^{t,x,U}, U_s) \mathrm{d}s + v(\tau, X_\tau^{t,x,U}) \Big| \mathcal{F}_t \Big].$$

Taking the essential infimum with respect to U completes the proof. □

Corollary 4.1.11 (Martingale Optimality Principle)**.** *For any control $U \in \mathbb{U}$, the process defined by*

$$M_t^U = v(t, X_t^{0,x_0,U}) + \int_0^t g(s, X_s^{0,x_0,U}, U_s) \mathrm{d}s$$

is a submartingale, and is a martingale if and only if U is optimal.

Proof. For fixed $t < t'$, we know that

$$\begin{aligned}
M_t^U &= v(t, X_t^{0,x_0,U}) + \int_0^t g(s, X_s^{0,x_0,U}, U_s) \mathrm{d}s \\
&= \operatorname*{ess\,inf}_{U' \in \mathbb{U}} \mathbb{E}\Big[\int_t^{t'} g(s, X_s^{t, X_t^{0,x_0,U}, U'}, U_s') \mathrm{d}s + v(t', X_{t'}^{t, X_t^{0,x_0,U}, U'}) \Big| \mathcal{F}_t \Big] + \int_0^t g(s, X_s^{0,x_0,U}, U_s) \mathrm{d}s \\
&\leq \mathbb{E}\Big[\int_t^{t'} g(s, X_s^{0,x_0,U}, U_s) \mathrm{d}s + v(t, X_{t'}^{0,x_0,U}) \Big| \mathcal{F}_t \Big] + \int_0^t g(s, X_s^{0,x_0,U}, U_s) \mathrm{d}s \\
&= \mathbb{E}[M_{t'}^U | \mathcal{F}_t].
\end{aligned}$$

so M is a submartingale. If (and only if) U is optimal, then this is an equality, in which case M is a martingale. □

4.2 Hamilton–Jacobi–Bellman equations

Given we know our value function v satisifes the dynamic programming equation, the natural next step is to derive a PDE which we expect it to satisfy. As we did in the deterministic setting, we first give a heuristic derivation of the result.

Suppose v is smooth and independent of ω. Itô's lemma tells us that, for any process $X = X^{t,x,U}$ of the form we are considering,

$$v(t', X_{t'}) = v(t, X_t) + \int_t^{t'} \partial_t v \, \mathrm{d}s + \int_t^{t'} (D_x v)^\top \mathrm{d}X_s + \frac{1}{2}\int_t^{t'} \mathrm{Tr}\Big[(D^2_{xx} v)\frac{\mathrm{d}\langle X\rangle_t}{\mathrm{d}t}\Big]\mathrm{d}s$$

where $D_x v = \nabla v$ and $D^2_{xx} v$ is the Hessian of v, and all derivatives are evaluated at (s, X_s). We know that $\mathrm{d}X_s = f(s, X_s, U_s)\mathrm{d}t + \sigma(s, X_s, U_s)\mathrm{d}W_s$ and $\frac{\mathrm{d}\langle X\rangle_t}{\mathrm{d}t} = (\sigma\sigma^\top)(s, X_s, U_s)$, so, dropping the s, X_s arguments for simplicity,

$$v(t', X_{t'}) = v(t, X_t) + \int_t^{t'} \partial_t v + (D_x v)^\top f(U_s) + \frac{1}{2}\mathrm{Tr}\Big[(D^2_{xx} v)\big((\sigma\sigma^\top)(U_s)\big)\Big]\mathrm{d}s + \int_t^{t'} (D_x v)^\top \sigma(U_s)\mathrm{d}W_s$$

Taking an expectation, we drop the $\mathrm{d}W$ term (as this is a martingale), and so find

$$v(t,x) = \mathbb{E}\Big[v(t', X^{t,x,U}_{t'}) - \int_t^{t'} \partial_t v + (D_x v)^\top f(U_s) + \frac{1}{2}\mathrm{Tr}\Big[(D^2_{xx} v)\big((\sigma\sigma^\top)(U_s)\big)\Big]\mathrm{d}s\Big].$$

At the same time, from the dynamic programming principle, we know

$$v(t,x) = \inf_{U\in\mathbb{U}} \mathbb{E}\Big[\int_t^{t'} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s + v(t', X^{t,x,U}_{t'})\Big]$$

so by substitution, we get the equation (with all terms evaluated at $X^{t,x,U}_s$)

$$0 = \inf_{U\in\mathbb{U}} \mathbb{E}\Big[\int_t^{t'} g(U_s)\mathrm{d}s + \int_t^{t'} \partial_t v + (D_x v)^\top f(U_s) + \frac{1}{2}\mathrm{Tr}\Big[(D^2_{xx} v)\big((\sigma\sigma^\top)(U_s)\big)\Big]\mathrm{d}s\Big]$$

As this must hold for all t, t', we divide by $t' - t$ and take $t' \searrow t$. Approximating $U_s \approx u$ and $X^{t,x,U_s}_s \approx x$, so there is no randomness left in our equation, we simply have

$$0 = \inf_{u\in\mathcal{U}} \Big\{\partial_t v(t,x) + g(t,x,u) + \big(D_x v(t,x)\big)^\top f(t,x,u) + \frac{1}{2}\mathrm{Tr}\Big[\big(D^2_{xx} v(t,x)\big)\big((\sigma\sigma^\top)(t,x,u)\big)\Big]\Big\}$$

or equivalently

$$-\partial_t v = H(t, x, D_x v, D^2_{xx} v)$$

where $H : \mathbb{T}\times\mathcal{X}\times\mathbb{R}^d\times\mathbb{R}^{d\times d} \to \mathbb{R}$ is the Hamiltonian

$$H(t,x,p,q) = \inf_{u\in\mathcal{U}} \underbrace{\Big\{g(t,x,u) + p^\top f(t,x,u) + \frac{1}{2}\mathrm{Tr}\Big[q\,\big((\sigma\sigma^\top)(t,x,u)\big)\Big]\Big\}}_{\tilde{H}(t,x,p,q,u)}.$$

This equation is known as the Hamilton–Jacobi–Bellman (or HJB) equation, as it is a second-order extension of the Hamilton–Jacobi equation we have previously seen. We immediately notice that in the case $\sigma \equiv 0$, where our dynamics do not have any stochastic term, we recover the Hamilton–Jacobi equation we studied in the previous chapter.

We now verify that this PDE is the 'right' representation of our value process. We give a slightly more delicate version than we had in the deterministic case.

Theorem 4.2.1 (Verification Theorem)**.** *Let v be the value function, and $w : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ a $C^{1,2}\big([0,T) \times \mathbb{R}^d\big) \cap C^0\big([0,T] \times \mathbb{R}^d\big)$ function satisfying the polynomial growth condition that there exists $p, K > 0$ such that*

$$|w(t,x)| \leq K(1 + \|x\|^p) \text{ for all } (t,x) \in [0,T] \times \mathbb{R}^d.$$

(i) Suppose that

$$\begin{aligned} -\partial_t w(t,x) &\leq H(t,x,D_x w(t,x), D^2_{xx} w(t,x)) \text{ for all } (t,x) \in [0,T) \times \mathbb{R}^d, \\ w(T,x) &\leq \Phi(x) \text{ for all } x \in \mathbb{R}^d. \end{aligned}$$

Then $w(t,x) \leq v(t,x)$ almost surely, for all $(t,x) \in [0,T] \times \mathbb{R}^d$.

(ii) Suppose that

$$\begin{aligned} -\partial_t w(t,x) &\geq H(t,x,D_x w(t,x), D^2_{xx} w(t,x)) \text{ for all } (t,x) \in [0,T) \times \mathbb{R}^d, \\ w(T,x) &\geq \Phi(x) \text{ for all } x \in \mathbb{R}^d. \end{aligned}$$

Then $w(t,x) \geq v(t,x)$ almost surely, for all $(t,x) \in [0,T] \times \mathbb{R}^d$.

(iii) Suppose that both (i) and (ii) hold (so w satisfies the HJB equation, and $w = v$ almost surely), and there exists a Borel measurable function $u : [0,T) \times \mathbb{R}^d \to \mathcal{U}$ such that the SDE

$$\mathrm{d}X_t = f(t, X_t, u(t,X_t))\mathrm{d}t + \sigma(t, X_t, u(t,X_t))\mathrm{d}W_t; \qquad X_0 = x_0$$

admits a unique (strong) solution, and u achieves the minimization

$$H(t,x,D_x w(t,x), D^2_{xx} w(t,x)) = \tilde{H}(t,x,D_x w(t,x), D^2_{xx} w(t,x), u(t,x))$$

for all $(t,x) \in [0,T) \times \mathbb{R}^d$. Then $U_t = u(t,X_t)$ is an optimal control, and is a feedback control.

Proof. Expanding the PDE. To begin, choose an arbitrary control $U \in \mathbb{U}$. We know that for any stopping time $\tau \in [t,T]$ we can apply Itô's formula

$$\begin{aligned} w(t' \wedge \tau, X^{t,x,U}_{t' \wedge \tau}) = w(t,x) &+ \int_t^{t' \wedge \tau} (D_x w(s, X^{t,x,U}_s))^\top \sigma(s, X^{t,x,U}_s, U_s)\mathrm{d}W_s \\ &+ \int_t^{t' \wedge \tau} \partial_t w(s, X^{t,x,U}_s) + (D_x w(t,x))^\top f(t,x,u) + \frac{1}{2}\mathrm{Tr}\Big[(D^2_{xx} w(t,x))\big((\sigma\sigma^\top)(t,x,u)\big)\Big]\mathrm{d}s \end{aligned}$$

Choose $\tau_n = \inf\{t' > t : \int_t^{t'} \|(D_x w(s, X^{t,x,U}_s))^\top \sigma(s, X^{t,x,U}_s, U_s)\|^2 \mathrm{d}s \geq n\} \wedge T$, so that we know that the $\mathrm{d}W$ term in the above formula has finite quadratic variation on $[0, \tau_n]$, and is therefore a true martingale

(in particular with expected value zero). We also notice that $\tau_n \nearrow T$ as $n \to \infty$, as the integrands are continuous. Hence, for any stopping time $\tau \leq T$, with $\tau \wedge \tau_n = \min\{\tau, \tau_n\}$ we have

$$\begin{aligned}&\mathbb{E}\Big[w(\tau \wedge \tau_n, X^{t,x,U}_{t'\wedge\tau})\Big|\mathcal{F}_t\Big]\\&= w(t,x) + \mathbb{E}\Big[\int_t^{\tau\wedge\tau_n} \partial_t w(s, X^{t,x,U}_s) - g(s, X^{t,x,U}_s, U_s) + \tilde{H}(s, X^{t,x,U}_s, D_x w, D^2_{xx} w, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big].\end{aligned}$$

by rearrangment, we can write this in the same form as the dynamic programming equation

$$\begin{aligned}w(t,x) = \mathbb{E}\Big[w(\tau \wedge \tau_n, X^{t,x,U}_{\tau\wedge\tau_n}) + \int_t^{\tau\wedge\tau_n} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big]\\ - \mathbb{E}\Big[\int_t^{\tau\wedge\tau_n} \partial_t w(s, X^{t,x,U}_s) + \tilde{H}(t, X^{t,x,U}_s, D_x w, D^2_{xx} w, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big].\end{aligned}$$

As we know that $|w(t,x)| \leq K(1+\|x\|^p)$ and $g(t,x,u) \leq K(1+\|x\|^{k+1})$, uniform integrability (cf. Lemma 4.1.7) implies that we can take $n \to \infty$, and thus

$$\begin{aligned}w(t,x) = \mathbb{E}\Big[w(\tau, X^{t,x,U}_\tau) + \int_t^{\tau} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big]\\ - \lim_{n\to\infty} \mathbb{E}\Big[\int_t^{\tau\wedge\tau_n} \partial_t w(s, X^{t,x,U}_s) + \tilde{H}(s, X^{t,x,U}_s, D_x w, D^2_{xx} w, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big].\end{aligned} \tag{4.2.1}$$

Step 2: Case (i). Now suppose the conditions of (i) hold. Then

$$-\partial_t w(s, X^{t,x,U}_s) \leq H(s, X^{t,x,U}_s, D_x w, D^2_{xx} w) \leq \tilde{H}(s, X^{t,x,U}_s, D_x w, D^2_{xx} w, U_s),$$

so the second line of (4.2.1) is negative. Therefore, with $\tau = T$,

$$w(t,x) \leq \mathbb{E}\Big[w(T, X^{t,x,U}_T) + \int_t^T g(s, X^{t,x,U}_s, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big] \leq \mathbb{E}\Big[\Phi(X^{t,x,U}_T) + \int_t^T g(s, X^{t,x,U}_s, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big]$$

and as $U \in \mathbb{U}$ is arbitrary, by taking an infimum we see that $w \leq v$.

Step 3: Case (ii). Now suppose the conditions of (ii) hold. In order to establish the bound, we first need to find a control $U \in \mathbb{U}$ such that, for some fixed $\varepsilon > 0$,

$$\tilde{H}(s, X^{t,x,U}_s, D_x w, D^2_{xx} w, U_s) \leq H(s, X^{s,x,U}_s, D_x w, D^2_{xx} w) + \varepsilon.$$

For notational simplicity, write $h(t,x,u) = \tilde{H}(t,x,D_x w, D^2_{xx} w, u)$. We know that h is continuous with respect to (t,x), and in particular is uniformly continuous on $[0,T'] \times \mathcal{X}_K$, uniformly in $u \in \mathcal{U}$, for all $T' < T$ and $\mathcal{X}_K = \{x : \|x\| \leq K\}$ with $K < \infty$.

Using this uniform continuity, we partition $[0,T]\times\mathcal{X}$ into countably many pieces of the form $[t_i, t_{i+1}) \times A_i$, such that

$$|h(t,x,u) - h(t',x',u)| \leq \delta \text{ for all } (t,x),(t',x') \in [t_i, t_{i+1}) \times A_i$$

for all $u \in \mathcal{U}$ and all i. We fix some $x_i \in A_i$, and find $u_i \in \mathcal{U}$ such that $h(t_i, x_i, u_i) \leq \inf_u h(t_i, x_i, u) + \varepsilon/2$. We define $u^\varepsilon(t,x) = u_0 1_{t=0} + \sum_i u_i 1_{(t,x)\in[t_i,t_{i+1})\times A_i}$, and observe that $u^\varepsilon : \mathbb{T}\times\mathcal{X} \to \mathcal{U}$ is Borel measurable.

Now, for a given x_0, we define X^ε to be the solution of the state dynamics started at x_0, with constant control $U^\varepsilon_t := u^\varepsilon(0, x_0)$, up to the stopping time

$$\rho_1 = \min\Big\{\inf\{t : \|h(t, X^\varepsilon_t, U^\varepsilon_t) - h(t, x_0, U^\varepsilon_t)\| > \varepsilon/2\},\, T\Big\}.$$

With this definition, we know

$$h(t, X_t^\varepsilon, U_t^\varepsilon) \leq \inf_u h(t, X_t^\varepsilon, u) + \varepsilon \quad \text{for } t \leq \rho_1.$$

We now iterate this construction, by defining the piecewise constant control $U_t^\varepsilon = u^\varepsilon(\rho_n, X_{\rho_n}^\varepsilon)$ for $t \in (\rho_n, \rho_{n+1}]$, where X^ε is the controlled state and

$$\rho_{n+1} = \min\Big\{ \inf\{t : \|h(t, X_t^\varepsilon, U_t^\varepsilon) - h(t, X_{\rho_n}^\varepsilon, U_t^\varepsilon)\| > \varepsilon/2\}, T \Big\}.$$

As U^ε only depends on past values of X, it is easy to check that U^ε is progressive, that the state dynamics admit a unique continuous solution with control U^ε, and that, from condition (ii),

$$h(t, X_t^\varepsilon, U_t^\varepsilon) \leq \inf_u h(t, X_t^\varepsilon, u) + \varepsilon \leq -\partial_t w(t, X_t^\varepsilon) + \varepsilon \quad \text{for } t \leq \rho_n.$$

By the argument of Lemma A.2.4, as X^ε satisfies our moment bounds its paths do not explode, and hence we also know that $\rho_n \to T$ as $n \to \infty$, almost surely.

We now compare with (4.2.1), to observe that

$$\begin{aligned} w(t,x) &= \mathbb{E}\Big[w(\rho_n, X_{\rho_n}^\varepsilon) + \int_t^{\rho_n} g(s, X_s^\varepsilon, U_s^\varepsilon)\mathrm{d}s \Big| \mathcal{F}_t\Big] - \mathbb{E}\Big[\int_t^{\rho_n} \partial_t w(s, X_s^\varepsilon) + h(t, X_s^\varepsilon, U^\varepsilon)\mathrm{d}s \Big| \mathcal{F}_t\Big] \\ &\geq \mathbb{E}\Big[w(\rho_n, X_{\rho_n}^\varepsilon) + \int_t^{\rho_n} g(s, X_s^\varepsilon, U_s^\varepsilon)\mathrm{d}s \Big| \mathcal{F}_t\Big] - \mathbb{E}\Big[\int_t^{\rho_n} \varepsilon \mathrm{d}s \Big| \mathcal{F}_t\Big] \end{aligned}$$

We now take $n \to \infty$, so by dominated convergence (as in Step 1)

$$\begin{aligned} w(t,x) &\geq \mathbb{E}\Big[w(T, X_T^\varepsilon) + \int_t^T g(s, X_s^\varepsilon, U_s^\varepsilon)\mathrm{d}s \Big| \mathcal{F}_t\Big] - \mathbb{E}\Big[\int_t^T \varepsilon \mathrm{d}s \Big| \mathcal{F}_t\Big] \\ &\geq \mathbb{E}\Big[\Phi(X_T^\varepsilon) + \int_t^T g(s, X_s^\varepsilon, U_s^\varepsilon)\mathrm{d}s \Big| \mathcal{F}_t\Big] - \varepsilon(T-t) \\ &\geq v(t,x) - \varepsilon(T-t). \end{aligned}$$

Finally, we conclude by taking $\varepsilon \searrow 0$.

Step 4. Describing an optimizer. Under condition (iii), we see from (4.2.1) that, with $U_s = u(t, X_s)$ and X the solution to the state dynamics,

$$w(t,x) = v(t,x) = \mathbb{E}\Big[\Phi(X_T) + \int_t^T g(s, X_s, U_s)\mathrm{d}s \Big| \mathcal{F}_t\Big].$$

Therefore U is an optimal control, as stated.

□

Remark 4.2.2. We've actually shown quite a bit more in this proof than it seems. We know that the value function exists and is continuous, and this theorem tells us that it lies above every smooth subsolution of the PDE (Case (i)), and below every smooth supersolution of the PDE (Case (ii)). This is closely related to the fact that it is a *viscosity solution* of the HJB equation, *even if we don't know that it's smooth!* Furthermore, there is a unique continuous viscosity solution to the HJB equation of polynomial growth, so this characterizes the value function completely. In particular, this implies that *the value function is independent of ω, as it must be the (viscosity) solution of a deterministic PDE!*

This is one of the directions that this theory develops in – showing that value functions are generally viscosity solutions of the HJB equation, and then using this representation to find numerical methods with which to solve the equation and thus the control problem. For more details of this theory, a classic place to start is Crandall, Ishii, and Lions [3], and Fleming and Soner [5] give a nice summary for control problems.

Remark 4.2.3. Note that the verification argument doesn't use the boundedness of f and σ, except in ensuring uniform integrability of X^U, and hence the uniform integrability of the costs. This suggests that we can lift that assumption, provided we have enough restrictions on our controls to ensure this fact. For example, if g, Φ are bounded, we can take any f, σ such that X is well defined for piecewise-constant controls.

Remark 4.2.4. In the truly stochastic context, we also know that HJB equation usually admits $C^{1,2}$ solutions, at least under the 'uniform ellipticity' condition that there exists a constant $\delta > 0$ such that $\|\lambda^\top \sigma(t,x,u)\| \geq \delta \|\lambda\|$ for all $\lambda \in \mathbb{R}^n$. This is a result due to Evans and Krylov, see Krylov [7, Chapter 4] for details in this direction.

Remark 4.2.5. The approach we took (based on piecewise constant controls) is easy to work with, but a nicer result is that in many cases, there exists an ε-optimal feedback control (with enough smoothness to guarantee that the state dynamics admit a unique solution). Indeed, given a smooth solution to the HJB equation, and sufficient smooth invertibility of the Hamiltonian, we expect that this will be the case.

Remark 4.2.6. While we have derived all our results under the assumption that controls have a bounded impact, we could now consider taking a sequence of problems where we relax this condition. Suppose our coefficients f, σ, g admit growth bounds with respect to u (which now takes values in a metric space), then the problem with u restricted to a bounded set is as we have described. As we relax the bounds on this set, we are allowing more controls, and so obtain a sequence of (pointwise decreasing) value functions. Taking the limit (assuming this sequence does not diverge), we can show that the limit also satisfies a version of the Hamilton–Jacobi–Bellman equation (at least in viscosity sense).

The only question is whether this limit coincides with the problem where we allow unbounded controls. If we assume that there is an almost-surely bounded ε-optimal control (in the unbounded problem), for all $\varepsilon > 0$, then it is easy to see that this is indeed the case.

Appendix A

Some useful basic theory

We here give a summary of additional results in probability theory which we make use of in the course. Some of these we simply state without proof (where the proofs are given in either B8.1 or B8.2), other less familiar results we reproduce in full. While we reproduce the key definitions below, it would be much better (if you're not familiar with the material) to try and learn it from a more developed text, for example [2].

A.1 Filtrations, Conditional Expectations, and Martingales

The basic structure we will use in order to understand stochastic processes is that of a filtered probability space. This is an abstract axiomatization (essentially due to Kolmogorov) of probability theory, which enables us to study the flow of information through time.

Definition A.1.1. *A* measurable space *is a set S, together with a σ-algebra $\mathcal{F}$ (that is, a family of subsets of S which is closed under taking complements and countable unions, and contains S). The elements of $\mathcal{F}$ are called events (in probabilistic language).*

A measure *μ is a map $\mathcal{F} \to [0, \infty]$ with the properties that $\mu(\emptyset) = 0$ and μ is additive with respect to countable disjoint unions, that is, for disjoint sets $A_1, A_2, ...$, we know $\mu(\cup_n A_n) = \sum_n \mu(A_n)$. A measure μ is called a probability measure if $\mu(S) = 1$, and by convention, we write $S = \Omega$ in this case. A measurable space with a probability measure is called a probability space.*

We say a measure space is complete *if for every $A \in \mathcal{F}$ with $\mu(A) = 0$, we know $B \in \mathcal{F}$ for all $B \subseteq A$.*

Definition A.1.2. *A filtration on a measurable space is an increasing collection of sub-σ-algebras of $\mathcal{F}$, that is, a family $\{\mathcal{F}_t\}_{t \in \mathbb{T}}$ such that $\mathcal{F}_t \subseteq \mathcal{F}_{t'} \subseteq \mathcal{F}$ for all $t \leq t'$. This provides a precise way to model the events which are determined by time t – any such event is included in $\mathcal{F}_t$, and we assume that this is still a σ-algebra. A measurable space with a filtration is called a filtered space (and hence filtered probability spaces are defined).*

Definition A.1.3. *The Borel σ-algebra on a topological space $\mathcal{Y}$ is the smallest σ-algebra containing all open sets, and is written $\mathcal{B}(\mathcal{Y})$*

Definition A.1.4. *A function $f : S \to \mathcal{Y}$, where $\mathcal{Y}$ is a topological space*[1] *and $(S, \mathcal{F})$ is a measurable space, is said to be* measurable *if for all sets $A \in \mathcal{B}(\mathcal{Y})$, we know $X^{-1}(A) \in \mathcal{F}$. If S is also a topological space, we say X is* Borel measurable *if $X^{-1}(A) \in \mathcal{B}(S)$. Every continuous function is Borel measurable.*

If S is the set of random seeds Ω, we often call measurable functions by the name random variables*.*

Understanding σ-algebras is made somewhat easier by taking a concrete example, which is provided by the σ-algebra generated by a function f

Definition A.1.5. *Given a function f, the σ-algebra generated by f is the smallest σ-algebra on Ω such that f is measurable.*

This definition leads to the following basic result:

Theorem A.1.6 (Doob–Dynkin Lemma)**.** *Let f be a function from S to a topological space $\mathcal{Y}$, and let $\sigma(f)$ denote the σ-algebra generated by f. Let $g : S \to \mathbb{R}$ be a measurable function. Then g is $\sigma(f)$-measurable if and only if there exists a Borel measurable function $h : \mathcal{Y} \to \mathbb{R}$ such that*

$$g(s) = h \circ f(s).$$

In order to work with families of random variables, it is convenient to be able to take limits of them. A useful fact is that, for any countable sequence of measurable functions f_n, the pointwise limit $\lim_n f_n$ is also a measurable function (provided it exists). Similarly for $\sup_n f_n$ and $\inf_n f_n$. Unfortunately this does not extend directly to the suprema/infima of uncountable families, see Theorem A.1.20 below.

Once we have a measure μ on a measurable space, we can define integrals. We will assume that μ is σ-finite, that is, there is a sequence of sets $A_n \in \mathcal{F}$ with $\cup_{n\in\mathbb{N}} A_n = S$, such that $\mu(A_n) < \infty$ for all n.

Definition A.1.7. *For a function $f : \mathcal{S} \to \mathbb{R}$, the integral of f is written*[2] *$\int f(x)\mu(\mathrm{d}x)$.*

For simple functions, that is functions of the form $\phi(x) = \sum_{n\in\mathbb{N}} a_n 1_{A_n}(x)$ such that the $A_n \in \mathcal{F}$ are disjoint, and $a_i \in \mathbb{R}^+$, the integral is given by $\int_{\mathcal{S}} \phi(x)\mu(\mathrm{d}x) = \sum_{n\in\mathbb{N}} a_n \mu(A_n)$. For nonnegative functions f^+, the integral is the supremum of the integrals of all simple functions bounded above by f^+. For general functions, the integral is given by $\int f\mu(\mathrm{d}x) = \int \max\{f, 0\}\mu(\mathrm{d}x) - \int \max\{-f, 0\}\mu(\mathrm{d}x)$, provided at least one of these terms is finite.

In the case where we have a probability measure $\mu = \mathbb{P}$, and $S = \Omega$, we often write $\mathbb{E}[f] = \int_\Omega f(\omega)\mathbb{P}(\mathrm{d}\omega)$, and call this the expectation.

Once we have an integral, we quickly obtain a topology over functions.

Definition A.1.8. *We define the metric spaces L^p, which are (equivalence classes of) measurable functions $S \to \mathbb{R}$, with the metric*

$$\|f - g\|_p = \Big(\int_S |f(x) - g(x)|^p \mu(\mathrm{d}x) \Big)^2.$$

This is a metric provided we identify functions where $\|f - g\| = 0$. Such functions are said to be the same almost everywhere*. We say functions are in L^p provided $\|f\|_p < \infty$.*

[1] One can generalize this to having $\mathcal{Y}$ a measure space with an arbitrary σ-algebra, but this is not so interesting

[2] The notation $\int f(x)\mathrm{d}\mu(x)$ is also common.

Using this, we now have various notions of convergence of sequences of functions. We say $f_n \to f$ pointwise if $f_n(x) \to f(x)$ for all x. Weakening this slightly, we say $f_n \to f$ almost everywhere (or almost surely, if we are working with probability spaces), if the set $A = \{x : f_n(x) \not\to f(x)$ satisfies $\mu(A) = 0$. It is possible to check that this agrees with the terminology above, and we often abbreviate it to writing a.e. or a.s. We say that $f_n \to f$ in L^p if $\|f_n - f\|_p \to 0$. Generally speaking, L^p and almost everywhere convergence do not imply each other (but see Theorem A.1.15).

Radon–Nikodym theorem

The key insight of Kolmogorov's axiomatization of probability in terms of measure theory was the definition of the conditional expecation (and the existence of certain continuous processes as a consequence).

Definition A.1.9. *Take a random variable $X \in L^1$ on a probability space. Given a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$, we define the conditional expectation $Y = \mathcal{E}[X|\mathcal{G}]$ to be the (unique, up to equality almost everywhere) random variable such that, for all $A \in \mathcal{G}$, we know $\mathbb{E}[1_A Y] = \mathbb{E}[1_A X]$.*

The existence of the conditional expectation follows from the Riesz representation theorem in Hilbert spaces (for random variables in L^2), along with a convergence theorem for the integral, for example Vitali's convergence theorem below. We should note that the conditional expectation depends on the choice of probability measure $\mathbb{P}$, and is only uniquely defined $\mathbb{P}$-almost everywhere.

The key property of the conditional expectation that is satisfied is the following:

Lemma A.1.10 (The tower law of conditional expectation)**.** *Let $\{\mathcal{F}_t\}_{t\in\mathbb{T}}$ be a filtration on a probability space. Then for any random variable $X \in L^1$, and any $t \le t'$, we know*

$$\mathbb{E}[X|\mathcal{F}_t] = \mathbb{E}\Big[\mathbb{E}[X|\mathcal{F}_{t'}]\Big|\mathcal{F}_t\Big]$$

up to equality almost everywhere.

Bayes' rule

The following inequality is one of the most useful inequalities in this area (together with Cauchy–Schwarz).

Lemma A.1.11 (Jensen's Inequality)**.** *Suppose ϕ is a convex map of $\mathbb{R}$ into $\mathbb{R}$ and suppose X is an integrable random variable such that $\phi \circ X$ is integrable. Let $\mathcal{G}$ be any sub-σ-algebra of $\mathcal{F}$. Then*

$$\phi\big(\mathbb{E}[X|\mathcal{G}]\big) \le \mathbb{E}[\phi \circ X|\mathcal{G}] \quad a.s.$$

Proof. Note that ϕ is the upper envelope of a countable family of affine functions

$$\lambda_n(x) = \alpha_n x + \beta_n, \quad x \in \mathbb{R}, \quad n \in \mathbb{N},$$

that is, $\phi(x) = \sup_n\{\lambda_n(x)\}$. The random variables $\lambda_n \circ X$ are integrable and

$$\lambda_n \circ \mathbb{E}[X|\mathcal{G}] = \mathbb{E}[\lambda_n \circ X|\mathcal{G}] \le E[\phi \circ X|\mathcal{G}] \quad \text{a.s.}$$

Taking the supremum with respect to n, the result follows. □

As convergence almost surely does not usually imply convergence in L^1 (or any other L^p), it is helpful to study cases where this does hold. The required condition is *uniform integrability*.

Definition A.1.12. *Suppose K is a set of random variables. Then K is said to be a* $\mathbb{P}$-uniformly integrable *set if*

$$\int_{\{|X|\geq c\}} |X(\omega)|\, d\mathbb{P}(\omega) = \mathbb{E}\big[1_{|X|\geq c}|X|\big]$$

converges to 0 uniformly in $X \in K$ as $c \to +\infty$.

A convenient reformulation is given be the following result, see [4, Theorem 2.5.4] for a proof.

Theorem A.1.13. *Suppose K is a subset of L^1. Then K is uniformly integrable if and only if both*

(i) there is a number $k < \infty$ such that for all $X \in K$, $\mathbb{E}\big[|X|\big] < k$, and

(ii) for any $\varepsilon > 0$ there is a $\delta > 0$ such that, for all $A \in \mathcal{F}$ with $\mathbb{P}(A) \leq \delta$, we have $\mathbb{E}\big[1_A|X(\omega)|\big] < \varepsilon$ for all $X \in K$.

Lemma A.1.14 (de la Vallée Poussin criterion)**.** *Let K be a set of random variables. Suppose there is a positive function ϕ defined on $[0,\infty)$ such that $\lim_{t\to\infty} t^{-1}\phi(t) = +\infty$ and $\sup_{X\in K} E\big[\phi(|X|)\big] < \infty$. (Common examples are $\phi(x) = x^p$ for $p > 1$, or $\phi(x) = x\log x$.) Then K is uniformly integrable.*

Proof. Write $\lambda = \sup_{X\in K} \mathbb{E}[\phi(|X|)]$ and fix $\varepsilon > 0$. Put $a = \varepsilon^{-1}\lambda$ and choose c large enough that $t^{-1}\phi(t) \geq a$ if $t \geq c$. Then, for $x > c$, we have $x \leq a^{-1}\phi(x)$, so

$$\sup_{X\in K} \mathbb{E}\big[1_{|X|>c}|X|\big] \leq a^{-1} \sup_{X\in K} \mathbb{E}\big[1_{|X|>c}\phi(|X|)\big] \leq a^{-1} \sup_{X\in K} E\big[\phi(|X|)\big] \leq \varepsilon.$$

Taking $c \to \infty$ we see that

$$\limsup_{c\to\infty} \sup_{X\in K} \mathbb{E}\big[1_{|X|>c}|X|\big] \leq \varepsilon$$

and as ε was arbitrary we conclude K is uniformly integrable. □

The power of the uniform integrability condition is due to the following result, which generalizes the dominated convergence theorem. (See [4, Theorem 2.5.8] for a proof.)

Theorem A.1.15 (Vitali convergence theorem)**.** *Suppose $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of integrable random variables which converge in probability to a random variable X. Then the following are equivalent:*

(i) X_n converges to X in L^1 that is $\mathbb{E}\big[|X_n - X|\big] \to 0$ (which easily implies $\mathbb{E}[X_n] \to \mathbb{E}[X]$),

(ii) the collection $K = \{X_n\}_{n\in\mathbb{N}}$ is uniformly integrable.

In either case, the limit X is also integrable.

A key property we make use of is that uniform integrability is not changed by taking conditional expectations.

Theorem A.1.16. *Let K be a uniformly integrable set, and $\mathfrak{G}$ be a (possibly uncountable) family of sub-σ-algebras of $\mathcal{F}$. Then the family of random variables $\{\mathbb{E}[X|\mathcal{G}]\}_{X\in K,\mathcal{G}\in\mathfrak{G}}$ is uniformly integrable.*

Proof. We prove this using Theorem A.1.13. From Jensen's inequality, we know that for any $A \in \mathcal{G}$,

$$\mathbb{E}\big[1_A|\mathbb{E}[X|\mathcal{G}]|\big] \leq \mathbb{E}[1_A|X|] \quad \text{for all } X \in K, \mathcal{G} \in \mathfrak{G}.$$

Setting $A = \Omega$, we obtain a uniform bound on $\mathbb{E}\big[|\mathbb{E}[X|\mathcal{G}]|\big]$. For each $\delta > 0$, let $A_\delta(\mathcal{G})$ be the largest set of the form $\{|\mathbb{E}[X|\mathcal{G}]| > k\}$ such that $\mathbb{P}(A_\delta(\mathcal{G})) \leq \delta$, that is,

$$A_\delta(\mathcal{G}) = \bigcup_{\{k:\mathbb{P}(|\mathbb{E}[X|\mathcal{G}]|>k)\leq\delta\}} \{\omega : |\mathbb{E}[X|\mathcal{G}]| > k\}.$$

Note that $A_\delta(\mathcal{G}) \in \mathcal{G}$ and by construction, for $A \in \mathcal{F}$ with $\mathbb{P}(A) \leq \delta$, we have $\mathbb{E}\big[1_A|\mathbb{E}[X|\mathcal{G}]|\big] \leq \mathbb{E}\big[1_{A_\delta(\mathcal{G})}|\mathbb{E}[X|\mathcal{G}]|\big]$. For any $\varepsilon > 0$, we can find a $\delta > 0$ such that $\mathbb{E}[1_{A_\delta(\mathcal{F})}|X|] < \varepsilon$, and hence, for any $A \in \mathcal{F}$ with $\mathbb{P}(A) \leq \delta$,

$$\begin{aligned}\mathbb{E}\big[1_A|\mathbb{E}[X|\mathcal{G}]|\big] &\leq \mathbb{E}\big[1_{A_\delta(\mathcal{G})}|\mathbb{E}[X|\mathcal{G}]|\big] \leq \mathbb{E}[1_{A_\delta(\mathcal{G})}|X|] \\ &\leq \mathbb{E}[I_{A_\delta(\mathcal{F})}|X|] < \varepsilon\end{aligned}$$

for all $X \in K, \mathcal{G} \in \mathfrak{G}$. By Theorem A.1.13, we see that the family $\{\mathbb{E}[X|\mathcal{G}]\}_{X\in K,\mathcal{G}\in\mathfrak{G}}$ is uniformly integrable. □

Given this structure, we can define some useful types of processes.

Definition A.1.17. *We call functions $X : \Omega \times \mathbb{T} \to \mathcal{Y}$ (for $\mathcal{Y}$ a topological space), random processes. We write $X_t(\omega) = X(\omega, t)$ and often simplify to X_t. We say X is* adapted *if $\omega \mapsto X(\omega, t)$ is $\mathcal{F}_t$-measurable for all t. This is a fairly weak condition though, as it doesn't tell us anything about the regularity of X in time.*

We say X is progressive *if for all $t \in \mathbb{T}$, the restricted map $\Omega \times [0, t] \to \mathcal{Y}; (s, \omega) \mapsto X_s(\omega)$ is $\mathcal{B}([0, t]) \otimes \mathcal{F}_t$-measurable. This ensures regularity with respect to time, as well as ω. If X has continuous paths, that is, $t \mapsto X(t, \omega)$ is continuous for all ω, and X is adapted, then X is progressively measurable (but this is only sufficient, for example having right-continuous or left-continuous paths is also enough).*

Definition A.1.18. *We say an process X is a* submartingale *if, for all $t \leq t'$, we have $X_t \leq \mathbb{E}[X_{t'}|\mathcal{F}_t]$, and $X_t \in L^1$. We say X is a* supermartingale *if $-X$ is a submartingale, and a martingale if it is both a sub- and super-martingale.*

There are many beautiful properties of these processes.

Optional stopping

Martingale convergence

Existence of cadlag modifications.

Doob–Meyer

Lemma A.1.19. *Let Y be an adapted right-continuous process. Then Y is a submartingale if and only if, for all stopping times $\tau \leq \tau'$, we know $\mathbb{E}[Y_\tau] \leq \mathbb{E}[Y_{\tau'}]$. In particular, Y is a martingale if and only if $\mathbb{E}[Y_\tau]$ is constant for all τ.*

Proof. If Y is a submartingale, this is an immediate consequence of the Doob–Meyer decomposition and optional stopping. We prove the converse. Let $t < t'$, and define the set

$$A = \big\{\mathbb{E}[Y_{t'}|\mathcal{F}_t] < Y_t\big\} \in \mathcal{F}_t.$$

Then $\tau = 1_A t + 1_{A^c} t' \leq t'$ is a stopping time, and $Y_{t'} - Y_\tau = 1_A(Y_{t'} - Y_t)$. Therefore,

$$0 \leq \mathbb{E}[Y_{t'} - Y_\tau] = \mathbb{E}[1_A(Y_{t'} - Y_t)] = \mathbb{E}[\underbrace{1_A(\mathbb{E}[Y_{t'}|\mathcal{F}_t] - Y_t)}_{\leq 0}] \leq 0.$$

Therefore, A is a null set and

$$Y_t \leq \mathbb{E}[Y_{t'}|\mathcal{F}_t]$$

that is, Y is a submartingale. The martingale statement follows by considering Y and $-Y$. □

A.1.1 Existence of essential suprema

Theorem A.1.20. *[[4], Theorem 1.3.40] Let (S, Σ, μ) be a σ-finite measure space. Let $\mathcal{F}$ be a (possibly uncountable) collection of Σ-measurable functions. Then there exists a Σ-measurable function f^* such that*

(i) $f^* \geq f$ μ-a.e. for all $f \in \mathcal{F}$,

(ii) $f^* \leq g$ μ-a.e. for all measurable g satisfying '$g \geq f$ μ-a.e. for all $f \in \mathcal{F}$'.

Suppose in addition that $\mathcal{F}$ is directed upwards, that is, for $f, f' \in \mathcal{F}$ there exists $\tilde{f} \in \mathcal{F}$ with $\tilde{f} \geq f \vee f'$ μ-a.e. Then there exists an increasing sequence $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ such that $f^* = \lim_n f_n$ μ-a.e.

We call the function f^* the essential supremum of $\mathcal{F}$, and write $f^* = \operatorname{ess\,sup} \mathcal{F}$. Similarly $\operatorname{ess\,inf} \mathcal{F} = -\operatorname{ess\,sup}\{-\mathcal{F}\}$. If we need to specify the sets involved, we will say that the essential infimum is taken over $\mathcal{F}$, in the Σ-measurable functions, and defined μ-a.e.

Proof. First assume that the functions in $\mathcal{F}$ are uniformly bounded above and μ is finite. If $\mathcal{F}$ is countable, then $f^*(x) := \sup_{f\in\mathcal{F}} f(x)$ is measurable, and satisfies the requirements. Now consider the quantity

$$c := \sup\Big\{ \int_S \Big(\sup_{f\in\mathcal{G}} f(x)\Big) d\mu \,\Big|\, \mathcal{G} \subset \mathcal{F} \text{ countable}\Big\} < \infty.$$

Let $\mathcal{G}_n$ be a sequence of countable subsets of $\mathcal{F}$ approaching the outer supremum, that is, $\int \big(\sup_{f\in\mathcal{G}_n} f(x)\big)d\mu \uparrow c$. Then $\mathcal{G}^* = \cup_n \mathcal{G}_n$ is a countable subset of $\mathcal{F}$ which attains the supremum, that is, $\int \big(\sup_{f\in\mathcal{G}^*} f(x)\big)d\mu = c$. Now let $f^*(x) := \sup_{f\in\mathcal{G}^*}\{f(x)\}$ for every x, and note that f^* is Σ-measurable.

To show this f^* satisfies the requirements of the theorem, observe that if we have $f' \in \mathcal{F}$ with $\mu(\{f' > f^*\}) > 0$ then $\{f'\} \cup \mathcal{G}^*$ is a countable subset of $\mathcal{F}$ and

$$\int_S \Big(\sup_{f\in\{f'\}\cup\mathcal{G}} f(x)\Big) d\mu = \int_S \big(f'(x) \vee f^*(x)\big) d\mu > c$$

giving a contradiction. Furthermore, if g satisfies $g \geq f$ μ-a.e. for all $f \in \mathcal{F}$, then $g(x) \geq \sup_{f\in\mathcal{G}^*} f(x) = f^*$. Finally, if $\mathcal{F}$ is upward directed, then $\mathcal{G}^*$ can be replaced by an increasing sequence of functions, and the result follows.

If the functions are not uniformly bounded, then the monotonic transformation $f(x) \mapsto \arctan(f(x))$ gives a uniformly bounded family. Using this,

$$f^* = \tan(\operatorname*{ess\,sup}_{f\in\mathcal{F}}\{\arctan \circ f)\}$$

gives the essential supremum of the original unbounded family. If μ is not finite but σ-finite, then decomposing into finite sections and constructing the essential supremum on each gives the result. □

A.1.2 Almost supermartingales and stochastic approximation

Here we prove some useful results related to stochastic approximation theory.

Definition A.1.21. *An adapted process X is called an* almost supermartingale *if there exist nonnegative adapted integrable processes A, B, C such that*

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] \le X_t(1+B_t) + A_t - C_t.$$

Theorem A.1.22 (Robbins–Siegmund convergence theorem)**.** *Let X be a nonnegative almost supermartingale. Then $\lim_{t\to\infty} X_t$ exists and is finite, and $\sum_t C_t < \infty$, on the event $\{\omega : \sum_t (A_t + B_t) < \infty\}$.*

Proof. We will rearrange X to construct a supermartingale bounded below, and then apply Doob's convergence theorem. Define

$$X'_t = X_t \prod_{n=1}^{t-1} (1+B_n)^{-1}$$

and

$$A'_t = A_t \prod_{n=1}^{t} (1+B_n)^{-1}, \qquad C'_t = C_t \prod_{n=1}^{t} (1+B_n)^{-1}$$

and then, for some $\beta > 0$,

$$U_t = X'_t - \sum_{n=1}^{t-1} (A'_n - C'_n), \qquad \tau = \min\Big\{t \in \mathbb{N} : \sum_{n\le t} A'_t > \beta\Big\}.$$

We claim that $U^\tau := \{U_{t\wedge\tau}\}_{t\in\mathbb{N}}$ is a supermartingale bounded below. To show this, observe that τ is a stopping time, so U^τ is adapted. As X, A, C are all integrable, and $0 \le Z' \le Z$ for $Z \in \{X, A, C\}$, we see that U is integrable. We also know that, as X is an almost supermartingale,

$$\mathbb{E}[X'_{t+1}|\mathcal{F}_t] = \mathbb{E}[X_{t+1}|\mathcal{F}_t] \prod_{n=1}^{t} (1+B_n)^{-1} \le \Big(X_n(1+B_t) + A_t - C_t\Big) \prod_{n=1}^{t} (1+B_n)^{-1} = X'_t + A'_t - C'_t$$

and hence

$$\begin{aligned}
\mathbb{E}[U^\tau_{t+1}|\mathcal{F}_t] &= U_\tau 1_{t\ge\tau} + \mathbb{E}[U_{t+1}|\mathcal{F}_t]1_{t<\tau} = U_\tau 1_{t\ge\tau} + \mathbb{E}\Big[X'_{t+1} - \sum_{n=1}^{t}(A_n - C_n)\Big|\mathcal{F}_t\Big]1_{t<\tau} \\
&= U_\tau 1_{t\ge\tau} + \Big(\mathbb{E}[X'_{t+1} - A_t + C_t|\mathcal{F}_t] - \sum_{n=1}^{t}(A_n - C_n)\Big)1_{t<\tau} \\
&\le U_\tau 1_{t\ge\tau} + \Big(X'_t - \sum_{n=1}^{t}(A_n - C_n)\Big)1_{t<\tau} = U_\tau 1_{t\ge\tau} + U_t 1_{t<\tau} \\
&= U^\tau_t
\end{aligned}$$

It follows that U^τ is a supermartingale. We also know that X', A', C' are nonnegative, so

$$U^\tau_t \ge U^\tau_t - \sum_{t\le\tau-1} C'_t = X'_t - \sum_{t\le\tau-1} A'_t \ge -\beta.$$

Therefore, as U^τ is a supermartingale bounded below, by Doob's convergence theorem it must converge to a finite limit a.s. In other words, U converges a.s. to a finite limit on the event $\{\tau = \infty\}$.

As $\beta > 0$ was arbitrary (and each defines a corresponding τ_β), we now see that we have convergence of U on the event

$$\bigcup_{\beta>0}\{\tau_\beta = \infty\} = \Big\{\sum_t A'_t < \infty\Big\}.$$

On this event, as X' is nonnegative, we know

$$\sum_{n=1}^{t-1} C'_n - \sum_{n=1}^{t-1} A'_n \leq X'_t - \sum_{n=1}^{t-1}(A'_n - C'_n) = U_t \not\to \infty,$$

so we must also have $\sum_t C'_t < \infty$, and also that X' is convergent to a finite limit.

Finally, observe that

$$0 \leq \prod_{n=1}^{t}(1+B_n) \leq \exp\Big(\sum_{n=1}^{t} B_n\Big)$$

On the event $\{\sum_t (A_t + B_t) < \infty\} \subseteq \{\sum_t B_t < \infty\}$ this remains finite, and hence $X_t = X'_t \prod_{n=1}^{t-1}(1+B_n)$ is convergent and $\sum_t C_t = \sum_t (C'_t \prod_{n=1}^{t}(1+B_n)) < \infty$, as desired. □

This allows us to easily prove the following version of the Robbins–Monro result (essentially due to Dvoretsky (1956)):

Lemma A.1.23. *Consider an adapted random process Y with values in $\mathbb{R}^n$, with dynamics (for each i an index in $\{1, ..., n\}$)*

$$Y_{t+1}(i) = (1 - \alpha_t(i))\beta_t(i)Y_t(i) + \alpha_t(i)\varepsilon_{t+1}(i)$$

where $\alpha_t(i), \beta_t(i) \in [0, 1]$ are adapted, and for all i,

- $\mathbb{E}[\varepsilon_{t+1}(i)|\mathcal{F}_t] = 0$
- $\mathbb{V}[\varepsilon_t(i)|\mathcal{F}_t] \leq c(1 + \|Y_t\|_\infty^2)$ *for $c > 0$.*

Then $Y \to 0$ a.s. on the event

$$\Big\{\sum_{t\in\mathbb{N}} \alpha_t(i) = \infty \text{ and } \sum_{t\in\mathbb{N}} \alpha_t^2(i) < \infty \text{ for all } i\Big\}.$$

Proof. Let $\tau = \min\{t : \|Y_t\|_\infty > k\}$ for some $k > 0$, and consider the stopped process $Y_t^\tau = 1_{t<\tau}Y_t + 1_{t\geq\tau}Y_\tau$. We know that, omitting the argument i for clarity,

$$\begin{aligned}
\mathbb{E}[(Y^\tau_{t+1})^2|\mathcal{F}_t] &\leq (1-\alpha_t)^2\beta_t^2(Y_t^\tau)^2 + 2(1-\alpha_t)\beta_t|(Y_t^\tau)|\alpha_t\mathbb{E}[\varepsilon_{t+1}|\mathcal{F}_t] \\
&\quad + \alpha_t^2\mathbb{V}[\varepsilon^2_{t+1}|\mathcal{F}_t] + \alpha_t^2\mathbb{E}[\varepsilon_{t+1}|\mathcal{F}_t]^2 \\
&\leq (1-\alpha_t)^2(Y_t^\tau)^2 + \alpha_t^2\mathbb{V}[\varepsilon^2_{t+1}|\mathcal{F}_t] \\
&\leq (1-\alpha_t)^2(Y_t^\tau)^2 + \alpha_t^2 c(1 + \|(Y_t^\tau)\|_\infty^2) \\
&\leq (1+\alpha_t^2)(Y_t^\tau)^2 + \alpha_t^2 c(1+k^2) - 2\alpha_t(Y_t^\tau)^2.
\end{aligned}$$

That is, $(Y^\tau(i))^2$ is a nonnegative almost supermartingale, with

$$A_t = (\alpha_t(i))^2 c(1+k^2), \qquad B_t = (\alpha_t(i))^2, \qquad C_t = 2(\alpha_t(i))(Y_t^\tau(i))^2.$$

We immediately see that, on the event of interest, $\sum_t (A_t + B_t) < \infty$, so $(Y_t^\tau(i))^2$ converges to a finite limit, as does

$$\sum_t C_t = 2 \sum_t \alpha_t(i)(Y_t^\tau(i))^2.$$

But as $\sum_t \alpha_t(i) = \infty$, this implies that $(Y_t^\tau(i))^2 \to 0$, and so is bounded for all t. As this holds for all i (and i takes finitely many values), it must be the case that $\|Y_t\|_\infty$ is a.s. bounded. Therefore, taking the union over all $k > 0$ we have the result. □

In order to prove the convergence of Q-learning and related algorithms, a slightly more involved version (due to Jaakkola, Jordan and Singh [6], whose proof is a variant of the approach below) is useful.

Lemma A.1.24 (cf. Lemma 2.3.4)**.** *Consider an adapted random process Y with values in $\mathbb{R}^n$, with dynamics (for each i an index in $\{1, ..., n\}$)*

$$Y_{t+1}(i) = (1 - \alpha_t(i))Y_t(i) + \alpha_t(i)Z_{t+1}(i)$$

where, for all i,

- $\alpha_t(i) \in [0, 1]$, $\sum_{t\in\mathbb{N}} \alpha_t(i) = \infty$, $\sum_{t\in\mathbb{N}} \alpha_t^2(i) < \infty$,
- $|\mathbb{E}[Z_{t+1}(i)|\mathcal{F}_t]| \leq \gamma\|Y_t\|_\infty$, *with* $\gamma < 1$,
- $\mathbb{V}[Z_t(i)|\mathcal{F}_t] \leq c(1 + \|Y_t\|_\infty^2)$ *for* $c > 0$.

Then $\|Y_t\|_\infty \to 0$ a.s. as $t \to \infty$.

Proof. We consider the rescaled process defined by

$$U_{t+1}(i) = (1 - \alpha_t(i))\beta_t U_t(i) + \alpha_t(i)\beta_t Z_{t+1}(i); \quad U_0 = Y_0$$

where $\beta_t = \min\{1, 1/\|U_t\|_\infty\}$, which has the property that $\beta_t U_t \leq 1$. Observe that $U_t = \big(\prod_{0\leq s<t} \beta_s\big)Y_t$. We write $\varepsilon_t = Z_t - \mathbb{E}[Z_t|\mathcal{F}_{t-1}]$. We can then decompose $U_t = \Delta_t + \Gamma_t$, where $\Delta_0 = U_0$, $\Gamma_0 = 0$ and

$$\begin{aligned}\Delta_{t+1}(i) &= (1 - \alpha_t(i))\beta_t\Delta_t(i) + \alpha_t(i)\beta_t\mathbb{E}[Z_{t+1}(i)|\mathcal{F}_t],\\ \Gamma_{t+1}(i) &= (1 - \alpha_t(i))\beta_t\Gamma_t(i) + \alpha_t(i)\beta_t\varepsilon_{t+1}(i).\end{aligned}$$

In particular, applying Lemma A.1.23 to Γ, we see that $\Gamma, \beta\varepsilon \to 0$ a.s.

In order to bound Δ_t, we again fix $k > 0$, and $T > 0$, and define

$$\rho_{T,k} = \min\Big\{t \geq T : \|\Gamma_t\|_\infty > \frac{1-\gamma}{2\gamma}k \text{ or } \|\Delta_t\|_\infty > k\Big\}.$$

Observe that we can write (using our assumption to bound $\mathbb{E}[Z_{t+1}|\mathcal{F}_t]$)

$$|\Delta_{t+1}(i)| \leq (1 - \alpha_t(i))\beta_t|\Delta_t(i)| + \alpha_t(i)\gamma\|\beta_t U_t\|_\infty \leq (1 - \alpha_t(i))|\Delta_t(i)| + \alpha_t(i)\beta_t\gamma(\|\Delta_t\|_\infty + \|\Gamma_t\|_\infty).$$

Define the event

$$A_{T,k} = \Big\{\|\Gamma_t\|_\infty \leq \frac{1-\gamma}{2\gamma}k \text{ for all } t \geq T\Big\} \cap \Big\{\|\Delta_T\|_\infty \leq k\Big\}.$$

On $A_{T,k}$, for $T \leq t < \rho_{T,k}$, by the triangle inequality we know $(\|\Delta_t\|_\infty + \|\Gamma_t\|_\infty) \leq \eta k$, where $\eta = 1 + \frac{1-\gamma}{2\gamma} < \frac{1}{\gamma}$. Thus, as $\beta_t \leq 1$,

$$\begin{aligned} |\Delta_{t+1}(i)| - \gamma\eta k &\leq (1-\alpha_t(i))\Big(|\Delta_t(i)| - \gamma\eta k\Big) \\ &\leq \Big(|\Delta_T(i)| - \gamma\eta k\Big)\prod_{s=T}^{t}(1-\alpha_t(i)) \leq (k-\gamma\eta k)\exp\Big(-\sum_{s=T}^{t}\alpha_t(i)\Big). \end{aligned} \tag{1.1.1}$$

This must hold for all i, which implies that

$$\|\Delta_{t+1}\|_\infty \leq k \text{ on } A_{T,k}, \text{ for all } T \leq t < \rho_{T,k}.$$

However, if $\rho_{T,k}(\omega) < \infty$ for some $\omega \in A_{T,k}$, taking $t = \rho_{T,k} - 1$ it follows that $\Delta_{\rho_{T,k}} \leq k$ and $\Gamma_{\rho_{T,k}} \leq \frac{1-\gamma}{2\gamma}k$, which gives a contradiction. Therefore $\rho_{T,k} = \infty$ on $A_{T,k}$, for all $T > 0$. Coupled with the convergence of Γ and $\beta\varepsilon$, we conclude that the sets

$$\cup_{T,k}\{\rho_{T,k} = \infty\} \subseteq \{\|H_t\varepsilon_t\|_\infty + \|\Gamma_t\|_\infty \to 0\} \subseteq \cup_{T,k}A_{T,k} \subseteq \cup_{T,k}\{\rho_{T,k} = \infty\}$$

are equal and have probability one.

Going back to (1.1.1), on $\cup_T A_{T,k}$ we can take $t \to \infty$ to see that

$$\limsup_{t\to\infty} \|\Delta_t\|_\infty \leq \gamma\eta k < k.$$

But, as $\|\Gamma_t\|_\infty \to 0$ on $\cup_T A_{T,k}$, this implies that $\cup_T A_{T,k} \subseteq \cup_T A_{T,\gamma\eta k}$. Conversely, as $\gamma\eta < 1$, we know $\cup_T A_{T,k} \supseteq \cup_T A_{T,\gamma\eta k}$. Therefore,

$$\Big\{\limsup_{t\to\infty} \|\Delta_t\|_\infty < \infty\Big\} = \cup_{T,k}A_{T,k} = \cap_{k>0}\cup_T A_{T,k} = \Big\{\limsup_{t\to\infty} \|\Delta_t\|_\infty = 0\Big\}.$$

To finish, we observe that we now have shown that $\|\Delta_t\|_\infty + \|\Gamma_t\|_\infty \to 0$, and therefore $\|U_t\|_\infty \to 0$. However, this means that $\beta_t \neq 1$ for a.s. only finitely many t, and so $(\prod_{s<t}\beta_s)$ converges a.s. to a strictly positive value. Consequently, as $U_t = \big(\prod_{s<t}\beta_s\big)Y_t$, we conclude that $\|Y_t\|_\infty \to 0$ a.s. □

A.2 A summary of stochastic calculus

progressive measurability

continuity of integral paths

Itô's lemma

BDG inequality

A.2.1 Lipschitz SDEs

This appendix is taken, with slight modification, from [4, Chapter 16]. Consider an SDE of the form

$$X_t = X_0 + \int_{[0,t]} \mu(\omega, s, X_s)\mathrm{d}s + \int_{[0,t]} \sigma(\omega, s, X_s)\mathrm{d}W_s, \tag{1.2.1}$$

for W a Brownian motion. Throughout this section, we write $\|\cdot\|$ for the Euclidean norm and, by extension, for a matrix $\|A\|^2 = \sum_{ij} A_{ij}$.

Theorem A.2.1. *Let μ and σ be uniformly Lipschitz stochastic functions (that is, maps $\mu : \Omega \times [0,T] \times \mathbb{R}^n \to \mathbb{R}^d$ and $\sigma : \Omega \times [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ with $\|\mu(t,x) - \mu(t,x')\| \le K\|x - x'\|$. Suppose*

$$\int_{[0,T]} \mathbb{E}\big[\|\mu_s(0)\|^2 + \|\sigma_s(0)\|^2\big]\mathrm{d}s < \infty$$

Then (1.2.1) *has a unique*[3] *(strong) solution X, with the predetermined Brownian motion.*

Our method of proof depends on establishing a useful stability result for this equation, under some additional assumptions.

Lemma A.2.2. *Let X be a solution of* (1.2.1) *with μ, σ functions satisfying the linear growth condition*

$$\|\mu_s(x)\| \le \tilde{\mu}_s + K\|x\|, \qquad \|\sigma_s(x)\| \le \tilde{\sigma}_s + K\|x\|,$$

for some constant K and some processes $\tilde{\mu}$ and $\tilde{\sigma}$. (Note that if μ and σ are uniformly Lipschitz, then $\tilde{\mu} = \|\mu(0)\|$ and $\tilde{\sigma} = \|\sigma(0)\|$ satisfy these requirements, with K the Lipschitz constant of the functions.)

Then X is continuous and for any deterministic time T and any $p \ge 2$, there exists a real constant C depending on T, K and p such that

$$\mathbb{E}\Big[\sup_{t\le s\le T} \|X_s\|^p \Big| \mathcal{F}_t\Big] < C\Big(\|X_t\|^p + \int_{[t,T]} \mathbb{E}\big[\|\tilde{\mu}_s\|^p + \|\tilde{\sigma}_s\|^p \big| \mathcal{F}_t\big] ds\Big).$$

Proof. Continuity of X follows immediately from the continuity of the integrals in (1.2.1). If $\|X_t\|^p + \int_{[t,T]} E\big[\|\tilde{\mu}_s\|^p + \|\tilde{\sigma}\|^p \big| \mathcal{F}_t\big] ds = \infty$, then the result is trivial, so we can assume this quantity is finite. In the following, C denotes a constant which can depend on T, K and p, and may vary from line to line. We observe, for $t \le t' \le T$,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{t\le t'\le T} \|X_{t'}\|^p \Big| \mathcal{F}_t\Big] \\
&= \mathbb{E}\Big[\sup_{r\in[t,t']} \Big\|X_t + \int_{[t,r]} \mu_s(X_s)\mathrm{d}s + \int_{[t,r]} \sigma_s(X_s)\mathrm{d}W_s\Big\|^p \Big| \mathcal{F}_t\Big] \\
&\le C\|X_t\|^p + C\int_{[t,t']} \mathbb{E}\big[\|\mu_s(X_s)\|^p \big| \mathcal{F}_t\big]\mathrm{d}s + C\mathbb{E}\Big[\Big(\sup_{r\in[t,t']} \Big|\int_{[t,r]} \sigma_s(X_s)\mathrm{d}W_s\Big|\Big)^p \Big| \mathcal{F}_t\Big] \\
&\le C\|X_t\|^p + C\int_{[t,t']} \mathbb{E}\big[\|\mu_s(X_s)\|^p \big| \mathcal{F}_t\big]\mathrm{d}s + C\mathbb{E}\Big[\Big(\int_{[t,t']} \|\sigma_s(X_s)\|^2]\mathrm{d}s\Big)^{p/2} \big| \mathcal{F}_t\Big] \\
&\le C\|X_t\|^p + C\int_{[t,t']} \mathbb{E}\big[\|\tilde{\mu}_s\|^p + K^p\|X_s\|^p \big| \mathcal{F}_t\big]ds + C\int_{[t,t']} \mathbb{E}\big[\|\tilde{\sigma}\|^p + K^p\|X_s\|^p \big| \mathcal{F}_t\big]\mathrm{d}s \\
&\le C\Big(\|X_t\|^p + \int_{[t,t']} \mathbb{E}\big[\|\tilde{\mu}_s\|^p + \|\tilde{\sigma}\|^p \big| \mathcal{F}_t\big]\mathrm{d}s\Big) + C\int_{[t,t']} \mathbb{E}\big[\sup_{s\le t} \|X_s\|^p \big| \mathcal{F}_t\big]\mathrm{d}s
\end{aligned}$$

where on the third and fifth lines we have used Jensen's inequality, and on the fourth we have used the Burkholder–Davis–Gundy inequality. By Grönwall's inequality, this implies that

$$\sup_{t'\in[t,T]} \|X_{t'}\|^p \le C\Big(\|X_t\|^p + \int_{[t,T]} \mathbb{E}\big[\|\tilde{\mu}_s\|^p + \|\tilde{\sigma}\|^p \big| \mathcal{F}_t\big]\mathrm{d}s\Big)e^{CT} < \infty.$$

Replacing C by Ce^{CT} gives the result. □ □

[3]Here and elsewhere, when stating that an equation has a unique solution, we mean both that a solution exists and that the solution is unique. By a unique strong solution, we mean that it is the only solution adapted to the predetermined filtration in which we pose our problem

One approach to solving SDEs is to apply the above argument to the difference of two SDEs, and then use the resulting estimate to solve the SDE over a short time interval. The existence of a solution for all time follows by pasting. In this setting, we can instead give a more elegant approach using the following, more careful, estimate, which we will also use elsewhere.

Lemma A.2.3. *Let $\mu, \tilde{\mu}, \sigma, \tilde{\sigma}$ be uniformly Lipschitz functions satisfying the conditions of Theorem A.2.1. Let X and $\tilde{X}$ be solutions of (1.2.1) with coefficients (μ, σ) and $(\tilde{\mu}, \tilde{\sigma})$ respectively. For any $\beta \geq 0$,*

$$\begin{aligned}
&\mathbb{E}\big[e^{-\beta T}\|X_T - \tilde{X}_T\|^2\big|\mathcal{F}_t\big] \\
&\leq e^{-(\beta-1-4K^2)(T-t)}\Big(\|X_t - \tilde{X}_t\|^2 + \int_{[t,T]} 2e^{-\beta s}\mathbb{E}[\|\mu_s(X_s) - \tilde{\mu}_s(X_s)\|^2\big|\mathcal{F}_t] \\
&\qquad + 2e^{-2\beta s}\mathbb{E}[\|\sigma_s(X_s) - \tilde{\sigma}_s(X_s)\|^2\big|\mathcal{F}_t]\mathrm{d}s\Big).
\end{aligned}$$

Proof. Write $Y_s = e^{-\beta s}\|X_s - \tilde{X}_s\|^2$. As our processes are continuous, using the Itô product rule we see

$$\begin{aligned}
Y_T = \|X_t - \tilde{X}_t\|^2 &- \beta\int_{[t,T]} e^{-\beta s}\|X_s - \tilde{X}_s\|^2\mathrm{d}s \\
&+ 2\int_{[t,T]} e^{-\beta s}(X_s - \tilde{X}_s)^\top(\mu_s(X_s) - \tilde{\mu}_s(\tilde{X}_s))\mathrm{d}s \\
&+ 2\int_{[t,T]} e^{-\beta s}(X_s - \tilde{X}_s)^\top(\sigma_s(X_s) - \tilde{\sigma}_s(\tilde{X}_s))\mathrm{d}W_s \\
&+ \int_{[t,T]} e^{-2\beta s}\|\sigma_s(X_s) - \tilde{\sigma}_s(\tilde{X}_s)\|^2\mathrm{d}s.
\end{aligned} \tag{1.2.2}$$

Calculating the quadratic variation of Y, we have

$$\langle Y\rangle_t \leq 4\int_{[t,T]} e^{-2\beta s}\|X_s - \tilde{X}_s\|^2\|\sigma(\omega, s, X_s) - \tilde{\sigma}(\omega, s, \tilde{X}_s)\|^2\mathrm{d}s.$$

From Lemma A.2.2, we see that $\mathbb{E}[\sup_{s\in[t,T]}\|X_s - \tilde{X}_s\|^2] < \infty$, so $\mathbb{E}[\int_{[t,T]}\|X_s - \tilde{X}_s\|^2\mathrm{d}s] < \infty$ and

$$\begin{aligned}
\mathbb{E}[\langle Y\rangle_T^{1/2}|\mathcal{F}_t] &\leq 4\mathbb{E}\Big[\Big(\sup_{s\in[t,T]}\|X_s - \tilde{X}_s\|\Big)\Big(\int_{[t,T]}\big(\|\sigma_s(0)\|^2 + K^2\|X_s - \tilde{X}_s\|^2\big)\mathrm{d}s\Big)^{1/2}\Big|\mathcal{F}_t\Big] \\
&\leq 2\mathbb{E}\Big[\Big(\sup_{s\in[t,T]}\|X_s - \tilde{X}_s\|\Big)^2 + \int_{[t,T]}\big(\|\sigma_s(0)\|^2 + K^2\|X_s - \tilde{X}_s\|^2\big)\mathrm{d}s\Big|\mathcal{F}_t\Big] \\
&< \infty.
\end{aligned}$$

By the BDG inequality we see that the '$\mathrm{d}W$' term in (1.2.2) is a true martingale.

Write $\delta\mu_s = \mu_s(X_s) - \tilde{\mu}_s(X_s)$ and $\delta\sigma_s = \sigma_s(X_s) - \tilde{\sigma}_s(X_s)$. Taking an expectation and applying the

Cauchy–Schwarz inequality to (1.2.2), we know that

$$\begin{aligned}
\mathbb{E}[Y_T|\mathcal{F}_t] &\leq Y_t - \beta \int_{[t,T]} \mathbb{E}[Y_s|\mathcal{F}_t]\mathrm{d}s + \int_{[0,t]} \mathbb{E}[Y_s|\mathcal{F}_t]\mathrm{d}s \\
&\quad + \int_{[t,T]} e^{-\beta s}\mathbb{E}[\|\mu_s(X_s) - \tilde{\mu}_s(\tilde{X}_s)\|^2|\mathcal{F}_t]\mathrm{d}s \\
&\quad + \int_{[t,T]} e^{-2\beta s}\mathbb{E}[\|\sigma_s(X_s) - \tilde{\sigma}_s(\tilde{X}_s)\|^2|\mathcal{F}_t]\mathrm{d}s \\
&\leq Y_t - (\beta - 1)\int_{[t,T]} \mathbb{E}[Y_s|\mathcal{F}_t]ds + \int_{[t,T]} \big(2\mathbb{E}[\|\delta\mu_s\|^2|\mathcal{F}_t] + 2K^2\mathbb{E}[Y_s|\mathcal{F}_t]\big)\mathrm{d}s \\
&\quad + \int_{[t,T]} \big(e^{-\beta s}2\mathbb{E}[\|\delta\sigma_s\|^2|\mathcal{F}_t] + 2K^2\mathbb{E}[Y_s|\mathcal{F}_t]\big)\mathrm{d}s \\
&\leq Y_t - (\beta - 1 - 4K^2)\int_{[t,T]} \mathbb{E}[Y_s|\mathcal{F}_t]ds \\
&\quad + \int_{[t,T]} \big(2e^{-\beta s}\mathbb{E}[\|\delta\mu_s\|^2|\mathcal{F}_t] + e^{-2\beta s}2\mathbb{E}[\|\delta\sigma_s\|^2|\mathcal{F}_t]\big)\mathrm{d}s.
\end{aligned}$$

Applying Grönwall's inequality, we conclude

$$\mathbb{E}[Y_T|\mathcal{F}_t] \leq \Big(Y_t + \int_{[t,T]} 2e^{-\beta s}\mathbb{E}[\|\delta\mu_s\|^2|\mathcal{F}_t] + e^{-2\beta s}2\mathbb{E}[\|\delta\sigma_s\|^2|\mathcal{F}_t]\mathrm{d}s\Big)e^{-(\beta-1-4K^2)t}.$$

□

Using this estimate, we now prove existence and uniqueness of the solution.

Proof of Theorem A.2.1. Fix the initial condition $X_0 = x_0$. Consider the map F defined by

$$F(X)_t = x_0 + \int_{[0,t]} \mu_s(X_s)\mathrm{d}s + \int_{[0,t]} \sigma_s(X_s)\mathrm{d}W_s.$$

The process $F(X)$ then satisfies an SDE of the form (1.2.1), with μ and σ which do not depend on $F(X)$. From Lemma A.2.3, taking an expectation we can see that, for any $X, \tilde{X}$ and any $\beta > 0$,

$$\begin{aligned}
&\mathbb{E}[e^{-\beta t}\|F(X)_t - F(\tilde{X})_t\|^2] \\
&\leq 2e^{-(\beta-1)t}\int_{[0,t]} \Big(e^{-\beta s}\mathbb{E}[\|\mu_s(X_s) - \mu_s(\tilde{X}_s)\|^2] \\
&\qquad\qquad + e^{-2\beta s}\mathbb{E}[\|\sigma_s(X_s) - \sigma_s(\tilde{X}_s)\|^2]\Big)ds \\
&\leq 4e^{-(\beta-1)t}\int_{[0,t]} K^2\mathbb{E}[e^{-\beta s}\|X_s - \tilde{X}_s\|^2]ds.
\end{aligned}$$

and hence, by Fubini's theorem,

$$\begin{aligned}
&\int_{[0,T]} \mathbb{E}[e^{-\beta t}\|F(X)_t - F(\tilde{X})_t\|^2]\mathrm{d}t \\
&\leq \int_{[0,T]} 4e^{-(\beta-1)t}\int_{[0,t]} K^2\mathbb{E}[e^{-\beta s}\|X_s - \tilde{X}_s\|^2]\mathrm{d}s\,\mathrm{d}t \\
&\leq \int_{[0,T]} \frac{4K^2}{\beta - 1}\mathbb{E}[e^{-\beta s}\|X_s - \tilde{X}_s\|^2]\mathrm{d}s
\end{aligned}$$

Therefore, for $\beta > 4K^2 + 1$, F is a contraction on the space of progressive processes $X : \Omega \times [0,T] \to \mathbb{R}$, under the norm

$$\|X\|_\beta = \int_{[0,T]} \mathbb{E}[e^{-\beta t}\|X_t\|^2]\mathrm{d}t.$$

As this is simply a weighted L^2 norm, the space is complete. By Banach's fixed point theorem for contractions, we know that there is a unique process which satisfies (1.2.1), up to equality in this norm. By continuity of the integrals, $F(X)$ is continuous, which implies the solution satisfies (1.2.1), and is unique, up to indistinguishability. □

The following lemma is sometimes useful when building approximations.

Lemma A.2.4. *Let X be the solution of an SDE*

$$\mathrm{d}X_t = \mu_t(X_t)\mathrm{d}t + \sigma_t(X_t)\mathrm{d}W_t$$

where W is a Brownian motion, and μ and σ are random Lipschitz functions, as above. For any $\delta > 0$, define a sequence of stopping times by $\tau_0 = 0$ and

$$\tau_{n+1} = \min\Big\{ \inf\{t : \|X_t - X_{\tau_n}\| > \delta\},\, \tau_n + \delta,\, T\Big\}.$$

Then $\tau_n \to T$ almost surely.

Proof. As X has continuous solutions which do not explode (with probability one), its paths are uniformly continuous on $[0,T]$. Therefore, there exists a random variable ε such that $\tau_{n+1} - \tau_n > \varepsilon 1_{\{\tau_{n+1} < T\}}$. The result follows. □

Bibliography

[1] J. Bhandari and D. Russo, Global Optimality Guarantees for Policy Gradient Methods, *Operations Research*, 72:5, https://doi.org/10.1287/opre.2021.0014, 2024

[2] M. Capiński and P.E. Kopp, *Measure, Integral and Probability* (2nd Ed.). Springer, 2004.

[3] M.G. Crandall, H. Ishii, and P.-L. Lions, *user's guide to viscosity solutions of second order partial differential equations*, https://arxiv.org/abs/math/9207212, 1992

[4] S.N. Cohen and R.J. Elliott, *Stochastic Calculus and Applications* (2nd Ed.). Birkhaüser, 2015.

[5] W.H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions* (2nd Ed.). Springer, 2006.

[6] T. Jaakkola, M.I. Jordan, and S.P. Singh, *On the Convergence of Stochastic Iterative Dynamic Programming Algorithms*, MIT AI Memo 1441, https://dspace.mit.edu/handle/1721.1/7205, 1993.

[7] N.V. Krylov, *Controlled Diffusion Processes*. Springer, 1980.

[8] L.C.G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales* (2nd Ed., Volume 1: Foundations), Cambridge University Press, 2000.

[9] D.J. White, Real Applications of Markov Decision Processes, *Interfaces*, 15:6, 1985, http://www.jstor.org/stable/25060766

[10] Y. Ye, The Simplex and Policy-Iteration Methods Are Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate, *Mathematics of Operations Research*, 36:4, https://doi.org/10.1287/moor.1110.0516, 2011