Stochastic Control and Reinforcement Learning
A mathematical introduction

Samuel N. Cohen, Christoph Knochenhauer, Leandro Sánchez-Betancourt

This version: March 9, 2026

Contents

Contents **3**

Introduction **5**

1 Discrete-time Deterministic Control **9**
1.1 Mathematical formulation 10
1.1A Possible optimization approaches 11
1.2 Building a dynamic programming problem 12
1.3 Some examples 14
1.4 Pontryagin's principle 22
1.5 Exercises 24

2 Discrete-time Stochastic Control **33**
2.1 Weak formulation 34
2.1A The martingale principle and dynamic programming . . . 38
2.2 Strong formulation and measurable selection 42
2.3 Some examples 48
2.4 Exercises 54

3 Markov Decision Problems **59**
3.1 Controlled Markov chains 59
3.1A Finite horizon MDPs 60
3.2 Infinite-horizon Discounted MDPs 64
3.2A Ergodic control 66
3.3 Aside: Entropy-regularized control 70
3.4 Exercises 72

4 Classical numerical methods **81**
4.1 Value iteration 81
4.2 Policy iteration 85
4.3 Exercises 90

5 Reinforcement Learning **97**
5.1 Q-learning . . . 98
5.1A SARSA and Actor–Critic methods . . . 101
5.1B TD learning . . . 102
5.2 Policy gradients . . . 104
5.2A Convergence of policy gradients . . . 107
5.3 Exercises . . . 115

6 Continuous-time Deterministic Control **123**
6.1 Notation and problem formulation . . . 123
6.2 Dynamic Programming and Hamilton–Jacobi . . . 126
6.3 Pontryagin's principle . . . 135
6.4 Exercises . . . 142

7 Continuous-time Stochastic Control **155**
7.1 Notation and problem formulation . . . 155
7.2 Dynamic programming . . . 158
7.2A A first set of assumptions . . . 158
7.2B Useful estimates . . . 159
7.2C Describing the value function . . . 165
7.3 Hamilton–Jacobi–Bellman equations . . . 167
7.4 Exercises . . . 175

A Some useful basic theory **185**
A.1 Filtrations, Conditional Expectations, and Martingales . . . 185
A.1A Existence of essential suprema . . . 192
A.1B Almost supermartingales and stochastic approximation . 193
A.2 A summary of stochastic calculus . . . 197
A.2A Lipschitz SDEs . . . 197
A.3 Properties of Viscosity solutions . . . 201

Bibliography **205**

Notation **208**

Introduction

This book grew out of an undergraduate masters course developed by Sam Cohen at the Mathematical Institute, University of Oxford. The book is aimed at mathematicians and it does not assume any prior knowledge on optimal control (deterministic nor stochastic). The book also introduces some of the mathematical results supporting the growing field of reinforcement learning.

The order of authorship was determined alphabeticcally.

We thank Xiaolu Tan, Lingyi Yang and Wojtek Anyszka for comments and pointing out errors in early versions of these notes.

Assumed knowledge

Although we aimed for the book to be self contained, we assume that the reader has some familiarity with:

(i) measure theoretic probability (some background material is in the appendix),

(ii) stochastic differential equations (some background material is in the appendix),

(iii) basic first and second order PDE theory and numerical methods (e.g., finite differences),

(iv) and fundamentals of coding for scientific computing in Python (as needed for the implemented examples we have online).

Notation

We will try and be consistent with notation throughout this book.

(1) For the avoidance of doubt, $0 \notin \mathbb{N}$, but $0 \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

(2) 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.

(3) 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.

(4) 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.

(5) The indicator function will be written $\mathbf{1}_A$, where A is some event or condition (so $\mathbf{1}_A = 1$ if A occurs, and $\mathbf{1}_A = 0$ otherwise).

(6) 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.

(7) 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 ∇ for the column vector with components ∂_{x_i}, so ∇ is the gradient of v . For a differentiable map $F : \mathbb{R}^n \to \mathbb{R}^m$ with $F(x) = \big(F_1(x), \dots, F_m(x)\big)$, we write $D_x F(x) \in \mathbb{R}^{m \times n}$ (or $\partial_x F$) for its Jacobian matrix at x, that is,

$$D_x F = \big(\partial_{x_j} F_i(x)\big)_{1 \leq i \leq m,\ 1 \leq j \leq n}.$$

Similarly, we write $D^2_{xx} v$ for the Hessian of v. Consequently, for a function $v : \mathbb{R}^n \to \mathbb{R}$, we know $\nabla v = (D_x v)^\top$ and $D^2_{xx} v = D_x(\nabla v) = D_x(D_x v)^\top$. Fortunately, we will not need any higher order differentials in this text, so this matrix-vector notation will be sufficient.

(8) The set of symmetric $d \times d$ matrices will be written $\mathbb{S}^{d \times d}$, and we has the partial order: $M \leq N$ if $N - M$ is a positive semi-definite matrix.

(9) The Euclidean norm will be denoted $\|x\|$, the ℓ^∞ norm denoted $\|x\|_\infty = \max_i\{|x_i|\}$, and the (Euclidean) inner products will be denoted either $\langle x, y \rangle$ or $x^\top y$. Unless otherwise stated, for M a matrix, $\|M\|$ will denote the operator norm $\|M\| = \sup_x \|Mx\|/\|x\|$.

(10) 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 sources (in a roughly increasing level of complexity/rigour) are:

1. Sutton and Barto. *Reinforcement Learning: An Introduction (2nd edition).* MIT Press, 2018 [42].
2. Whittle. *Optimal Control: Basics and Beyond.* Wiley, 1996 [46].
3. Meyn. *Control Systems and Reinforcement Learning.* Cambridge University Press, 2022 [37].
4. Bertsekas. *A Course in Reinforcement Learning (2nd Edition).* Athena Scientific, 2024 [7].
5. Szepesvári. *Theoretical Foundations of Reinforcement Learning.* `https://rltheory.github.io/`
6. Puterman. *Markov decision processes: discrete stochastic dynamic programming.* Wiley, 2014 [39].
7. Bensoussan. *Estimation and Control of Dynamical Systems.* Springer, 2018 [6].
8. Pham. *Continuous-time Stochastic Control and Optimization with Financial Applications.* Springer 2009 [38].
9. Bertsekas and Shreve. *Stochastic Optimal Control: The Discrete-time case.* Athena Scientific, 1996 [8].
10. Yong and Zhou. *Stochastic Controls: Hamiltonian Systems and HJB equations.* Springer 1999 [48].
11. Touzi. *Optimal Stochastic Control, Stochastic Target Problems and Backward SDE.* Springer, 2010 [43].
12. Fleming and Soner. *Controlled Markov Processes and Viscosity Solutions.* Springer 2006 [20].
13. Krylov. *Controlled Diffusion Processes.* Springer 1980 [33].

Chapter 1

Discrete-time Deterministic Control

In the first five chapters of the book, we will look at discrete-time optimal control problems. In this chapter we will begin by considering deterministic problems, and then, in the next chapter, introduce randomness in our system.

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 begin with a simple motivating example:

Example 1.0.1 (Lemonade stall)**.** *Suppose you are running a lemonade stall for a week (days $t = 0, 1, ..., 6$). Each day, you choose the quantity of new ingredients to buy, and you set the price of lemonade that day. Depending on the price you charge, you will sell a variable amount of lemonade, which reduces your inventory of ingredients. Your aim is to make the most profit after a week and to keep the inventory level "close" to a given baseline quantity.*

To build a mathematical model of this, we set up the following notation:

(i) Write X_t for the ingredients you have at the start of day t (where, for simplicity, we describe ingredients using a single variable which tells us how many servings we can make). We know the initial inventory $X_0 = x_0$. We call X the state*.*

(ii) On day t, you choose $U_t = (\delta_t, p_t)$, where δ_t is the quantity of ingredients you purchase (you cannot buy a negative amount and there is a maximum amount you can buy), and p_t is the price you charge during the day. We call U the control*.*

(iii) The demand for lemonade we model using a (deterministic) decreasing

function D, where $D(p_t)$ is the number of servings we sell if the price is p_t.

From this, we know that our ingredients satisfy the state dynamics

$$X_{t+1} = X_t - D(p_t) + \delta_t,$$

and we have the practical requirement that $X_t \geq 0$.

We now need to describe our costs:

(i) Denote the cost of ingredients[1] by $C(\delta)$ for a given function C, that is, at time t we will spend $C(\delta_t)$ to buy enough ingredients for δ_t servings.

(ii) We suppose it may be expensive to hold inventory, which is described by a function Γ which adds a cost $\Gamma(X_t)$ at time t.

(iii) These costs are offset by our revenue from sales, which is given by $p_t\, D(p_t)$ (as we sell $D(p_t)$ servings at a price of p_t).

We combine these to give an objective which we want to minimize, that is,

$$J(x_0) = \sum_{t=0}^{6} \Big(C(\delta_t) + \Gamma(X_t) - p_t\, D(p_t) \Big)$$

subject to the requirement that X_t satisfies our dynamics and is nonnegative, and starts at the value $X_0 = x_0$.

We are now ready to state the problem: how can we describe and compute the optimal choice of $U_t = (\delta_t, p_t)$ and the minimal cost.

Motivated by the example we have just seen (which we will return to later), we set up the mathematical notation needed to model a general decision problem.

1.1 Mathematical formulation

Consider a discrete-time finite-horizon model, where time is in $\bar{\mathbb{T}} = \{0, 1, ..., T\}$ for some $T \in \mathbb{N}$; we write $\mathbb{T} = \{0, 1, ..., T-1\}$ for the set of time points at which decisions need to be made.

We suppose we have a *state process* $X = \{X_t\}_{t \in \bar{\mathbb{T}}}$, which describes all (relevant) properties of the world. 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 $\mathcal{U}$. For now we do not make any assumptions about $\mathcal{U}$.

[1] We allow C to be non-linear, which captures the idea that it may be very expensive to purchase a very large quantity of ingredients. When C is quadratic in δ, we obtain a model that is similar to the 'temporary price impact' framework in algorithmic trading; see e.g., the books [14, 24].

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 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.

Remark 1.1.1. 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 $g : \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R} \cup \{\infty\}$, so that $g(t, X_t, U_t)$ represents the cost which the agent must pay at time t, in state X_t, if they choose control U_t. We also include a separate terminal cost $\Phi : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, which depends on the final state. A possible cost of ∞ is included, as this is often convenient to represent strategies which are not permissible. We seek to find an optimal control, that is, a control which minimizes

$$J(x, U) = \sum_{t \in \mathbb{T}} g(t, X_t^U, U_t) + \Phi(X_T^U) \tag{1.1}$$

with respect to $U = \{U_t\}_{s \in \mathbb{T}}$, 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 $t \in \mathbb{T}$) with control U.

Remark 1.1.2. 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 is 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 and U). In a stochastic setting, this is closely related to a Markov assumption (but this is made 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.

1.1A Possible optimization approaches

Now that we have specified our problem, there are a few ways that we could try and resolve the optimization problem.

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

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

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

(ii) We could consider minimizing J with respect to $\{X_t, U_t\}_{t \in \mathbb{T}}$, by treating $X_{t+1} = f(t, X_t, U_t)$ 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 in Section 1.4).

(iii) We can embed our optimization within a family of optimization problems, by considering the behaviour over a single step. This exploits the dynamic nature of our problem, allowing us to reduce our high-dimensional problem (of finding the best control at all times) to a sequence of low-dimensional problems (of finding the control at each time, given the future controls). This will lend itself to stochastic problems as well.

1.2 Building a dynamic programming problem

Instead of just optimizing J in (1.1), we will consider the family of problems given by minimizing the *cost-to-go* (or *remaining-cost*), which we abuse notation and write as

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

where $X^{t,x,U}$ solves the state dynamics with control U and initial value $X_t = x$. With this notation, $J(0, x, U)$ is our original optimization objective in (1.1). If we minimize this, we get the *value function*[2]

$$v(t, x) = \inf_U J(t, x, U).$$

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, for $t < T$,

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

[2] We should note that this terminology is standard in the optimal control literature, but the reinforcement learning community often uses these terms slightly differently, with J being referred to as a value function, and V as its optimized value.

So, with a further abuse of notation

$$J(t,x,U) = g(t,x,U_t) + J\big(t+1, f(t,x,U_t), \{U_s\}_{s\geq t+1}\big). \tag{1.2}$$

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

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

In principle, this allows us to compute v sequentially backward in time t. Using v, we can then identify U_t as the $\arg\min$ in the Bellman equation, which describes the (set of) optimal controls. We know that at the terminal time T, we can just write

$$v(T,x) = \Phi(x).$$

We can then use backward induction to construct $v(T-1,x)$ by solving the Bellman equation, and so on to obtain $v(t,x)$ for all values of t and x. This reduces the problem of finding our optimal control (which involves searching overall possible choices of $\{U_t\}_{t\in\mathbb{T}}$) to a sequence of optimization problems given by the Bellman equation.

This 'principle of optimality' was expressed by Bellman in the following terms:

> Principle of Optimality: An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. — Bellman [?]

We will see (Theorem 2.1.17) that a slightly more nuanced statement is possible, and is needed in order to guarantee that a construction using backward induction will give an optimal solution (see Exercise 1.5.4).

The challenge is now to actually compute the function v defined by this equation. As v is a function of x, whether we can do this depends on how we wish to describe such functions – in special cases we may be able to show that v has a nice structure (for example, is a quadratic), in which case a closed form solution is available. More generally, we have to approximate v numerically, which is a concern we will return to later.

Remark 1.2.1. Even in this simple setting, there are some interesting things to say about dynamic programming, which we will explore in more detail later.

(i) 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 the context.

(ii) 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 (1.2), which ensures that J is monotone with respect to the future cost-to-go – there's no situation where you don't want to minimize tomorrow's costs unless doing so is expensive today.

1.3 Some examples

Example 1.3.1 (Lemonade stall continued)**.** *Let us explore Example 1.0.1 further. We fix $T = 7$ and $\Phi(x) = 0$ (so there is no cost or value in holding inventory at the end of the week).*

Assume the demand for lemonade is piecewise linear and of the form $D(p) = \max\{0, \bar{D} - d\,p\}$ with $\bar{D} = 10$ and $d = 5$. Let the cost per unit of ingredient be linear, so $C(\delta) = \delta(q + r\,\delta)$ for $q = 3$ and $r = 0.5$. The inventory cost function $\Gamma(X_t)$ is taken to be $\Gamma(x) = \gamma\,(x - \bar{x})^2$, where $\gamma = 1$ describes our eagerness to keep inventory close to $\bar{x} = 5$. Figure 1.1 shows the functions $C(\delta)$, $D(p)$, and $\Gamma(x)$.

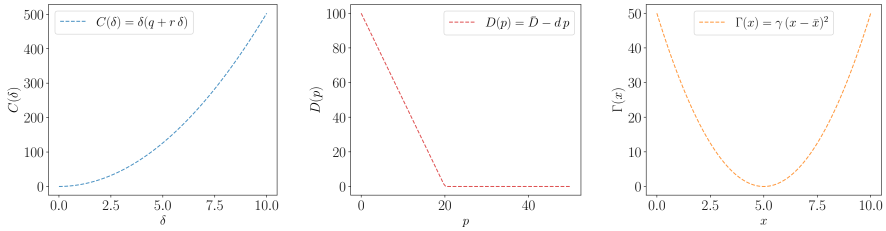


Figure 1.1: Left panel: cost function $C(\delta)$. Middle panel: demand function $D(p)$. Right panel: penalty function $\Gamma(x)$. Model parameters are $q = 3$, $r = 0.5$, $\bar{D} = 10$, $d = 5$, $\gamma = 1$, and $\bar{x} = 5$.

As before, we write $v(t, x)$ for the value function and $U = (\delta, p)$ for the control. We will temporarily ignore the requirement that the inventory is non-

negative, and drop the nonlinearity in demand, and attempt to find an optimal control without these restrictions – provided our demand and inventory stay positive, this will still give the optimal solution for our problem.

Under these simplifications, the functions f and g are given by

$$\begin{aligned} f(t,x,U) &= x-(\bar{D}-d\,p)+\delta\,,\\ g(t,x,U) &= \delta(q+r\,\delta)+\gamma\,(x-\bar{x})^2-p(\bar{D}-d\,p)\,. \end{aligned}$$

From the Bellman equation, we have that for $t\in\mathbb{T}=\{0,1,\ldots,6\}$, the value function satisfies

$$v(t,x)=\inf_{(\delta,p)\in[0,\bar{\delta}]\times\mathbb{R}+}\Big\{\delta(q+r\,\delta)+\gamma(x-\bar{x})^2-p(\bar{D}-d\,p)+v(t+1,x-(\bar{D}-d\,p)+\delta)\Big\},$$

with the terminal value $v(7,x)=\Phi(x)=0$. The challenge is to find the function v. Given the quadratic behaviour of the functions g and f, we employ an ansatz[3]

$$v(t,x)=h_0(t)+h_1(t)\,x+h_2(t)\,x^2$$

where h_0,h_1,h_2 are functions to be determined. Given we used $v(7,x)\equiv 0$, this implies that $h_0(7)=h_1(7)=h_2(7)=0$. Applying the Bellman equation,

$$\begin{aligned} &h_0(t)+h_1(t)\,x+h_2(t)\,x^2\\ &\quad=\inf_{(\delta,p)}\Big\{\delta(q+r\,\delta)+\gamma\,(x-\bar{x})^2-p(\bar{D}-d\,p)+h_0(t+1)\\ &\qquad\qquad+h_1(t+1)\big(x-(\bar{D}-d\,p)+\delta\big)\\ &\qquad\qquad\qquad+h_2(t+1)\big(x-(\bar{D}-d\,p)+\delta\big)^2\Big\}\,. \end{aligned}$$

Using the first order conditions to find the optimizer $U_t=(\delta_t^*,p_t^*)$, we obtain

$$\begin{aligned} p_t^* &= \frac{\bar{D}\,r-d\,r\,h_1(t+1)+(\bar{D}+d\,q+2\,d\,\bar{D}\,r-2\,d\,r\,x)h_2(t+1)}{2\,d(r+(1+d\,r)h_2(t+1))}\,,\\ \delta_t^* &= -\frac{q+h_1(t+1)+(-\bar{D}+d\,q+2\,x)h_2(t+1)}{2(r+(1+d\,r)h_2(t+1))}\,. \end{aligned}$$

Plugging these back in the Bellman equation and taking all terms to one side, we observe that the result remains a quadratic in x (which makes us happy, as it suggests our ansatz was a good guess of the form of the solution). We then combine coefficients of powers of x to obtain an equation of the form

$$0=\{\cdots\}+x\,\{\cdots\}+x^2\{\cdots\}.$$

From here, given that this equation should be zero for all values of x, we conclude that each of the expressions in the curly brackets should be zero and obtain a

[3] That is, an educated guess of the form of v, which we bravely hope will be general enough for us to solve the equation.

system of equations[4] *that characterize* h_0, h_1, *and* h_2. *For example, the equation for* h_2 *is*

$$h_2(t) = \frac{r\,\gamma + (r + \gamma + d\,r\,\gamma)h_2(t+1)}{r + (1 + d\,r)h_2(t+1)}\,,$$

similarly,

$$h_1(t) = \frac{-2\,r\,\bar{x}\,\gamma + r\,h_1(t+1) - (q + \bar{D}r + 2\,\bar{x}\,\gamma + 2\,d\,r\,\bar{x}\,\gamma)h_2(t+1)}{r + (1 + d\,r)h_2(t+1)},$$

and we do not show the value of h_0 *because it is not needed to determine the control. We can then solve these equations backwards (starting from* $h_1(7) = h_2(7) = 0$*); Figure 1.2 shows the optimal strategy (top panels), inventory trajectory (bottom left), and costs (bottom right), for a variety of initial inventory levels. Note that negative costs are profits.*

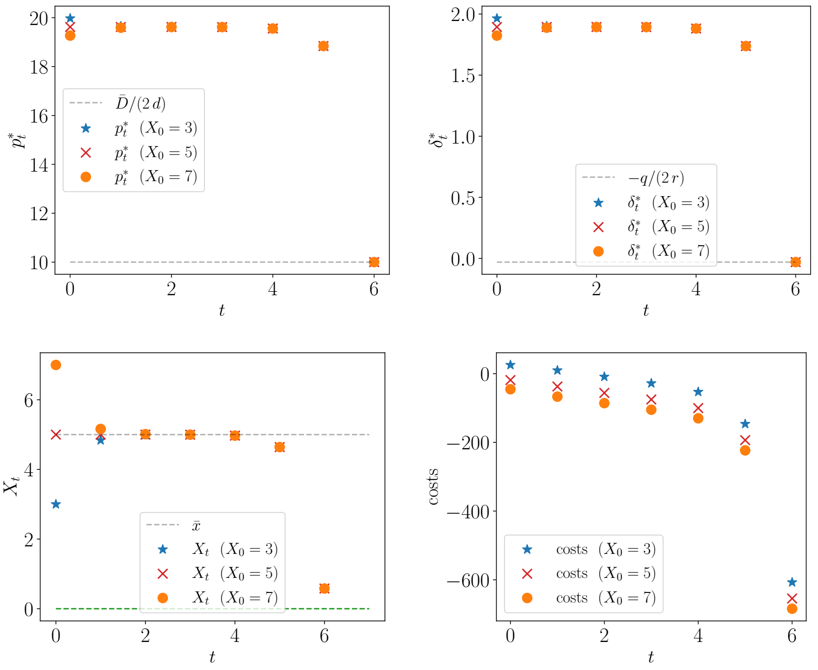


Figure 1.2: Implementation of the lemonade problem. Top panels: optimal strategies (p_t^*, δ_t^*). Bottom left panel: inventory trajectory. Bottom right panel: cumulative costs (negative values are profits).

We observe that in all these scenarios, the inventory is positive at all times, and the price charged is below 20, *and so these requirements (which were omitted*

[4] In the GitHub repository for the book there is a Mathematica notebook with the details.

in our derivation) are automatically satisfied.[5]

We can also consider an abstract version of the problem above, for which we can give a generic closed-form solution.

Example 1.3.2. *Consider the one-dimensional Linear-Quadratic problem, where* $\mathcal{X} = \mathcal{U} = \mathbb{R}$ *and for* $t < T$*,*

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

and

$$\Phi(x) = \pi_T + \rho_T(x - \xi_T)^2.$$

We make an ansatz 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*. The Bellman equation then is*

$$\begin{aligned} v(t,x) = \inf_{U_t} \Big\{ &\underbrace{\alpha + \beta(x-\mu)^2 + \gamma(U_t - \nu)^2}_{g(t,x,U_t)} \\ &\qquad + \underbrace{\pi_{t+1} + \rho_{t+1}\big(a + bx + U_t - \xi_{t+1}\big)^2}_{v(t+1, f(t,x,U_t))} \Big\}. \end{aligned}$$

Basic calculus shows that the optimal strategy is of the form

$$U_t^* = \frac{\gamma\nu + \rho_{t+1}(\xi_{t+1} - a)}{\gamma + \rho_{t+1}} - \frac{b\rho_{t+1}}{\gamma + \rho_{t+1}} x =: h_t + k_t x \tag{1.3}$$

and hence, by substitution in the Bellman equation,

$$\begin{aligned} v(t,x) &= \alpha + \beta(x-\mu)^2 + \gamma(U_t^* - \nu)^2 + \pi_{t+1} + \rho_{t+1}\big(a + bx + U_t^* - \xi_{t+1}\big)^2 \\ &= \alpha + \beta(x-\mu)^2 + \gamma(h_t + k_t x - \nu)^2 \\ &\quad + \pi_{t+1} + \rho_{t+1}\big(a + bx + h_t + k_t x - \xi_{t+1}\big)^2 \\ &= \alpha + \pi_{t+1} + \beta(x-\mu)^2 + k_t^2\gamma\Big(x - \frac{\nu - h_t}{k_t}\Big)^2 \\ &\quad + \rho_{t+1}(b + k_t)^2\Big(x - \frac{\xi_{t+1} - a - h_t}{b + k_t}\Big)^2, \end{aligned}$$

[5] Of course, this is not always the case. In fact, if we take $X_0 = 1$, the value of p_0^* is slightly greater than 20. To deal with such a case, we would have to solve the equations above using an indicator function and treat the various cases carefully.

which one can rearrange to obtain

$$\begin{aligned} v(t,x) = \Big[\beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2\Big] \times \\ \left(x - \frac{\beta\mu + k_t\gamma(\nu - h_t) + \rho_{t+1}(b+k_t)(\xi_{t+1} - a - h_t)}{\beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2}\right)^2 \\ + \frac{\beta\gamma(k_t\mu - \nu + h_t)^2 + \beta\rho_{t+1}((b+k_t)\mu - \xi_{t+1} + a + h_t)^2}{\beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2} \\ + \frac{\gamma\rho_{t+1}(b(\nu - h_t) + k_t(\nu - \xi_{t+1} - a))^2}{\beta + k_t^2\gamma + \rho_{t+1}(b+k_t)^2} + \alpha + \pi_{t+1}\,. \end{aligned}$$

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

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

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

In order to make the above abstract results more concrete, we give another computational example.

Example 1.3.3. *Consider a control problem where the controller wishes to keep X_t as close as possible to $\mu = 5$ and where U_t different from zero induces a cost. The dynamics of the system in this example are*

$$X_{t+1} = 0.5 + 0.5\,X_t + U_t,$$

that is, in the absence of interventions (i.e., if $U_t = 0$ for all t), the dynamics of the system bring X_t back to the equilibrium level $x = 1$. The costs are given by

$$g(t,x,u) = (x-5)^2 + \gamma\,(u-0)^2, \qquad \Phi(x) = 0$$

thus, the controller incurs costs if $U_t \neq \nu = 0$ and if $X_t \neq \mu = 5$. Below, we evaluate the solutions for $X_0 = 0$, $a = b = 0.5$, $T = 20$, $\alpha = 0$, $\beta = 1$, and $\gamma \in \{0.01, 0.5, 1, 2\}$.

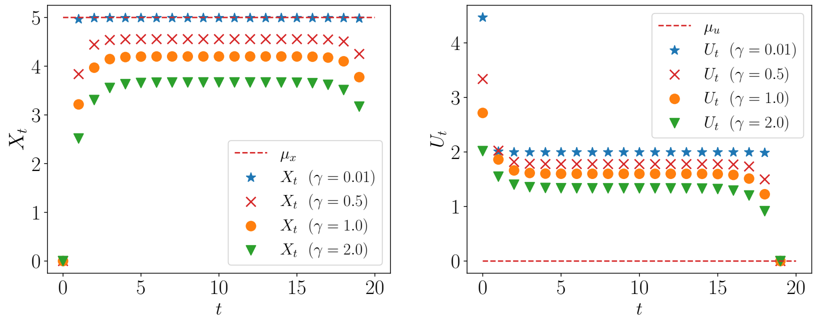


Figure 1.3: Implementation of the one-dimensional deterministic LQ problem. Left panel: optimal trajectory for the state X_t. Right panel: optimal control U_t.

As expected, when the cost parameter γ *is small* ($\gamma = 0.01$ *in the example), the controller takes* X_t *close to* $\mu = 5$ *from the start and compensates any decrease from the mean-reversion to one to keep* X_t *around* 5. *In the graphs we can also appreciate the so-called turnpike phenomenon, where the trajectory of* X *remains fairly constant for most of the time window* $[0, 20]$ *with deviations at the beginning and at the end (see Exercise 3.4.9 for a formal result in this direction with randomness, also the review paper [18]).*

Next, we look at the coefficients h_t *and* k_t *from the optimal control in* (1.3).

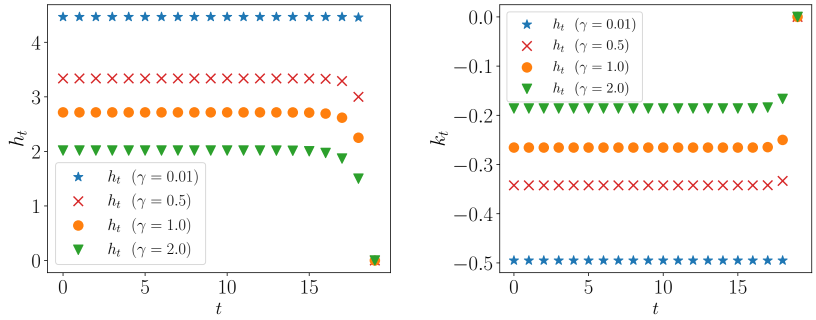


Figure 1.4: Implementation of the one-dimensional deterministic LQ problem. Left panel: trajectory for the auxiliary variable h_t. Right panel: trajectory for the auxiliary variable k_t.

For the trajectory with the least costly controls (the case $\gamma = 0.01$*), we see that the optimal strategy is almost of the form* $U_t \approx 4.5 - 0.5\, X_t$ *making the next step of the state be* $X_{t+1} \approx 0.5 + 0.5\, X_t + 4.5 - 0.5\, X_t = 5$ *for most of the trajectory. As expected, as we approach* T*, the benefit of staying near* $X = 5$ *is reduced, so our controls approach the minimal cost control* $u = 0$.

The following example shows another application of discrete-time deterministic control involving graphs.

Example 1.3.4 (Shortest path in a directed graph)**.** *Consider a finite directed graph, that is, a set of nodes $V = \{1, 2, ..., N\}$ and edges $E \subset V \times V$. We assume that all self-connections are possible, so $(x, x) \in E$ for all $x \in V$. Suppose the graph is connected, that is, for any $x, x' \in V$ there exists $m \in \mathbb{N}$ and a sequence $x = x_0, x_1, x_2, ..., x_m = x'$ with $(x_i, x_{i+1}) \in E$ for all i. We call such a sequence a path from x to x'. Figure 1.5 shows an example of a directed graph, with seven nodes, that is connected and where all self-connections are possible.*

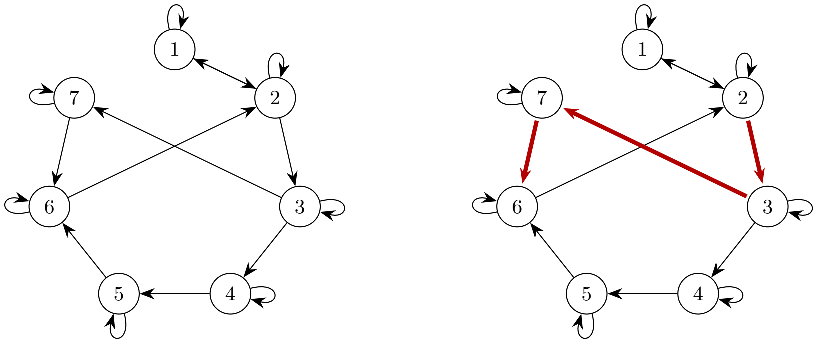


Figure 1.5: Example of a directed graph with seven nodes. The graph is connected and all self-connections are possible. In this example the nodes are $V = \{1, 2, 3, 4, 5, 6, 7\}$, and the edges are $E = \{(1, 1), (1, 2), (2, 1), (2, 2), (2, 3), \ldots, (7, 7)\}$. The right panel highlights the shortest path from $x_0 = 2$ to $x^* = 6$ with length equal to 3.

From such simple setup, we can already infer that for any $x, x' \in V$ there exists a path from x to x' with exactly N steps. To see this, observe that if there exists a path from x to x', then there exists a path without repeated nodes. As there are N nodes in total, we know that a path without repeats will have length at most N. Now allowing repeats in the final node, we see that there must be a path with exactly N steps.

With such a setup a classical problem is to find the shortest path between two nodes; we formulate this in our optimal control notation, leading to a version of the Bellman–Ford algorithm [4].

We define the cost of following a path of length $T \geq N$ to be

$$J(0, x_0, U) = \left(\sum_{t=1}^{T-1} f(X_t, U_t) \right) + \Phi(X_T)$$

where $U_t \in V$ determines the next step in our path (so $X_{t+1}^U = U_t$), $X_0 = x_0 \in$

V, and for a fixed state $x^* \in V$

$$f(x,u) = \begin{cases} 0 & \textit{if } x = x^*, \\ 1 & \textit{if } (x,u) \in E, x \neq x^*, \\ \infty & \textit{if } (x,u) \notin E, \end{cases} \qquad \Phi(x) = \begin{cases} 0 & \textit{if } x = x^*, \\ \infty & \textit{if } x \neq x^*. \end{cases}$$

These costs are chosen so that $\inf_U J(t, x_0, U)$ is the length of the shortest path from x_0 to x^*. More precisely, as there exists a path from x_0 to x^* with at most N steps, and we can repeat the final node with zero cost, we see that $V(0, x_0) < N$. For U minimizing $J(0, x_0, U)$, we know that $\Phi(X_T^U) < \infty$, so $X_T^U = x^*$. In this case, the costs are simply the number of non-repeating steps in the path. The dynamic programming formula then states that the length of the shortest path is $V(0, x_0) = \min_U J(0, U)$, and we have

$$V(t,x) = \min_{u:(x,u)\in E} \Big\{ f(x,u) + V_{t+1}(u) \Big\}$$

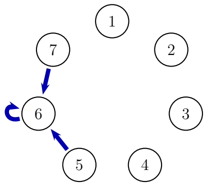
with terminal value $V(T, x) = \Phi$. We also observe that $V(t, x) < N$ if and only if there is a path from x to x^* in at most $T - t$ steps.

Next, we employ the dynamic programming formula above to find the shortest path in the directed graph of Figure 1.5; by inspection it is easy to see that the actual shortest path from $x_0 = 2$ to $x^* = 6$ is the one in red on the right panel of the figure. To solve this mathematically (or computationally) we start by computing $V(T, x)$ for $x \in V$ and $T = 7$. From the terminal condition it follows that

$$V(7,x) = \begin{cases} 0 & \textit{if } x = 6, \\ \infty & \textit{otherwise.} \end{cases} \tag{1.4}$$

Then, from the dynamic programming formula we have that at time $t = 6$ the value function is given by $V(6, x) = \min_{u:(x,u)\in E}\{f(x, u) + V(7, u)\}$. It follows that

$$V_6(x) = \begin{cases} 0 & \text{if } x = 6, \\ 1 & \text{if } x \in \{5,7\}, \\ \infty & \text{otherwise.} \end{cases}$$

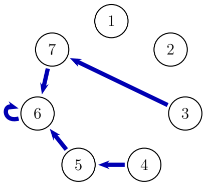


Here we draw (on the right) the set of transitions which are known to be optimal[6] *at time $t = 6$.*

Similarly, using $V(5, x) = \min_{u:(x,u)\in E}\{f(x, u) + V(6, u)\}$ we have

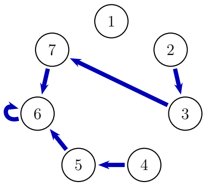
[6] At time $t = 6$, all possible transitions from states in $\{1, 2, 3, 4\}$ are equally bad, and so are not shown.

$$V(5,x) = \begin{cases} 0 & \text{if } x = 6, \\ 1 & \text{if } x \in \{5,7\}, \\ 2 & \text{if } x \in \{3,4\}, \\ \infty & \text{otherwise.} \end{cases}$$



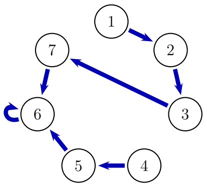
Proceeding in the same way, we have that

$$V(4,x) = \begin{cases} 0 & \text{if } x = 6, \\ 1 & \text{if } x \in \{5,7\}, \\ 2 & \text{if } x \in \{3,4\}, \\ 3 & \text{if } x = 2, \\ \infty & \text{if } x = 1, \end{cases}$$



and lastly, for $t \in \{0,1,2,3\}$

$$V(t,x) = \begin{cases} 0 & \text{if } x = 6, \\ 1 & \text{if } x \in \{5,7\}, \\ 2 & \text{if } x \in \{3,4\}, \\ 3 & \text{if } x = 2, \\ 4 & \text{if } x = 1. \end{cases}$$



Thus, the shortest path from $x_0 = 2$ *to* $x^* = 6$ *has length* $V(0,2) = 3$ *and this is achieved by following the (allowed) path* $x \to x'$ *that satisfies* $V(t,x) = V(t+1,x') - 1$. *In our example this would be* $x_0 = 2 \to 3 \to 7 \to 6$.

In the above example, we might also want to consider the case where, at each time step, the connections available are random. To solve such a problem we will need a little more theory, which we will develop in the next section. We will return to this in Example 2.3.1.

We finish this chapter with a classical result known as Pontryagin principle.

1.4 Pontryagin's principle

In this section we consider a discrete time deterministic control problem, on a finite horizon, where the cost function g and the state dynamics f are both differentiable with respect to the pair (x,u). We also assume that $\mathcal{U} \subset \mathbb{R}^m$ and recall that $\mathcal{X} \subset \mathbb{R}^d$. Below, for any $(t,x,u) \in \mathbb{T} \times \mathbb{R}^d \times \mathbb{R}^m$, we have that $D_x f(t,x,u) \in \mathbb{R}^{d\times d}$ and $D_u f(t,x,u) \in \mathbb{R}^{d\times m}$ are Jacobian matrices. Finally, $\nabla g(t,x,u) = (D_x g(t,x,u))^\top \in \mathbb{R}^{d\times 1}$ and $(D_u g(t,x,u))^\top \in \mathbb{R}^{m\times 1}$. Suppose that the value function is a differentiable function of the state, and there exists a differentiable function $\mathbf{u}^* : \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ such that $\mathbf{u}^*(t,x)$ is an optimal control when $X_t = x$. Given that $\mathbf{u}^*$ is itself differentiable, we let $D_x\mathbf{u}^*(t,x) \in \mathbb{R}^{m\times d}$ be its Jacobian matrix.

Proposition 1.4.1. *Define X_t^* to be the trajectory when following the control $U_t^* = \mathbf{u}^*(t, X_t^*)$. In order for U^* and X^* to be an optimal control–trajectory pair, they must be part of a fixed point to the following forward-backward system of equations:*

$$\begin{aligned} X_{t+1}^* &= f(t, X_t^*, U_t^*), & X_0^* &= x,\\ Q_t &= \nabla g(t, X^*, U_t^*) + \big(D_x f(t, X_t^*, U_t^*)\big)^\top Q_{t+1}, & Q_T &= \nabla\Phi(T, X_T^*),\\ 0 &= \big(D_u g(t, X_t, U_t^*)\big)^\top + \big(D_u f(t, X_t, U_t^*)\big)^\top Q_{t+1}. \end{aligned}$$

Proof. From the Bellman equation we see that

$$V(t,x) = \min_u \left\{ g(t,x,u) + V(t+1, f(t,x,u) \right\}, \tag{1.5}$$

and by hypothesis, $\mathbf{u}^*(t,x)$ minimizes the expression inside the $\min\{\cdot\}$ operator. Thus, using the multivariate chain rule,

$$\Big(D_u g\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top + \Big(D_u f\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top \nabla V\big(t+1, f(t,x,\mathbf{u}^*(t,x)\big) = 0\,. \tag{1.6}$$

Next, observe that (1.5) evaluated at the optimum becomes

$$V(t,x) = g\big(t,x,\mathbf{u}^*(t,x)\big) + V\big(t+1, f(t,x,\mathbf{u}^*(t,x)\big)\,,$$

and differentiating with respect to x (which is possible because of the differentiability assumptions), we obtain

$$\begin{aligned} \nabla V(t,x) = \nabla g\big(t,x,\mathbf{u}^*(t,x)\big) &+ \big(D_x\mathbf{u}^*(t,x)\big)^\top \Big(D_u g\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top\\ &+ \bigg(\Big(D_x f\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top\\ &+ \big(D_x\mathbf{u}^*(t,x)\big)^\top \Big(D_u f\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top\bigg) \nabla V\Big(t+1, f\big(t,x,\mathbf{u}^*(t,x)\big)\Big). \end{aligned}$$

Collecting terms with $\big(D_x\mathbf{u}^*(t,x)\big)^\top$ we see that

$$\begin{aligned} &\nabla V(t,x)\\ &= \nabla g(t,x,\mathbf{u}^*(t,x)) + \Big(D_x f\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top \nabla V(t+1, f(t,x,\mathbf{u}^*(t,x))\\ &\quad + \big(D_x\mathbf{u}^*(t,x)\big)^\top \bigg(\Big(D_u g\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top\\ &\qquad\qquad + \Big(D_u f\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top \nabla V(t+1, f(t,x,\mathbf{u}^*(t,x))\bigg)\,. \end{aligned}$$

Using (1.6) we see that the last term is zero and we obtain

$$\nabla V(t,x) = \nabla g(t,x,\mathbf{u}^*(t,x)) + \Big(D_x f\big(t,x,\mathbf{u}^*(t,x)\big)\Big)^\top \nabla V(t+1, f(t,x,\mathbf{u}^*(t,x))) \,,$$

which yields

$$\nabla V(t,X_t^*) = \nabla g(t,X_t^*,U_t^*) + \Big(D_x f\big(t,X_t^*,U_t^*\big)\Big)^\top \nabla V\big(t+1, f(t,X_t^*,U_t^*)\big) \,, \tag{1.7}$$

substituting $X_t^* = x$ and $U_t^* = \mathbf{u}^*(t,x)$. Set $Q_t = \nabla V(t, X_t^*)$, which is the gradient of the value function evaluated along the optimal trajectory X_t^*. By definition,

$$X_{t+1}^* = f\big(t, X_t^*, \mathbf{u}^*(t, X_t^*)\big) \,,$$

and from the definition of Q_t we have

$$Q_{t+1} = \nabla V(t+1, X_{t+1}^*) = \nabla V(t+1, f(t, X_t^*, U_t^*)),$$

which together with (1.7) means that

$$Q_t = \nabla g(t, X_t^*, U_t^*) + \big(D_x f(t, X_t^*, U_t^*)\big)^\top Q_{t+1} \,,$$

and the terminal condition follows from

$$Q_T = \nabla V(T, X_T^*) = \nabla \Phi(T, X_T^*) \,.$$

Finally, from (1.6) we have that

$$\big(D_u g(t, X_t^*, U_t^*)\big)^\top + \big(D_u f(t, X_t^*, U_t^*)\big)^\top Q_{t+1} = 0 \,.$$

Collecting the results above, we obtain the required forward-backward system which concludes the proof. □

In the Pontryagin principle, Q is called the *adjoint* or *costate* process, and can be interpreted as a marginal value associated with changing the state, or as a Lagrange multiplier arising from treating the state equation as a constraint (see Exercise 1.5.11).

1.5 Exercises

Exercise 1.5.1 (Two-step lemonade problem)**.** *Consider the lemonade stall model of Example 1.0.1 with the parameter choices*

$$\bar{D} = 10, \quad d = 5, \quad q = 3, \; r = 0.5, \quad \gamma = 1, \; \bar{x} = 5,$$
$$C(\delta) = \delta(q + r\delta), \quad \Gamma(x) = \gamma(x - \bar{x})^2,$$

and horizon $T = 2$ (so decisions are made at $t = 0, 1$ and $\Phi \equiv 0$). As before, ignore the nonnegativity and truncation constraints and treat $p_t \in \mathbb{R}, \delta_t \in \mathbb{R}$.

The one-step dynamics and running cost are

$$X_{t+1} = X_t - (\bar{D} - dp_t) + \delta_t, \qquad g(t, x, (\delta, p)) = \delta(q + r\delta) + \gamma(x - \bar{x})^2 - p(\bar{D} - dp).$$

(i) *Using backward induction, solve the control problem at time $t = 1$ for an arbitrary state $X_1 = x$. That is, find the minimizers $(\delta_1^*(x), p_1^*(x))$ of*

$$g(1, x, (\delta, p)) + \Phi(X_2) = \delta(q + r\delta) + \gamma(x - \bar{x})^2 - p(\bar{D} - dp),$$

and write the value $v(1, x)$ in explicit quadratic form $v(1, x) = h_0 + h_1 x + h_2 x^2$, identifying h_0, h_1, h_2.

(ii) *Using the quadratic form found in (i), solve the Bellman problem at time $t = 0$, which is*

$$v(0, x) = \inf_{(\delta_0, p_0)} \Big\{ \delta_0(q+r\delta_0)+\gamma(x-\bar{x})^2-p_0(\bar{D}-dp_0)+v(1, x-(\bar{D}-dp_0)+\delta_0) \Big\}.$$

Show that the problem separates into two scalar quadratic minimizations and derive the first-order conditions that give the optimal $(\delta_0^*(x), p_0^*(x))$. Give the explicit formulas for $\delta_0^*(x)$ and $p_0^*(x)$ in terms of the parameters and the coefficients h_1, h_2 from part (i).

(iii) *Substitute the expressions for $\delta_0^*(x), p_0^*(x)$ back into the objective to compute $v(0, x)$. Show that $v(0, x)$ is a quadratic polynomial in x and identify its quadratic coefficient.*

(iv) *Identify the minimal level of initial inventory X_0 such that the assumption that inventories are nonnegative is satisfied at all times.*

Exercise 1.5.2. *Show that the Bellman equation, together with the terminal value of $v(T, x)$, uniquely defines the value function; that is, if v and v' are both solutions of the Bellman equation with $v(T, \cdot) = v'(T, \cdot)$, then $v(t, x) = v'(t, x)$ for all t and x.*

Exercise 1.5.3. (A 'time-inconsistent' problem, where the DPP fails.) *Consider a deterministic control problem, where at each time t, our agent's preferences are described by a modified cost-to-go function*

$$J(t, x, U) = \sum_{s=t}^{T} \frac{1}{1 + (s - t)} g(s, X_s, U_s),$$

which is sometimes known as hyperbolic discounting. *We assume $f(t, x, u) = x + u$ and $g(t, x, u) = -u + x^2$, where $x_0 \in [0, 1]$ and $u \in \mathcal{U} = [0, 2]$, with a horizon $T = 2$.*

Show that the dynamic programming principle fails, i.e., there are controls U which optimize $J(t, X_t, U)$ but do not optimize $J(t + 1, X_{t+1}, U)$.

Exercise 1.5.4. *In Homer's Odyssey, Odysseus must sail past the island of the Sirens, whose song tempts sailors to steer toward them, leading to shipwreck and death. Knowing this danger in advance, Odysseus considers tying himself to the mast so that, even if he later wishes to steer toward the Sirens, he will be unable to do so.*

We model this as a simple two-stage deterministic decision problem with times $\mathbb{T} = \{0, 1, 2\}$, and will analyse whether it satisfies the dynamic programming principle, in the sense of Theorem 2.1.17.

At time 0*, Odysseus chooses whether to* commit *by tying himself to the mast* $(C = 1)$ *or remain free* $(C = 0)$*. His preferences are described by the following costs:*

- *If $C = 0$, then at time 1 (after hearing the Sirens) he chooses $A \in \{Stay, Leave\}$.*
- *If $C = 1$, then his choice A is irrelevant (as he is unable to leave), but he pays a cost $\alpha \geq 0$.*
- *If he leaves, the ship is destroyed and he incurs loss $L > 0$.*
- *If he stays, he survives with no loss.*

Now suppose that, at time 1, when Odysseus hears the Sirens, he revises his views and evaluates all policies with $A = Stay$ as having cost $\beta \geq 0$, and with $A = Leave$ as having cost 0. In particular, he assigns no cost to losing the ship.

We write $U = (C, A)$ for the possible policies Odysseus can consider, and $J(t, U)$ for the cost he associates with them at time t.

(i) Show that, if $\alpha, \beta > 0$, there is no policy which is optimal at both time 0 and time 1 (i.e. which minimizes both $J(0, U)$ and $J(1, U)$ simultaneously).

(ii) Show that, if $\alpha > 0$, then the policy $C = 1$ is never optimal at time 0, (that is, there is a U such that $J(0, U) < J(0, (1, A))$ for all A).

(iii) Show that, if $\beta = 0$, every policy which is optimal at time 0 is optimal at time 1 (that is, if U minimizes $J(0, U)$ then it also minimizes $J(1, U)$).

(iv) Show that, if $\beta = 0$, there is a policy U^0 which is optimal at time 0, and a policy U^1 which is optimal at time 1, such that the pasted policy $\mathbf{1}_{t=0}U^0 + \mathbf{1}_{t=1}U^1$ is not optimal at time 0. Explain why this may cause problems if you attempt to solve such a problem using dynamic programming.

(v) Explain why, if Odysseus takes into account the decisions he will make at time $t = 1$, he would naturally consider optimizing by choosing U^0 such that the cost of $\mathbf{1}_{t=0}U^0 + \mathbf{1}_{t=1}U^1$ is minimized, where U^1 is an optimal decision from the perspective of time $t = 1$. Explain why, with this additional reasoning, he might be willing to commit $C = 1$ despite the additional cost[7].

[SC]$_1$:add citations

Exercise 1.5.5. *Consider the following general linear-quadratic control problem where the state is n-dimensional and the control is m-dimensional. The state dynamics are linear and given by*

$$X_{t+1} = A\,X_t + B\,U_t, \qquad X_0 = x \in \mathbb{R}^n,$$

[7] This idea develops into 'playing a strategic game against your future self', and was classically explored in the economics literature by [?, ?, ?, ?].

with $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. The cost function g is

$$g(x,u) = x^\top Q\, x + u^\top R\, u,$$

where $Q = Q^\top \succeq 0$, and $R = R^\top \succ 0$. The performance criterion over horizon $T \in \mathbb{N}$ with terminal weight $S = S^\top \succeq 0$ is

$$J(t,x,U) = \sum_{s=t}^{T-1} \Big(X_s^\top Q\, X_s + U_s^\top R\, U_s \Big) + X_T^\top S\, X_T.$$

(i) *Write down the Bellman equation satisfied by $v(t,x) = \inf_U J(t,x,U)$, together with its terminal value $v(T,x)$.*

(ii) *Use the ansatz $v(t,x) = x^\top P_t\, x$ with symmetric matrices $P_t = P_t^\top \succeq 0$ and $P_T = S$ in the Bellman equation from the previous part.*

(iii) *By expanding $x^\top Q\, x + u^\top R\, u + V(t+1, A\, x + B\, u)$ and completing the square in u, derive the optimal control at time t as*

$$U_t^* = -K_t X_t, \qquad K_t := \big(R + B^\top P_{t+1} B\big)^{-1} B^\top P_{t+1} A,$$

with the Riccati backward recursion

$$P_t = Q + A^\top P_{t+1} A - A^\top P_{t+1}\, B\big(R + B^\top P_{t+1}\, B\big)^{-1} B^\top P_{t+1} A, \qquad P_T = S.$$

(iv) *Suppose that S satisfies the algebraic Riccati equation*

$$S = Q + A^\top S A - A^\top S\, B\big(R + B^\top S\, B\big)^{-1} B^\top S A.$$

Show that the value function is independent of time, and that the optimal control is a linear-feedback rule $U_t^* = -K X_t$ for a fixed matrix K.

Remark 1.5.6. The following exercise is a discrete-time version of Pontryagin's principle, which we will return to in Chapter 6. The practical use of this result is that, in many cases, this system of equations can be solved numerically and, if the solution is unique, then this allows us to identify the optimal control process without solving the full Bellman equation.

Exercise 1.5.7. *Consider the scalar deterministic control problem with horizon $T = 2$ and decisions at times $t = 0, 1$. The dynamics of the state process are given by*

$$X_{t+1} = a\, X_t + U_t, \qquad a \in \mathbb{R},$$

and the controller faces quadratic costs

$$g(t,x,u) = (x-\mu)^2 + \frac{r}{2} u^2, \qquad \Phi(x) = q(x-\mu)^2,$$

where $\mu \in \mathbb{R}$, $r > 0$, and $q \geq 0$ are constants. The initial state $X_0 = x_0$ is given.

(i) (Bellman approach) *Apply the results of Exercise 1.5.5 to determine the value function.*

(ii) (Pontryagin approach) *Solve the same problem using the discrete-time Pontryagin minimum principle (as in Section 1.4).*

(iii) Verify that the optimal controls obtained in (i) and (ii) coincide.

Exercise 1.5.8 (Uniform versus trajectory-wise approximation error)**.** *Consider a finite-horizon deterministic control problem with horizon T, dynamics*

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

running cost $g(t,x,u)$, terminal cost $\Phi(x)$, and value function $v(t,x)$ satisfying the Bellman equation

$$v(t,x) = \inf_{u \in \mathcal{U}} \big\{ g(t,x,u) + v(t+1, f(t,x,u)) \big\}, \qquad v(T,x) = \Phi(x).$$

Suppose we are given an approximate value function $\tilde{v}(t,x)$ with $\tilde{v}(T,x) = \Phi(x)$. By "approximate" we mean that there exists $\varepsilon > 0$ such that for all $t < T$ and all $x \in \mathcal{X}$,

$$\left| \tilde{v}(t,x) - \inf_u \Big\{ g(t,x,u) + \tilde{v}\big(t+1, f(t,x,u)\big) \Big\} \right| \le \varepsilon.$$

Lastly, define the greedy policy

$$\tilde{U}_t(x) \in \arg\min_u \Big\{ g(t,x,u) + \tilde{v}\big(t+1, f(t,x,u)\big) \Big\}.$$

(i) Show, by backward induction, that

$$\sup_{x \in \mathcal{X}} |\tilde{v}(t,x) - v(t,x)| \;\le\; (T-t)\,\varepsilon, \qquad t = 0, \dots, T.$$

(ii) Now suppose instead that we only know the following weaker property: for a given initial condition x_0, the approximate Bellman equation holds with error at most ε along the trajectory generated by the greedy policy*, that is,*

$$\left| \tilde{v}(t, X_t^{\tilde{U}}) - \inf_u \Big\{ g(t, X_t^{\tilde{U}}, \tilde{U}) + \tilde{v}\big(t+1, f(t, X_t^{\tilde{U}}, \tilde{U})\big) \Big\} \right| \le \varepsilon$$

for all t along the path $X^{\tilde{U}}$.

Construct a simple deterministic control problem for which this trajectory-wise bound holds with $\varepsilon = 0$, but the resulting policy $\tilde{U}$ is strictly suboptimal at time 0*.*

Remark 1.5.9. The next exercise introduces differential dynamic programming (DDP), which seeks to approximate a general deterministic control problem using a linear–quadratic problem. There are a variety of methods of this type, see for example [**?**], [**?**].

[SC]$_2$:we need to add references here

Exercise 1.5.10 (Linear–Quadratic Approximation)**.** *Consider the finite-horizon deterministic control problem*

$$X_{t+1} = f(t, X_t, U_t), \qquad X_0 = x,$$

with running cost $g(t, x, u)$ and terminal cost $\Phi(x)$. Assume f, g, Φ are twice continuously differentiable. Here we take $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{U} = \mathbb{R}^m$.

Fix a trajectory (which does not need to be optimal) $(X_t^*, U_t^*)_{t=0}^{T-1}$ with terminal state X_T^*. Define perturbations

$$\delta x_t := X_t - X_t^*, \qquad \delta u_t := U_t - U_t^*.$$

(i) To build an approximation, we linearize the dynamics around (X_t^*, U_t^*) as

$$\delta x_{t+1} \approx D_x f(t, X_t^*, U_t^*)\,\delta x_t + D_u f(t, X_t^*, U_t^*)\,\delta u_t,$$

where $D_x f(t, X_t^*, U_t^*) \in \mathbb{R}^{d\times d}$, and $D_u f(t, X_t^*, U_t^*) \in \mathbb{R}^{d\times m}$ are Jacobian matrices.

a. Write the second-order Taylor expansion of the running cost as

$$g(t, X_t, U_t) \approx g_t^* + g_x^\top \delta x_t + g_u^\top \delta u_t + \tfrac{1}{2}\begin{pmatrix}\delta x_t\\ \delta u_t\end{pmatrix}^\top \begin{pmatrix} g_{xx} & g_{xu}\\ g_{ux} & g_{uu}\end{pmatrix}\begin{pmatrix}\delta x_t\\ \delta u_t\end{pmatrix},$$

and state precisely what each term denotes.

b. Assume that at time $t+1$, the value function admits the quadratic local approximation

$$v(t+1, X_{t+1}^* + \delta x_{t+1}) \approx V_{t+1}^* + \left(V_{t+1}^x\right)^\top \delta x_{t+1} + \tfrac{1}{2}\delta x_{t+1}^\top V_{t+1}^{xx}\,\delta x_{t+1}.$$

Substitute the linearized dynamics for δx_{t+1} into this expression.

c. Define $Q_t = g(t, X_t, U_t) + v(t+1, X_{t+1})$, and use the previous part together with the quadratic expansion of g to write Q_t as a function of δx_t and δu_t. More precisely, show that

$$Q_t(\delta x_t, \delta u_t) = Q_t^* + Q_x^\top \delta x_t + Q_u^\top \delta u_t + \tfrac{1}{2}\begin{pmatrix}\delta x_t\\ \delta u_t\end{pmatrix}^\top \begin{pmatrix} Q_{xx} & Q_{xu}\\ Q_{ux} & Q_{uu}\end{pmatrix}\begin{pmatrix}\delta x_t\\ \delta u_t\end{pmatrix},$$

and derive explicit formulas for

$$Q_x, \quad Q_u, \quad Q_{xx}, \quad Q_{xu}, \quad Q_{uu}$$

in terms of derivatives of f, g and V_{t+1}^x, V_{t+1}^{xx}.

d. Assuming Q_{uu} is invertible, minimize $Q_t(\delta x_t, \delta u_t)$ with respect to δu_t to show that

$$\delta u_t^* = k_t + K_t\,\delta x_t,$$

with $k_t = -Q_{uu}^{-1} Q_u^\top$ and $K_t = -Q_{uu}^{-1} Q_{ux}$.

e. *Show that the quadratic approximation of the value function satisfies the backward recursion defined by*

$$\begin{aligned} V_t^{xx} &= Q_{xx} - Q_{xu}Q_{uu}^{-1}Q_{ux}, \\ V_t^{x} &= Q_{x} - Q_{xu}Q_{uu}^{-1}Q_{u}, \end{aligned}$$

with terminal conditions $V_T^x = \Phi_x(X_T^*)$ *and* $V_T^{xx} = \Phi_{xx}(X_T^*)$.

(ii) *Given processes* (k_t, K_t) *and an initial trajectory* (X_t^*, U_t^*), *define an updated trajectory* $(X_t^{\mathrm{new}}, U_t^{\mathrm{new}})$ *through the induction*

$$\begin{aligned} U_t^{new} &= U_t^* + \alpha\Big(k_t + K_t\big(X_t^{new} - X_t^*\big)\Big), \\ X_{t+1}^{\mathrm{new}} &= f(t, X_t^{\mathrm{new}}, U_t^{\mathrm{new}}), \qquad X_0^{\mathrm{new}} = x, \end{aligned}$$

where $0 < \alpha \le 1$ *is a regularization parameter. We can then find determine a new choice of* k_t, K_t *by applying part (i).*

a. *Explain why this process converges in a single backward–forward sweep for the linear-quadratic problem in Exercise 1.5.5*

b. *Explain why this algorithm can be interpreted as repeatedly solving locally approximating linear-quadratic problems.*

(iii) *Recall that, in the discrete-time Pontryagin principle, an optimal trajectory* (X_t^*, U_t^*) *satisfies*

$$\begin{aligned} X_{t+1}^* &= f(t, X_t^*, U_t^*), \qquad X_0^* = x_0 \\ Q_t &= \nabla g(t, X_t^*, U_t^*) + \big(D_x f(t, X_t^*, U_t^*)\big)^\top Q_{t+1}, \qquad Q_T = \nabla\Phi(X_T^*), \\ 0 &= \big(D_u g(t, X_t^*, U_t^*)\big)^\top + \big(D_u f(t, X_t^*, U_t^*)\big)^\top Q_{t+1}. \end{aligned}$$

a. *Show that if the value function is differentiable, then along an optimal trajectory*

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

b. *Compare the Pontryagin first-order condition*

$$0 = (D_u g)^\top + (D_u f)^\top Q_{t+1}$$

with the DDP formula

$$k_t = -Q_{uu}^{-1}Q_u.$$

Show that, at a locally optimal trajectory, we know $Q_u = 0$ *and* $k_t = 0$.

c. *Explain why this algorithm may be interpreted as a Newton-type method applied to the Pontryagin optimality system.*

Exercise 1.5.11 (Deriving Pontryagin's principle via Lagrange multipliers)**.** *Suppose* f *and* g *are continuously differentiable in* (x, u) *and* $\mathcal{U} \subset \mathbb{R}^m$ *is open.*

Define the function

$$\begin{aligned}\mathcal{L} :&(\mathcal{X} \times \mathcal{U} \times \mathbb{R}^m)^T \to \mathbb{R};\\ \{X_t, U_t, Q_t\}_{t<T} &\mapsto \sum_{t=0}^{T-1} g(t, X_t, U_t) + \Phi(X_T)\\ &\qquad + \sum_{t=0}^{T-1} \big(f(t, X_t, U_t) - X_{t+1}\big)^\top Q_{t+1}.\end{aligned}$$

(i) *Show that $\mathcal{L}$ is a Lagrangian for the optimization problem of minimizing $J(U)$ subject to the constraint $X_{t+1} = f(t, X_t, U_t)$ for all $t < T$, where $\{X_t, U_t\}$ are treated as free variables.*

(ii) *Compute the first-order optimality condition obtained by differentiating $\mathcal{L}$ with respect to U_t for $t = 0, \dots, T-1$, and show that any interior minimizer satisfies*

$$0 = \big(D_u g(t, X_t^*, U_t^*)\big)^\top + \big(D_u f(t, X_t^*, U_t^*)\big)^\top Q_{t+1}.$$

(iii) *Compute the first-order condition obtained by differentiating $\mathcal{L}$ with respect to X_t for $t = 1, \dots, T-1$, and show that*

$$Q_t = \nabla g(t, X_t^*, U_t^*) + \big(D_x f(t, X_t^*, U_t^*)\big)^\top Q_{t+1}.$$

(iv) *Differentiate $\mathcal{L}$ with respect to X_T and show that the terminal condition for the multipliers is*

$$Q_T = \nabla\Phi(X_T^*).$$

(v) *Collect the results above and show that any optimal trajectory-control pair (X^*, U^*) must satisfy the forward–backward system*

$$\begin{aligned} X_{t+1}^* &= f(t, X_t^*, U_t^*), && X_0 = x,\\ Q_t &= \nabla g(t, X_t^*, U_t^*) + \big(D_x f(t, X_t^*, U_t^*)\big)^\top Q_{t+1}, && Q_T = \nabla\Phi(X_T^*),\\ 0 &= \big(D_u g(t, X_t^*, U_t^*)\big)^\top + \big(D_u f(t, X_t^*, U_t^*)\big)^\top Q_{t+1}, \end{aligned}$$

which is precisely the discrete-time Pontryagin minimum principle stated in Section 1.4.

Chapter 2

Discrete-time 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. Because of this, we will 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 will also be needed when we take our results into the more difficult (continuous time) setting. Some introductory comments about this theory are in the appendix, but for a more general introduction, the textbook [12] may be of use.

Before going into the details of the formulation, we give a couple of further motivating examples.

Example 2.0.1. *Suppose you wake up one morning and find a frog has appeared in your bathroom. You want to get the frog out of your apartment before you leave for work, but cannot catch it – instead you can encourage the frog to jump to another room, until eventually it leaves through the front door. Every ten minutes, you can move furniture to change the probability that the frog jumps in a given direction, and pay a small cost if the frog continues to be inside (as you want it to leave quickly), and a large cost if you have to leave for work with the frog still inside. The challenge is to find the optimal strategy, which we expect will depend on the frog's current (random) position, and how long you have left until you have to leave.*

Example 2.0.2. *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). At the start of each period, the reward (negative cost) is the number of subscribers 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, 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).*

2.1 Weak formulation

We start with a model of the world described by a filtered space given by $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{T}})$. 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).

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).

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. Because of this, we won't focus on the role of a state variable, but on a more abstract (and generic) class of stochastic control problems.

Remark 2.1.1. In the next chapter, we will focus on the case where we have a state variable X_t described by 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 'weak' formulation we now present.

The alternative 'strong' approach (which we will consider in Section 2.2) is to have our controls directly changing X, but not altering the probability measure at all. As we will see, this will require more measure-theoretic care, as our controls then need to depend on the current value of the state process (which will vary based on the controls used previously), not simply on past information.

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, $U \in \mathcal{U} = \{\text{advertising levels, price levels}\}$.

We write the expectation when using control U as $\mathbb{E}^U$. As we see below, each control U will have a probability measure $\mathbb{P}^U$ associated with it.

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 simplify notation.

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

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

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 2.1.2. *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 $\{X_s\}_{s\leq t}$ (by the Doob–Dynkin lemma).

We note here that we have first fixed the filtration (that is, the flow of information), and then assumed that our strategies must be adapted to this filtration. This simplifies our framework, as it means that the control does not affect the information available, which makes dynamic programming easier to analyse.

We next need to formalize the impact of our control. As indicated above, we work under a 'weak' formulation of the stochastic control problem, that is, the controller affects the probabilities of events, rather than directly changing a state variable. Taking this approach will simplify the exposition quite a bit because we start off with a filtration, which we use to define the class of admissible controls.

Remark 2.1.3. There are natural problems where this approach is insufficient, in particular when we are learning while controlling. In these settings, the choice of control affects the information available in the future, and a more complicated approach is needed. (See, for example, [6, Chapter 9] or [17] and references therein.)

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 2.1.4. *Let $\mathcal{G}$ be a family of measurable functions from Ω into $\mathbb{R}$, on the 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 $g'' \geq g'$ almost everywhere), 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.1.24). 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.

To avoid measure-theoretic issues, we make the following two assumptions.

Assumption 2.1.5. *There is some measure $\mathbb{P}^{\mathrm{ref}}$ such that, for all $U \in \mathbb{U}$, we know that $\mathbb{P}^U(A) = 0$ only if $\mathbb{P}^{\mathrm{ref}}(A) = 0$, that is $\mathbb{P}^U$ is absolutely continuous with respect to $\mathbb{P}^{\mathrm{ref}}$ for all $U \in \mathbb{U}$.*

Definition 2.1.6. *For fixed U, the Radon–Nikodym derivative of $\mathbb{P}^U$ with respect to $\mathbb{P}^{\mathrm{ref}}$ is a random variable, denoted $\frac{\mathrm{d}\mathbb{P}^U}{\mathrm{d}\mathbb{P}^{\mathrm{ref}}}$, such that for all $\mathbb{P}^U$-integrable Z,*

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

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

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

Assumption 2.1.7. *The map $(\{u_t\}_{t\in\mathbb{T}}, \omega) \mapsto \frac{\mathrm{d}\mathbb{P}^U}{\mathrm{d}\mathbb{P}^{\mathrm{ref}}}(\omega)$ (where $U_t = u_t$) is measurable as a map $(\mathcal{U}^{\mathbb{T}}, \Omega) \to \mathbb{R}$.*

Remark 2.1.8. This 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 $\mathbb{P}^{\mathrm{ref}}$-a.e.).

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.

Definition 2.1.9. *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*[1]

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

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

Proposition 2.1.10. *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 $\mathbf{1}_{s\leq t}$, $\mathbf{1}_A$, 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 2.1.11. *We say our control problem is* dynamic *if, for any random variable ξ, with U'' as defined above, for $s \geq t$ we have*

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

and for $s < t$,

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

[1] Clearly, the first of these formulae is only valid if $\mathcal{U}$ is a set on which addition and multiplication by indicator functions is defined. We take this as given, purely for notational convenience, and all our results can be easily reexpressed for general (topological) sets using the latter formula without issues.

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 that our problem is dynamic.

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

Assumption 2.1.12. *The costs g and Φ are such that J satisfies the following bounds:*

(i) for all $t \in \mathbb{T}$ and $U \in \mathbb{U}$, we have that $\mathbb{E}^U[J(t,U)] > -\infty$, i.e., there are no infinitely desirable controls;

(ii) 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.

Remark 2.1.13. So far, our analysis has assumed that the running cost is known at time t, in particular it is given by $g_t(t,u)$ for an adapted function g. However, it is straightforward to include random costs in our setup.

For example, suppose the control at time t results in a cost known only at time $t+1$. We can write $G(t+1,U_t)$ for these costs, and then set $g(t,U_t) := \mathbb{E}^U[G(t+1,U_t)|\mathcal{F}_t]$. Using the tower law, we consider an agent who wishes to minimize

$$\mathbb{E}^U\left[\sum_{t=0}^{T-1} G(t+1,U_t) + \Phi\right] = \mathbb{E}^U\left[\sum_{t=0}^{T-1} g(U_t) + \Phi\right].$$

It is then clear that this is 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.

2.1A 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 2.1.14. *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 2.1.15. *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 $U \in \mathbb{U}$,*

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

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.1.24, 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 2.1.12, 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

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

It follows that

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

□

Theorem 2.1.16 (Bellman equation / Martingale optimality principle)**.** *The value process V satisfies the Bellman equation*

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

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

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

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

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

for U' an arbitrary admissible control. Then we see that, by pasting and Lemma 2.1.15,

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

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

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

so M^{U^*} is a martingale. The converse follows similarly. □

One can show that the martingale optimality principle is, in some sense, a sufficient condition to identify the value process. The proof of this follows exactly as we will see in Theorem 2.2.5, when we consider a strong formulation. 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 2.1.17 (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 $t' \geq t$,*

(i) if U is optimal at t, then U is optimal at t';

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

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

is optimal at t.

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

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

and we see that U is optimal at t if and only if this is an equality, or equivalently, the process $M_s' := \mathbf{1}_{\{s\geq t\}}(M_s^U - M_t^U)$ 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 statement, 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. □

Remark 2.1.18. It's worth pointing out that both of the statements in Theorem 2.1.17 are needed, if we want to construct controls using backward induction, as we did in the deterministic case. To see this, suppose we have a problem with two time steps, $t \in \{0, 1, 2 = T\}$.

When we run backward induction, we first consider time $t = 1$ and find a control U_1 to minimize[2] $J(1, U_1)$ (and we can do this so that U_1 is optimal $\mathbb{P}^{\text{ref}}$-almost everywhere, using an essential infimum, as our controls are closed under pasting). Next, proceeding to time $t = 0$, we freeze the choice of U_1, and construct a control U_0 to optimize $u \mapsto J(0, (u, U_1))$. We need to guarantee that the resulting pair (U_0, U_1) is also optimal at time $t = 0$.

For comparison, we could instead choose a pair $(\tilde{U}_0, \tilde{U}_1)$ which is optimal at $t = 0$, that is, which minimizes $(u_0, u_1) \mapsto J(0, (u_0, u_1))$ among admissible controls. Theorem 2.1.17(i) then implies that $\tilde{U}_1$ is also optimal at time $t = 1$. Furthermore, by Theorem 2.1.17(ii), we are allowed to switch to a different optimal control at time $t = 1$, and hence we can guarantee that the control $(\tilde{U}_0, U_1)$ is optimal at both time $t = 1$ and $t = 0$. This then implies that $\tilde{U}_0$ will also be optimal at $t = 0$ *with* U_1 *fixed*, so freezing U_1 doesn't make our solution at $t = 0$ any worse. Therefore, the pair (U_0, U_1) constructed by backward induction will also be optimal at all times. A concrete example discussing when this fails is given in Exercise 1.5.4.

We have focused in this chapter on the most straightforward version of the optimal control problem, where the controller simply wishes to minimize the expected value of their costs. However, various alternatives are possible, in particular where the controller is concerned by the risk of their position. Some of these are considered in Exercises 1.5.3, 2.4.3, and 2.4.6.

2.2 Strong formulation and measurable selection

In the presentation above, we have assumed that the control works through a 'weak' formulation, where it is the distribution of future costs which is modified by changing the control. In many cases, one also wants to use an explicit state variable, which is directly modified by the control. In this section, we will highlight the key steps in setting up this problem.

[2] Our notation suggests that we should really write $J(1, (U_0, U_1))$, as J is a map from $\mathbb{T} \times \mathbb{U} \to \mathbb{R}$. However, we know that $J(1, (U_0, \cdot))$ does not depend on U_0, so dropping U_0 makes good notational sense.

Remark 2.2.1. Suppose we have an investor deciding how much to invest in a stock. We suppose that the stock values are unaffected by the investor's actions, but can depend on its own past values (for example, if large changes on one day are related to large changes on subsequent days). If the investor chooses not to invest, then their wealth remains constant (with probability one), but if they invest, then it will change randomly through time. The investor makes a new decision of how much to invest at each time point, in order to optimize the utility of their terminal wealth.

In this setting, the underlying probability space describes the values of the stock, so we cannot typically model the effects of investment decisions on wealth by changing probabilities, as in a weak formulation – we need both the right dynamics of the stock price (which is unaffected by investment decisions) and wealth (which is affected by investment decisions). This leads us to seek a strong formulation of this problem, despite this requiring a little more technical detail.

In order to set up a strong formulation, we will assume $\mathbb{T} = \{0, 1, ..., T\}$, we have a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}_{t\in\mathbb{T}})$ and we have a state process $X \in \mathcal{X} \subset \mathbb{R}^d$ with controlled dynamics

$$X_{t+1} = f(\omega, t, X_t, u) \qquad \text{for } u \in \mathcal{U}$$

We use similar notation to the deterministic case, as $f : \Omega \times \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ gives our state dynamics.

Remark 2.2.2. One way of seeing the difference between the strong and weak formulation of the problem is through the dependence of the control on the state. In the weak formulation, any 'state variable' isn't directly affected by the controls (as only the probabilities are changed). In particular, this means that all the required information to make decisions is available in the filtration $\mathcal{F}_t$. In the strong formulation things are a little different, as the state $X_t^{0,x_0,\mathbf{u}}$ depends not only on the filtration $\mathcal{F}_t$, but also on the control that we have used prior to time t. This means that we expect to need the additional information of the current value $X_t = x$ in order to make good decisions.

We define admissible controls essentially as in Definition 2.1.2, however we now also allow our control to have explicit (measurable) dependence on the state variable X, as well as on ω.

Definition 2.2.3. *A control* $\mathbf{u} : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ *is said to be* admissible *if* $(\omega, x) \to u(\omega, t, x)$ *is* $\mathcal{F}_t \otimes \mathcal{B}(\mathcal{X})$*-measurable for all* $t \in \mathbb{T}$. *We write* $\mathbb{U}$ *for the set of admissible controls.*

We write $X^{t,x,\mathbf{u}}$ for the state initialized with $X_t = x$, and following a control $\mathbf{u} \in \mathbb{U}$, that is,

$$X_{s+1}^{t,x,\mathbf{u}}(\omega) = f\Big(\omega, t, X_s^{t,x,u}(\omega), \mathbf{u}\big(\omega, s, X_s^{t,x,u}(\omega)\big)\Big).$$

Remark 2.2.4. To avoid confusion, we will use the notational convention that

(i) $u \in \mathcal{U}$ is a general control value we could choose,

(ii) $U_t = \mathbf{u}(\omega, t, X_t^{0,x,\mathbf{u}})$ is the realized control given our state variable at each time (so $U : \Omega \times \mathbb{T} \to \mathcal{U}$), and

(iii) $\mathbf{u} : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ is the (stochastic) *policy*, that is, the map that determines the control given the state X.

We will often omit the explicit dependence on ω for notational simplicity.

As before, we seek to minimize the total cost

$$J(x, \mathbf{u}) = \mathbb{E}\Big[\sum_{t=0}^{T-1} g(t, X_t^{0,x,\mathbf{u}}, U_t) + \Phi(X_T^{0,x,\mathbf{u}})\Big].$$

We note that, as X depends on ω, we assume that our control impacts X directly, rather than also changing the probabilities of different ω.

As before, for an admissible control, we define the cost-to-go

$$\begin{aligned} J(\omega, t, x, \mathbf{u}) &:= \mathbb{E}\Big[\sum_{s=t}^{T-1} g(\omega, s, X_s^{t,x,\mathbf{u}}, U_s) + \Phi(\omega, X_T^{t,x,\mathbf{u}})\Big|\mathcal{F}_t\Big] \\ &= g(\omega, t, x, U_t) + \mathbb{E}\Big[J\Big(\omega, t+1, f(\omega, t, x, U_t), \mathbf{u}\Big)\Big|\mathcal{F}_t\Big]. \end{aligned}$$

Given this setting, we can give the following sufficient condition for optimality of a control, based on the martingale principle. We highlight that this result requires very few continuity or compactness conditions on our problem, however it assumes that we can guess a value function and a candidate optimal control.

Theorem 2.2.5 (Martingale verification)**.** *Consider a random field $v : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathbb{R}$ which is adapted with respect to (ω, t), Borel measurable with respect to x, and satisfies*

$$v(\omega, T, x) = \Phi(\omega, x) \text{ for all } x, \text{ almost surely.}$$

For each $(t, x) \in \mathbb{T} \times \mathcal{X}$ and each admissible control $\mathbf{u}$, with $U_s = \mathbf{u}(s, X_s^{(t,x,\mathbf{u})})$, define the process

$$M_{t'}^{t,x,\mathbf{u}}(\omega) = \sum_{s=t}^{t'-1} g(\omega, s, X_s^{t,x,\mathbf{u}}, U_s) + v(\omega, t', X_{t'}^{t,x,\mathbf{u}}).$$

Suppose that for every $(t, x, \mathbf{u})$, the process $M^{t,x,U}$ is a submartingale. Suppose also that for each (t, x) there exists some $\mathbf{u}^{(t,x)}$ such that $M^{t,x,\mathbf{u}^{(t,x)}}$ is a martingale. Then

$$v(\omega, t, x) = \operatorname*{ess\,inf}_{\mathbf{u} \in \mathbb{U}} \Big\{\mathbb{E}\Big[\sum_{s=0}^{T-1} g(t, X_t^{t,x,\mathbf{u}}, U_t) + \Phi(X_T^{t,x,\mathbf{u}})\Big|\mathcal{F}_t\Big]\Big\}$$

(so v is a version of the value function), and $\mathbf{u}^{(t,x)}$ is an optimal control for the problem started at (t, x).

Proof. For any $U \in \mathbb{U}$, as $M^{t,x,U}$ is a submartingale, we know that

$$\begin{aligned} v(\omega, t, x) &= M_t^{t,x,U} \\ &\leq \mathbb{E}[M_T^{t,x,U}|\mathcal{F}_t] = \mathbb{E}\Big[\sum_{s=t}^{T-1} g(s, X_s^{t,x,U}, U_s) + \Phi(X_T^{t,x,U})\Big|\mathcal{F}_t\Big]. \end{aligned}$$

On the other hand, we know that there exists $\mathbf{u}^{(t,x)}$ such that

$$v(\omega, t, x) = M_t^{t,x,\mathbf{u}^{(t,x)}} = \mathbb{E}\Big[\sum_{s=t}^{T-1} g(s, X_s^{t,x,U^{(t,x)}}, \mathbf{u}_s^{(t,x)}) + \Phi(X_T^{t,x,U^{(t,x)}})\Big|\mathcal{F}_t\Big].$$

It follows immediately that v is a version of the desired essential infimum, and that $\mathbf{u}^{(t,x)}$ achieves at least as low an expected cost (given $\mathcal{F}_t$, almost surely) as any other admissible policy, and hence is optimal for the problem started at (t, x). □

While these definitions give a well posed problem, the measurability requirement on U causes some difficulties when we seek to do backward induction. This is because we have a tension between finding the optimal $u \in \mathcal{U}$ (for a given t, x), and making sure we have enough measurability in x that $X^{t,x,\mathbf{u}}$ is a well defined process (which requires us to consider multiple x's simultaneously). In order to give a rigorous construction, we make the following assumption.

Assumption 2.2.6. *We make the following technical assumptions:*

(i) For all $x \in \mathcal{X}$ and $\mathbf{u} \in \mathbb{U}$, the random variables $g(\omega, t, X_t^{0,x,\mathbf{u}})$ and $\Phi(\omega, X_T^{0,x,\mathbf{u}})$ are integrable;

(ii) $\mathcal{U}$ is a compact subset of a metric space;

(iii) f, g and Φ are Lipschitz continuous in x, uniformly in (ω, t, u), and f and g are continuous in u, uniformly in (ω, t, x).

This is a stronger assumption than is truly needed, but allows us to apply a 'measurable selection theorem', in order to ensure measurability of our controls. There are many such theorems (see for example the review of Graf [23], or a good book on measure theory, such as Bogachev [10]), however one which is convenient for our purposes is as follows:

Theorem 2.2.7 (Filippov's Implicit Function Theorem[3])**.** *Let $(M, \mathcal{M})$ be a measurable space, A a separable metric space, and $\mathcal{U}$ a compact metrizable space (or more generally, the union of countably many compact metrizable sets). Consider a function $k : M \times \mathcal{U} \to A$ such that $k(\cdot, u)$ is $\mathcal{M}$-measurable for every $u \in \mathcal{U}$ and $k(m, \cdot)$ is continuous for every $m \in M$. Let $y : M \to A$ be an $\mathcal{M}$-measurable map such that,*

$$\textit{for all } m \in M, \textit{ there exists } u \in \mathcal{U} \textit{ such that } y(m) = k(m, u).$$

[3] A proof of this can be found in Appendix A.10 of [16], and is based on Beneš [5] and McShane and Warfield [36], generalizing results of Filippov [19].

Then there exists an $\mathcal{M}$-measurable map $u : M \to U$ such that

$$y(m) = k(m, u(m)).$$

Using this result, we can now obtain existence of an optimal control.

Theorem 2.2.8. *In the situation described, there exist functions $u^* : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ and $v : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathbb{R}$ such that*

(i) v is Lipschitz with respect to x, uniformly in (ω, t);

(ii) $\mathbf{u}^*$ is adapted with respect to (ω, t) for every x, and (Borel) measurable with respect to x for every (ω, t);

(iii) the cost-to-go is minimized by the policy $U^*_s = \mathbf{u}^*(\omega, s, X_s^{t,x,\mathbf{u}^*})$, and v is its minimal value, that is, for every t, x we have the almost-sure equality

$$J(\omega, t, x, \mathbf{u}^*) = v(\omega, t, x) := \operatorname*{ess\,inf}_{\mathbf{u} \in \mathbb{U}} J(\omega, t, x, \mathbf{u}).$$

(iv) v satisfies the Bellman equation

$$v(\omega, t, x) = \operatorname*{ess\,inf}_{u \in \mathcal{U}} \Big\{ g(\omega, t, x, u) + \mathbb{E}\Big[v\Big(\omega, t+1, f(\omega, t, x, u)\Big) \Big| \mathcal{F}_t \Big] \Big\}$$

with terminal value

$$v(\omega, T, x) = \Phi(\omega, x).$$

Proof. As is now familiar, we will work by backward induction. Fix some $t < T$ and suppose we have a construction of v and u^* for all $s > t$.

We first need to define $v(\omega, t, \cdot)$. Write

$$\tilde{J}_t(\omega, x, u) = \mathbb{E}\Big[g(\cdot, t, x, u) + v\Big(\cdot, t+1, f(\cdot, t, x, u)\Big) \Big| \mathcal{F}_t \Big]$$

for the cost-to-go when we vary only the immediate control $U_t = u$, and are optimal after then. As $x \mapsto v(\omega, t+1, x)$ is Lipschitz continuous, and compositions, sums and expectations of (uniformly) Lipschitz functions are Lipschitz, we know that $\tilde{J}$ is also Lipschitz with respect to x, and continuous with respect to u.

We set $M = \Omega \times \mathcal{X}$ with the σ-algebra $\mathcal{M} = \mathcal{F}_t \otimes \mathcal{B}(X)$ and $A = \mathbb{R}$. Given that $\tilde{J}_t$ is a continuous function of u, is Lipschitz in x and is $\mathcal{F}_t$-measurable in ω, we know that

$$v(\omega, t, x) := \operatorname*{ess\,inf}_{u \in \mathcal{U}} \tilde{J}_t(\omega, x, u)$$

is $\mathcal{M}$-measurable for each t; that is, we take the family $\{\tilde{J}_t(\cdot, \cdot, u)\}_{u \in \mathcal{U}}$ as a set of $\mathcal{M}$-measurable functions, and apply the definition to construct an $\mathcal{M}$-measurable essential infimum, which minimizes $d\mathbb{P} \times \mathrm{d}x$-almost everywhere.

As our set of controls $\mathcal{U}$ is compact, we know that this essential infimum is attainable, for every (ω, x). Therefore, applying Filippov's implicit function theorem, we obtain an $\mathcal{M}$-measurable map $\mathbf{u}^*_t : \Omega \times \mathcal{X} \to \mathcal{U}$ such that

$$v(\omega, t, x) = \tilde{J}_t\big(\omega, x, \mathbf{u}^*_t(\omega, x)\big).$$

We now compute, with k the appropriate Lipschitz constant,

$$\begin{aligned} v(\omega,t,x) - v(\omega,t,x') &= \underbrace{\tilde{J}_t\big(\omega,x,\mathbf{u}_t^*(\omega,x)\big) - \tilde{J}_t\big(\omega,x,\mathbf{u}_t^*(\omega,x')\big)}_{\leq 0} \\ &\quad + \underbrace{\tilde{J}_t\big(\omega,x,\mathbf{u}_t^*(\omega,x')\big) - \tilde{J}_t\big(\omega,x',\mathbf{u}_t^*(\omega,x')\big)}_{\leq k|x-x'|} \\ &\leq k|x-x'| \end{aligned}$$

and by reversing x and x', we see that v is also Lipschitz continuous[4] in x.

Finally, we complete this construction and define $\mathbf{u}^*(\omega,t,x) = \mathbf{u}_t^*(\omega,x)$ for each t. This is clearly an admissible control, and satisfies

$$\tilde{J}_t\big(\omega,x,\mathbf{u}_t^*(\omega,x)\big) = \operatorname*{ess\,inf}_{\mathbf{u}\in\mathbb{U}} \tilde{J}_t\big(\omega,x,\mathbf{u}(\omega,t,x)\big),$$

(as the right hand side does not depend on $\mathbf{u}(\omega,s,x)$ for any $s \neq t$). To conclude, we see that the optimal value for our problem is, for any t,

$$\begin{aligned} &\operatorname*{ess\,inf}_{\mathbf{u}\in\mathbb{U}} J(\omega,t,x,\mathbf{u}) \\ &= \operatorname*{ess\,inf}_{\mathbf{u}\in\mathbb{U}} \Big\{ g(\omega,t,x,\mathbf{u}(t,x)) + \mathbb{E}\Big[J\Big(\omega,t+1,f(\omega,t,x,\mathbf{u}(t,x)),\mathbf{u}\Big)\Big|\mathcal{F}_t\Big]\Big\} \\ &\geq \operatorname*{ess\,inf}_{\mathbf{u}\in\mathbb{U}} \Big\{ g(\omega,t,x,\mathbf{u}(t,x)) + \mathbb{E}\Big[v\Big(\omega,t+1,f(\omega,t,x,\mathbf{u}(t,x)),\mathbf{u}\Big)\Big|\mathcal{F}_t\Big]\Big\} \\ &= v(\omega,t,x) \\ &= \tilde{J}_t\big(\omega,x,\mathbf{u}_t^*(\omega,x)\big) \\ &= g(\omega,t,x,\mathbf{u}_t^*(\omega,x)) + \mathbb{E}\Big[J\Big(\omega,t+1,f(\omega,t,x,\mathbf{u}_t^*(\omega,x)),\mathbf{u}^*\Big)\Big|\mathcal{F}_t\Big] \\ &= J(\omega,t,x,\mathbf{u}^*). \end{aligned}$$

This shows that this $\mathbf{u}^*$ is indeed the optimal control and v is the optimal value function. □

Remark 2.2.9. In the case of a finite-state Markov decision process, much of the above discussion can be avoided as x takes only finitely many values (so all functions of x are measurable).

Remark 2.2.10. Given the above result, it is straightfoward to prove the martingale optimality principle is necessary (as before), that is, that the process

$$M_t^{\mathbf{u}} = \Big(\sum_{s=0}^{t-1} g(\omega,t,X_t^{\mathbf{u}},U_t)\Big) + v(\omega,t,X_t^{\mathbf{u}})$$

is a submartingale for all $\mathbf{u}\in\mathbb{U}$, and is a martingale if and only if $\mathbf{u}$ is optimal. Consequently, the dynamic programming principle also holds.

[4] You may wonder about whether we can really do this, as v is only defined almost everywhere when we take the essential infimum. If we wish to be really careful, this argument shows that we can define v on a dense set (say, points with rational coordinates in $\mathcal{X}$) then extend it by continuity to all of $\mathcal{X}$, using Kolmogorov's continuity theorem. This gives a Lipschitz continuous version of the desired function.

2.3 Some examples

Example 2.3.1 (Shortest expected path in directed graph)**.** *Recall the setup in Example 1.3.4. Now suppose that at each time t, we randomize whether a connection x, x' is available or not. That is, for each t and $(x, x') \in E$, there is a Bernoulli random variable $Y_{t,x,x'}$, such that*

(i) with probability p, the path from x to x' is available and we represent it by $Y_{t,x,x'} = 1$, and

(ii) with probability $1-p$ it is not available, which we represent with $Y_{t,x,x'} = 0$.

Assume all edges are independent, and whether an edge is available (at time t) is known only at time t. To find the expected shortest path we need to modify the cost function f. If we look for paths of length $N - t$ or less (and we place a penalty of $\Phi(x) = N/p$ on any path which does not reach x^* in at most N steps, rather than the infinite cost we used in Example 1.3.4), the dynamic programming equation becomes

$$V_t(x, \{y_{t,x,x'}\}_{x' \in V}) = \min_{u: y_{t,x,u}=1} \Big\{ f(t, x, u) + \mathbb{E}[V_{t+1}(u, \{Y_{t+1,u,x'}\}_{x' \in V})] \Big\}$$

because we know the Y's are independent, so we don't need conditional expectations. Here we used the convention $Y_{t,x,x'} = y_{t,x,x'} = 0$ for any $(x, x') \notin E$. This gives us our dynamic programming formalism, but with a penalty N/p if we don't reach state x^* in at most $T - t$ steps.

We now notice that if $(x, x') \in E$, then the number of times we have to wait until being able to move from x to x' is a geometric random variable, and so the expected value is at most $1/p$. By extension, the expected time to get to x^* from any state x is at most N/p. Therefore, we observe that $0 \leq v_t(x, y) \leq v_{t+1}(x, y')$ for any y, y'. Consequently, we can take the limit $t \to -\infty$ (or $T \to \infty$), and v will converge to the expected number of steps from x to x^*.

The following example is a discrete-time version of a classical control problem in algorithmic trading, known as the "market making" problem. Although the formulation here is new (to the best of our knowledge), we are inspired by [3, 25, 13] which builds on much earlier works such as [28].

Example 2.3.2. *A* market maker *(MM) is a trader that provides liquidity to the market of a given asset by displaying prices and volumes at which they are willing to buy or sell the asset.*[5] *Consider the case where a MM offers a quote*

[5] The currency exchange stalls at any airport can be thought as a retail version of the activity of a market maker. For any given currency pair, say GBP/USD, they offer the price at which they are willing to buy GBP in exchange for USD and a price at which they are willing to sell GBP in exchange for USD. These quotes are valid up to a given amount (e.g., 10,000 GBP). For example, one stall could be offering 1.34 USD per GBP when you give them GBP in exchange for USD, and they require 1.36 USD if you want to purchase GBP in exchange for USD. The difference between 1.36 and 1.34 is known as the spread. These quotes adjust throughout time because the consensus about the level of the fundamental exchange rate changes over time as new events occur and as supply and demand for these currencies gets processed around the world.

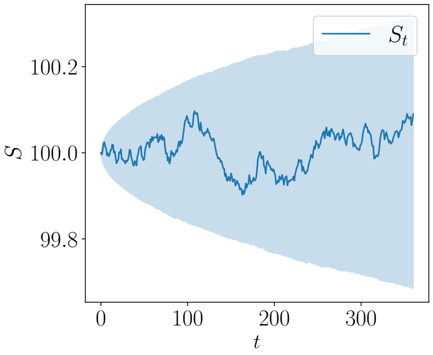


Figure 2.1: Sample path of the fundamental price process with $S_0 = 100$, $\sigma = 0.01$, and $T = 360$.

to buy and a quote to sell a given asset throughout a trading horizon of one hour. Ideally, the quote to buy is at a price that is lower than the price of the quote to sell (this will be the case once we compute the optimal quotes), and this difference is how the business makes money.

In this market, there are other traders, known as liquidity takers *(or LTs) who come and trade with the market maker. Typically, liquidity takers arrive randomly with probabilities that depend on how "generous" the quotes of the market maker are with respect to some "fundamental price" – if the price offered is close to the fundamental price, the MM will expect to trade more quickly.*

In our model, the MM revises their quotes at every timestep. At time t, the new pair of quotes are characterised by their displacements $\delta_t^{\pm}$ from the fundamental price S_t. Formally, we let $t \in \mathbb{T} := \{1, \dots, T\}$ with $T = 360$ (one hour has 360 seconds). The reference probability space $(\Omega, \mathcal{F}, \mathbb{P}^{\text{ref}})$ supports a collection $(\Delta W_t)_{t\in\mathbb{T}}$ of i.i.d. standard normal random variables that we use to model random changes in the fundamental price, and a collection $(n_t^-, n_t^+)_{t\in\mathbb{T}}$ of i.i.d. random vectors uniformly chosen from the set $\{(0,0),\ (1,0),\ (0,1)\}$, which we use to model the arrival of LTs.

We let S_t be the fundamental price, which has dynamics

$$S_{t+1} = S_t + \sigma\, \Delta W_{t+1}\,, \tag{2.1}$$

with $S_0 = 100$ and $\sigma = 0.01$. Figure 2.1 illustrates a sample path of the fundamental price process together with 90% confidence bands across time.

The MM uses controls δ^+ and δ^- to determine the quote to buy (at price $S_t - \delta_t^-$) and the quote to sell (at price $S_t + \delta_t^+$) a total of $\xi = 1$ units of the asset. The larger the values of $\delta^{\pm} = \{\delta^+, \delta^-\}$, the less likely it is LTs to trade at prices $S_t \pm \delta_t^{\pm}$.

Inspired by [3], we model the trades by liquidity takers with Bernoulli random variables $n_t^{\pm}$. Here $(n_t^+, n_t^-) = (1, 0)$ if an LT arrives to buy the asset at time t (so the MM makes a sale), while $(n_t^+, n_t^-) = (0, 1)$ if an LT arrives to sell the *asset. If $(n_t^+, n_t^-) = (0,0)$, then no trade occurs. The filtration is given by*

$$\mathcal{F}_t = \sigma\Big((\Delta W_s)_{s \le t}, (n_s^+)_{s \le t}, (n_s^-)_{s \le t} \Big) \,,$$

In order to use a reference probability approach, we say that the controlled probabilities are given by $\mathbb{P}^{\delta^\pm}$, under which $n_t^\pm$ have joint probabilities given by

$$\mathbb{P}^{\delta^\pm}\Big((n_{t+1}^-, n_{t+1}^+) = (x,y) \Big| \mathcal{F}_t \Big) = \begin{cases} p_t^- & \text{if } (x,y) = (1,0)\,, \\ p_t^+ & \text{if } (x,y) = (0,1)\,, \\ 1 - p_t^- - p_t^+ & \text{if } (x,y) = (0,0)\,, \end{cases} \tag{2.2}$$

with

$$p_t^\pm = \min\big\{ \lambda \, \exp(-\kappa\, \delta_t^\pm), 1/2 \big\}.$$

The parameters $\lambda, \kappa > 0$ model the underlying probability an LT trades (to buy or sell) and the sensitivity of LTs to displacements $\delta^\pm$ respectively. The control does not alter the fundamental price, so the distribution of ΔW_t remains unchanged. In other words, under $\mathbb{P}^{\delta^\pm}$, conditional on $\mathcal{F}_t$, we have independent outcomes $\Delta W_{t+1} \sim N(0,1)$ and (n_{t+1}^+, n_{t+1}^-) as in (2.2).

For $t \in \mathbb{T}$ the inventory of the market maker follows

$$Q_{t+1} = Q_t + n_{t+1}^- - n_{t+1}^+ \,, \tag{2.3}$$

with $Q_0 = 0$. Finally, the cash of the MM follows the dynamics

$$X_{t+1} = X_t - n_{t+1}^- \, (S_t - \delta_t^-) + n_{t+1}^+ \, (S_t + \delta_t^+) \,, \tag{2.4}$$

with $X_0 = 0$.

The objective of the market maker is to maximise profits, while minimizing a terminal inventory penalty; this penalty captures the eagerness of the market maker to not hold any assets at the end of the trading day. We write

$$J(t, \delta^\pm) = \mathbb{E}^{\delta^\pm}\Bigg[-\underbrace{(X_T + Q_T\, S_T - X_t - Q_t\, S_t)}_{\textit{change in inventory value}} + \underbrace{\gamma\, Q_T^2}_{\textit{inventory penalty}} \Bigg]. \tag{2.5}$$

We set

$$\begin{aligned} g(\omega, s, \delta_s^\pm) &= -n_{s+1}^-(\omega)\, \delta_s^- - n_{s+1}^+(\omega)\, \delta_s^+ \,, \\ \Phi(\omega) &= \gamma \Bigg(\underbrace{\sum_{s=1}^{T} n_s^-(\omega) - n_s^+(\omega)}_{Q_T(\omega)} \Bigg)^2 . \end{aligned}$$

and write $X_T + Q_T\, S_T$ as a telescoping sum

$$\sum_{t=0}^{T-1} \Big(X_{t+1} + Q_{t+1}\, S_{t+1} - X_t - Q_t\, S_t \Big) .$$

Using equations (2.3) *and* (2.4)*, we obtain*

$$\begin{aligned}&\sum_{t=0}^{T-1}\Big(-n_{t+1}^-(S_t-\delta_t^-)+n_{t+1}^+(S_t+\delta_t^+)\\&\qquad\qquad+(Q_t+n_{t+1}^- - n_{t+1}^+)(S_t+\sigma\,\Delta W_{t+1})-Q_t\,S_t\Big)\\&=\sum_{t=0}^{T-1}\Big(n_{t+1}^-\,\delta_t^- + n_{t+1}^+\,\delta_t^+ + (Q_t+n_{t+1}^- - n_{t+1}^+)\sigma\,\Delta W_{t+1}\Big)\,.\end{aligned}$$

From here, we recall that, under $\mathbb{P}^{\delta^\pm}$*,* ΔW_{t+1} *and* $n_{t+1}^\pm$ *are* $\mathcal{F}_t$*-conditionally independent and* $\mathbb{E}^{\delta^\pm}[\Delta W_{t+1}|\mathcal{F}_t]=0$*. Hence*

$$\mathbb{E}^{\delta^\pm}[(Q_t+n_{t+1}^- - n_{t+1}^+)\sigma\,\Delta W_{t+1}]=0\,.$$

Therefore, we have an alternative representation of our costs:

$$J(t,\delta^\pm)=\mathbb{E}^{\delta^\pm}\left[\left(\sum_{s=t}^{T-1}g(\omega,s,\delta^\pm)\right)+\Phi(\omega)\right]. \tag{2.6}$$

From Theorem 2.1.16, it follows that the value process

$$V_t=\operatorname*{ess\,inf}_{\delta^\pm}J(t,\delta^\pm)\,,$$

satisfies Bellman's equation

$$V_t=\operatorname*{ess\,inf}_{\delta^\pm}\mathbb{E}^{\delta^\pm}\big[g(t,\delta_t^\pm)+V_{t+1}\big|\mathcal{F}_t\big].$$

Recursively, we have that $V_T=\gamma\,Q_T^2$*, and*

$$\begin{aligned}V_{T-1}&=\operatorname*{ess\,inf}_{\delta^\pm}\mathbb{E}^{\delta^\pm}\big[-n_T^-\,\delta^- - n_T^+\,\delta^+ + \gamma\,Q_T^2\big|\mathcal{F}_{T-1}\big]\\&=\operatorname*{ess\,inf}_{\delta^\pm}\mathbb{E}^{\delta^\pm}\big[-n_T^-\,\delta^- - n_T^+\,\delta^+ + \gamma\,(Q_{T-1}+n_T^- - n_T^+)^2\big|\mathcal{F}_{T-1}\big]\\&=\operatorname*{ess\,inf}_{\delta^\pm}\Big\{-p^-\,\delta^- - p^+\,\delta^+ + \gamma\,Q_{T-1}^2(1-p^- - p^+)\\&\qquad\qquad+\gamma\,(Q_{T-1}+1)^2\,p^- + \gamma\,(Q_{T-1}-1)^2\,p^+\Big\}\,.\end{aligned}$$

Next, we assume that the minimum in $p_t^\pm$ *is attained at* $\lambda\,\exp(-\kappa\,\delta_t^\pm)$*, that is, we assume* $\lambda\,\exp(-\kappa\,\delta_t^\pm)\le 0.5$*. Of course, when deploying our numerical experiments below we would need to double check this condition is satisfied (otherwise the analysis below would be invalid). From here, a first order condition shows that*

$$\delta_{T-1}^{+*}=\frac{1}{\kappa}+\gamma(1-2\,Q_{T-1})\,, \tag{2.7}$$

$$\delta_{T-1}^{-*}=\frac{1}{\kappa}+\gamma(1+2\,Q_{T-1})\,, \tag{2.8}$$

and plugging this back in the expression for V_{T-1} yields

$$V_{T-1} = \gamma\, Q_{T-1}^2 - \frac{\lambda}{\kappa} \exp(-1 - \gamma\, \kappa (1 + 2\, Q_{T-1})) \big(1 + \exp(4\, \gamma\, \kappa\, Q_{T-1})\big)\,. \quad (2.9)$$

From this point is easy to see that if $\gamma = 0$, that is, if the market maker is not averse to holding inventory[6]*, the optimal strategy is $\delta_{T-1}^{\pm *} = 1/\kappa$. This is because V_{T-1} would be constant, and moreover, the recursive calculation would yield the same optimality condition at each step so that $\delta_t^{\pm *} = 1/\kappa$ for $t \in \mathbb{T}$.*

From here, we see that attempting to solve

$$V_{T-2} = \operatorname*{ess\,inf}_{\delta^\pm} \mathbb{E}^{\delta^\pm} \big[- n_{T-1}^- \, \delta^- - n_{T-1}^+ \, \delta^+ + V_{T-1} \big| \mathcal{F}_{T-2} \big]$$

is quite challenging and the recursive steps become increasingly challenging. If however, the market is such that the spreads $\delta^\pm$ remain small, we expect κ to be large and hence $1/\kappa \approx 0$, in this case, the problem simplifies substantially. Alternatively, if the trading horizon is fixed (e.g., one day or one hour), and the number of timesteps at which the market maker makes decisions is large, then we expect $\lambda \approx 0$. More precisely, for a fixed horizon, there is a number of steps T such that λ is as small as we wish. In either of these cases, the problem is simpler to solve because $V_{T-1} \approx \gamma\, Q_{T-1}^2$, and we see that the equation for V_{T-2} becomes tractable.[7] *More precisely,*

$$\begin{aligned} V_{T-2} &\approx \operatorname*{ess\,inf}_{\delta^\pm} \mathbb{E}^{\delta^\pm} \big[- n_{T-1}^- \, \delta^- - n_{T-1}^+ \, \delta^+ + \gamma\, (Q_{T-2} + n_{T-1}^- - n_{T-1}^+)^2 \big| \mathcal{F}_{T-2} \big] \\ &= \operatorname*{ess\,inf}_{\delta^\pm} \Big\{ - p^-\, \delta^- - p^+\, \delta^+ + \gamma\, Q_{T-2}^2 (1 - p^- - p^+) \\ &\qquad\qquad + \gamma\, (Q_{T-2} + 1)^2\, p^- + \gamma\, (Q_{T-2} - 1)^2\, p^+ \Big\}\,, \end{aligned}$$

which yields

$$\delta_{T-2}^{+*} \approx \frac{1}{\kappa} + \gamma (1 - 2\, Q_{T-2}), \quad (2.10)$$

$$\delta_{T-2}^{-*} \approx \frac{1}{\kappa} + \gamma (1 + 2\, Q_{T-2}). \quad (2.11)$$

Repeating these approximations, the strategy of the market maker should always be of the form

$$\delta_t^{+*} \approx \frac{1}{\kappa} + \gamma (1 - 2\, Q_t), \quad (2.12)$$

$$\delta_t^{-*} \approx \frac{1}{\kappa} + \gamma (1 + 2\, Q_t). \quad (2.13)$$

[6] This solution is well known in the continuous-time case; see e.g. [14, Chapter 10].

[7] The argument that $\lambda \to 0$ as the number of steps goes to infinity is at the core of why the continuous-time Avellaneda–Stoikov framework is tractable. Observe that if the λ/κ-term in (2.9) is not zero, the recursive equation for V_{T-2} is far from tractable.

We emphasise that these are approximate solutions because we made the simplifying assumption that λ/κ was small enough so that the corresponding term can be neglected in the recursive equations for V, furthermore, we also disregarded the minimum operator in the probabilities $\mathrm{p}^{\pm}$. In the numerical implementation below, the true values of $\lambda\,\exp(-\kappa\,\delta_t^{\pm})$ are indeed less than or equal to 0.5, so that our simplification of the minimum operator above is valid. We take $\kappa = 100$ so that the baseline spread $1/\kappa$ is 0.01, and the remainder of model parameters are $S_0 = 100$, $\sigma = 0.01$, $T = 360$, $\lambda = 0.1$, and $\gamma \in \{0.0001, 0.001, 0.01\}$.

Figure 2.2 shows how the histogram of terminal inventory Q_T changes as we increase the penalty γ on holding terminal inventory.

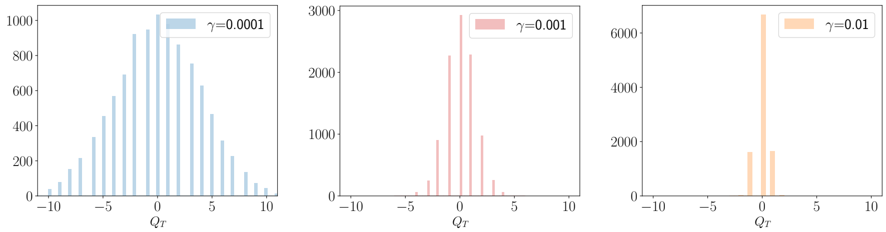


Figure 2.2: Histogram of the terminal inventory Q_T over 10,000 simulations for $\gamma = 0.0001$ (left), $\gamma = 0.001$ (middle), and $\gamma = 0.01$ (right).

Observe that higher aversion to holding inventory translates into more missed opportunities to trade, and a lower value $X_T + Q_T\,S_T$ of the inventory position at the end of the trading day. This is illustrated in Figure 2.3.

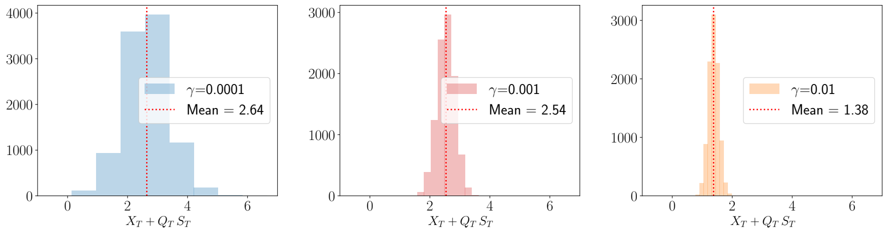


Figure 2.3: Histogram of the terminal wealth $X_T + Q_T\,S_T$ over 10,000 simulations for $\gamma = 0.0001$ (left), $\gamma = 0.001$ (middle), and $\gamma = 0.01$ (right).

Lastly, Figure 2.4 shows that the bound $\mathrm{p}_t^{\pm} \in (0, 1/2)$ for all $t \in \mathbb{T}$ is always satisfied in the simulations we run.

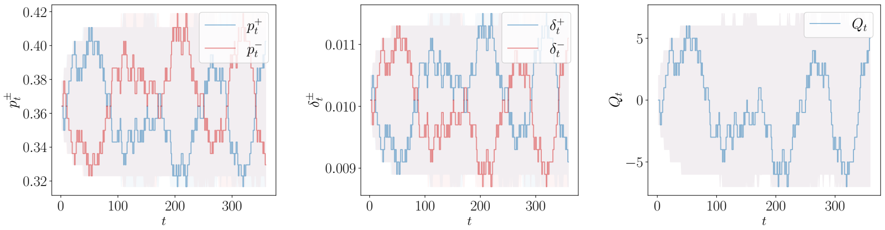


Figure 2.4: Sample path for the inventory Q (left panel), probabilities $p^\pm$ (middle panel), and displacements $\delta^\pm$ (right panel) for the case $\gamma = 0.0001$.

2.4 Exercises

Exercise 2.4.1. *Consider the following weak formulation version of a classical linear-quadratic-Gaussian (LQG) control problem, which adds randomness to the LQ problem we considered earlier.*

Let $T \in \mathbb{N}$ be the finite horizon. Our control[8] takes values in $\mathcal{U} = \mathbb{R}^n$. We introduce a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{T}}, \mathbb{P}^{\text{ref}})$ supporting T independent m-dimensional normal random vectors

$$Y_s \;\sim\; \mathcal{N}\big(0,\ \Sigma_{s+1}\big), \qquad s = 0, 1, \ldots, T-1,$$

with Σ_s a positive definite covariance matrix. Let $X_0 \in \mathbb{R}^m$ be given and define $X_{s+1} = A_s\, X_s + Y_s$. Define the filtration to be that generated by X and assume the matrices $A_s \in \mathbb{R}^{m\times m}$ are deterministic. For a given control U adapted to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{T}}$, we let $\mathbb{P}^U$ be a measure where

$$X_{s+1}\,\big|\,\mathcal{F}_s \;\sim\; \mathcal{N}\big(A_s X_s + B_s U_s,\ \Sigma_{s+1}\big), \qquad s = 0, 1, \ldots, T-1,$$

with given deterministic matrices $B_s \in \mathbb{R}^{m\times n}$. Finally, the performance criterion of the controller is

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

where

$$g(\omega, s, U) = X_s(\omega)^\top Q_s\, X_s(\omega) + U^\top R_s\, U$$

and the terminal cost

$$\Phi_T = X_T(\omega)^\top Q_T\, X_T(\omega),$$

with Q_s, R_s symmetric strictly positive definite matrices for all s. Define the value function of the problem to be

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

[8] Note that, while we do not have the assumed compactness of $\mathcal{U}$ in this problem, the martingale verification result (Theorem 2.2.5) shows that this is not an issue, if we can identify a candidate optimal control.

and using a quadratic ansatz or solving recursively, show that the optimal control is of the form

$$U_t^* = -(B_t^\top K_{t+1} B_t + R_t)^{-1} B_t^\top K_{t+1}^\top A_t X_t \, ,$$

where the symmetric positive semidefinite matrices K_t are recursively defined as $K_T = Q_T$ and $K_t = A_t^\top \big(K_{t+1} - K_{t+1} B_t (B_t^\top K_{t+1} B_t + R_t)^{-1} B_t^\top K_{t+1}\big) A_t + Q_t$.

Exercise 2.4.2. *Consider a stochastic control problem, where we seek to minimize*

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

Show that any optimal control for this problem is also optimal for the problem of minimizing

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

where h is an adapted process. Hence show that for any stochastic control problem, and any choice of adapted stochastic process $\hat{V}$ there exists an equivalent problem (i.e. one where $\hat{J}$ induces the same preference ordering of strategies) such that the value process is given by $\hat{V}$.

Exercise 2.4.3. *Consider a stochastic dynamic control problem where an agent has an exponential utility function, that is, they wish to minimize*

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

where $\gamma > 0$ is a risk aversion parameter. Assuming g is bounded, show that the value of this problem satisfies a dynamic programming equation of the form

$$V_t = \operatorname*{ess\,inf}_{u \in \mathbb{U}} \mathbb{E}^U\Big[\exp\Big(\gamma g(t,U_t) + \log V_{t+1}\Big)\Big| \mathcal{F}_t\Big]$$

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

Exercise 2.4.4. *Consider a stochastic control problem, as described above, and define the total cost process*

$$\hat{\Phi} = \sum_{s=0}^{T-1} g(s,U_s) + \mathbf{1}_{\{t=T\}}\Phi$$

and the accumulated cost $Y_t = \sum_{s=0}^{t-1} g(s,U_s)$.

Show that the control problem without running cost, but with terminal cost $\hat{\Phi}$, is equivalent to the original problem (in the sense of having the same preference ordering of controls), and that there is an adapted process $\hat{V}$ such that $\hat{V}_t + Y_t$ is the value process of the original control problem. Apply the martingale principle to describe the dynamics of $\hat{V}$.

Remark 2.4.5. The previous exercise highlights that there are a variety of different representations for the same control problem. We have generally written our analysis in *Bolza form* (with both a running cost and terminal cost), but this question shows that there is an equivalen *Mayer form* (with only a terminal cost).

If we think of the accumulated cost as part of our state process (even though we have not focussed on models with state variables in this chapter), then it becomes clear that we can always replace a problem in Bolza form with its equivalent Mayer form. However, this typically requires us to use a strong formulation of the control problem.

Exercise 2.4.6. *Suppose $\Phi : \Omega \times \mathcal{X} \to D \subset \mathbb{R}$ is a random function, and $A : D \to \mathbb{R}$ is a monotone strictly increasing function (examples of these are utility functions, which are often used to represent risk preferences). Consider a control problem in strong formulation and of Mayer form (that is, without a running cost) and with terminal cost $\Phi(X^{\mathbf{u}}_T)$, and consider an agent who seeks to minimize*

$$\mathbb{E}\Big[A(\Phi(X^{\mathbf{u}}_T))\Big].$$

Show that there is a value function $v : \Omega \times \mathcal{X} \to \mathbb{R}$ (called the 'certainty equivalent' value process) satisfying the dynamic programming equation

$$v(\omega, t, X^{\mathbf{u}}_t) = A^{-1}\Big(\operatorname*{ess\,inf}_{\mathbf{u}\in\mathbb{U}} \mathbb{E}\Big[A(v(t+1, X^{\mathbf{u}}_{t+1}))\Big|\mathcal{F}_t\Big]\Big),$$

and that a control $\mathbf{u}^*$ is optimal if and only if

$$v(\omega, t, X^{\mathbf{u}}_t) = A^{-1}\Big(\mathbb{E}\Big[A(v(t+1, X^{\mathbf{u}}_{t+1}))\Big|\mathcal{F}_t\Big]\Big).$$

Exercise 2.4.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space supporting an unknown real parameter Θ and a sequence of observations $(Y_t)_{t=1}^T$.*

Assume that

(i) $\Theta \sim \mathcal{N}(m_0, \sigma_0^2)$,

(ii) conditional on Θ, the observations are independent and $Y_t \mid \Theta \sim \mathcal{N}(\Theta, \sigma^2)$.

Let the filtration be defined by $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$, with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. At each time $t = 0, \dots, T-1$, after observing $\mathcal{F}_t$, the agent chooses an $\mathcal{F}_t$-measurable control U_t, which is interpreted as an estimate of Θ at time t.

We first consider the sequential quadratic-loss criterion

$$J(U) = \mathbb{E}\Big[\sum_{t=0}^{T-1}(U_t - \Theta)^2\Big].$$

(i) Show that this is a stochastic control problem in the weak formulation of this chapter, and that admissible controls are precisely the adapted processes $(U_t)_{t=0}^{T-1}$.

(ii) *For fixed* t*, with* $m_t := \mathbb{E}[\Theta \mid \mathcal{F}_t]$*, show that*

$$\mathbb{E}\big[(U_t - \Theta)^2 \mid \mathcal{F}_t\big] = (U_t - m_t)^2 + \mathbb{V}(\Theta \mid \mathcal{F}_t).$$

(iii) *Using the dynamic programming principle, show that the value process satisfies*

$$V_t = \operatorname*{ess\,inf}_{U_t \in L^2(\mathcal{F}_t)} \mathbb{E}\big[(U_t - \Theta)^2 + V_{t+1}\big|\mathcal{F}_t\big].$$

Deduce that the optimal control at time t *is* $U_t^* = \mathbb{E}[\Theta|\mathcal{F}_t]$*, that is, the conditional mean given our observations.*

(iv) *Show that the value process is*

$$V_t = \sum_{s=t}^{T-1} \mathbb{V}(\Theta \mid \mathcal{F}_s),$$

and conclude that Bayesian mean estimation arises as the solution of a sequential quadratic-loss minimisation problem.

Now consider the sequential quadratic-variation-loss criterion

$$J(U) = \mathbb{E}\Big[\Big(\sum_{t=0}^{T-1} \alpha(U_t - X_t)^2\Big) + (\Theta - X_T)^2\Big].$$

with $\alpha > 0$ *a constant and state dynamics*

$$X_{t+1}^U = U_t, \qquad X_0 = m_0$$

(vi) *Find the optimal control for this problem, and outline how it differs from the earlier quadratic-loss criterion.*

(vii) *Explain what happens as* $\alpha \to 0$*, and how this differs from the case* $\alpha = 0$*.*

Exercise 2.4.8 (Mean–variance control)**.** *Let* $\mathbb{T} = \{0, 1, \dots, T\}$*. We consider a strong formulation stochastic control problem, where the state* $X_t^U \in \mathcal{X} \subseteq \mathbb{R}$ *evolves under feedback controls* $U_t = \mathbf{u}(t, X_t)$ *with transition kernel*

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

Fix scalars $\beta, \gamma > 0$ *and consider the controller who wishes (at time* $= 0$*) to minimize the mean–variance objective*

$$J(U) \;=\; \mathbb{V}(X_T^U) \;-\; \gamma\,\mathbb{E}[X_T^U] + \beta\,\mathbb{E}[\sum_{t=0}^{T-1} U_t^2],$$

where $\mathbb{V}(X_T^U) = \mathbb{E}^U[X_T^2] - (\mathbb{E}[X_T^U])^2$*.*

(i) Show that

$$\inf_U J(U) = \inf_{m\in\mathbb{R}} \Big\{ \inf_{U:\,\mathbb{E}[X_T^U]=m} \mathbb{E}\big[(X_T^U - m)^2\big] - \gamma m + \beta\,\mathbb{E}\Big[\sum_{t=0}^{T-1} U_t^2\Big]\Big\}.$$

(ii) Introduce a Lagrange multiplier $\alpha \in \mathbb{R}$ and show (formally) that solving the inner constrained problem is equivalent to solving the unconstrained quadratic control problem of minimizing

$$J^\alpha(U) := \mathbb{E}\Big[(X_T^U)^2 - \alpha X_T^U + \beta \sum_{t=0}^{T-1} U_t^2\Big].$$

Explain briefly how minimising over α (or over m) recovers the original mean–variance cost.

(iii) Consider the linear model

$$X_{t+1} = a_t X_t + b_t U_t + \sigma_t \varepsilon_{t+1}, \qquad \varepsilon_{t+1} \text{ i.i.d. mean } 0,\ \mathrm{Var} = 1,$$

with given deterministic a_t, b_t, σ_t, $X_0 = x_0$, where U is also assumed to be a real scalar. Give an explicit solution for the problem in part (ii) in this context. (Hint: The approach of Exercise 2.4.1 may be helpful.)

(iv) Using the solution you obtained in part (iii), describe a control which provides the optimal mean-variance cost at the initial time.

(v) Show, using your explicit solution, that the mean-variance control problem does not *satisfy the dynamic programming principle, when considering the cost-to-go obtained by replacing the mean and variance with the conditional mean and variance.*

Chapter 3

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 obtain nicer numerical approaches. In this setting, it is natural that our controls will affect the transition probabilities, so our general results on the weak formulation of stochastic control are applicable.

3.1 Controlled Markov chains

Formally, we describe our problem through the use of the transition law (also known as kernel, density, generator) $p : \mathcal{X} \times \mathbb{T} \times \mathcal{X} \times \mathcal{U} \to [0, 1]$ defined as

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

where $\mathcal{X}$ is the (discrete) state space in which X takes values, $x, x' \in \mathcal{X}$, 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; in addition, this assumption is enough that we will later be able to include randomization of strategies (so we will have access to an independent random variable at each time) and of costs.

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

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

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.

Once we have this setup, we see that it's easy to compute various quantities, for example, as X is Markov, for any function $g : \mathcal{X} \to \mathbb{R}$, we have the expectation, for $s < t$, by the law of total probability

$$\mathbb{E}^U\big[g(X_t)\big|\mathcal{F}_s\big] = \sum_{x\in\mathcal{X}} p(x; s, X_s, U)\mathbb{E}^U\big[g(X_t)\big|X_{s+1} = x, \mathcal{F}_s\big].$$

If $t = s + 1$, then the right hand expectation can be dropped, giving

$$\mathbb{E}^U\big[g(X_{s+1})\big|\mathcal{F}_s\big] = \sum_{x\in\mathcal{X}} p(x; s, X_s, U)g(x).$$

If we restrict U to being of *feedback form*, that is $U_s = \mathbf{u}(s, X_s)$ (for some function $\mathbf{u}$, which does not depend on ω), then the left hand side only depends on the current value of X_s, so

$$\mathbb{E}^U\big[g(X_{s+1})\big|X_s\big] = \sum_{x\in\mathcal{X}} p(x; s, X_s, \mathbf{u}(s, X_s))g(x).$$

This is interesting, as we will see that we can usually find optimal controls of this type. Of course, if X takes infinitely many values, the right hand side changes from sums to integrals with respect to the conditional measure $p(\mathrm{d}x; s, X_s, U)$.

If X takes only $n < \infty$ values, it's often useful to assume that these values are simply the basis vectors of $\mathbb{R}^n$. In that case, we see that for any function g, we have

$$g(e_i) = \langle e_i, \mathbf{g}\rangle = \mathbf{g}_i,$$

where $\mathbf{g}$ is the vector with entries $\mathbf{g}_i$. We can then write a matrix $[P(s, U)]_{ij} = p(e_j; s, e_i, U)$, so that (for feedback controls) we have

$$\mathbb{E}^U\big[g(X_{s+1})\big|X_s = e_i\big] = e_i^\top P(s, U)\mathbf{g}$$

which allows us to use the tools of linear algebra to manipulate expectations. In particular, note that the conditional distribution of X_{t+1} is given by $X_t^\top P(t + 1, U)$. Essentially, $P(t, U)$ is the transition matrix for the time-inhomogeneous Markov chain X, when we use the control U.

3.1A 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

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

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 [45] for a classic selection. Two examples cited there are:

Example 3.1.1. *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 refills randomly each day, up to a maximum level. By discretizing the level in the dam, which we write X, we obtain an MDP.*

Example 3.1.2. *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 the following.

Example 3.1.3. *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 (which is controlled), 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 below.

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 3.1.4. *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 3.1.5. *Under Assumption 3.1.4,*

(i) 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 $V_t = v(t, X_t)$;

(ii) 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 $v(T,x) = \Phi(x)$;

(iii) there exists at least one optimal control which is of feedback *type, that is, $U_t^* = \mathbf{u}^*(t, X_t)$ for some function $\mathbf{u}^* : \mathbb{T} \times \mathcal{X} \to \mathcal{U}$, 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 $V_{t+1} = v(t+1, X_{t+1})$, where $v(t, \cdot)$ is the function defined by the Bellman equation. As the terms inside the minimum are all continuous with respect to u, and u takes values in the compact set $\mathcal{U}$, we know there exists[1] a minimizer

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

for each value of (t, x).

It remains to show show how v relates to the value process. We know from Theorem 2.1.16 that

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

As $\mathbf{u}^*$ was defined as the pointwise minimizer (which is always smaller than the essential infimum), it is clear that

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

On the other hand, it is also the case that $U_t = \mathbf{u}^*(t, X_t)$ is an admissible strategy, and hence we have the almost sure inequality

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

This completes the inductive step, and hence the proof. □

The following proposition deals with the cases when p and g have particular structures.

Proposition 3.1.6. *Consider a Markov decision problem with costs $g(t, x, u)$ and transition probabilities $p(x'; t, x, u)$. Under Assumption 3.1.4, we have that*

(i) If p is affine with respect to u, and g is strictly convex with respect to u, there is a unique optimal strategy.

(ii) If p and g are both affine with respect to u, and u takes values in an interval $[u_{\min}, u_{\max}]$, there is an optimal strategy which only ever takes the values $u_{\min}$ and $u_{\max}$ (so-called Bang-Bang solutions).

[1] As $\mathcal{X}$ is discrete, we don't need to worry about measurability in x here, but otherwise could appeal to a measurable selection result such as Filippov's implicit function theorem, (Theorem 2.2.7).

Proof. For (i), from the Bellman equation we know that a control U^* is optimal if and only if

$$U_t^* \in \arg\min_{u \in \mathcal{U}} \Big\{ g(t,x,u) + \sum_{x'} p(x';t,x,u)v(t+1,x') \Big\}.$$

We observe that the term inside the brackets is the sum of a strictly convex function and a collection of affine functions (with respect to u), and hence is strictly convex. However, strictly convex functions have unique minimizers, so there is a unique choice of u which satisfies this property. Therefore, the optimal control is unique.

For (ii), as above, we now see that U^* is optimal if and only if it is the minimizer of an affine function on an interval. Affine functions achieve their minima at the boundaries of their domains, so U^* will only take values on the boundary of $[u_{\min}, u_{\max}]$. □

Remark 3.1.7. 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* (in our general filtration). That is, when solving an MDP, there's no possible advantage to remembering past values of the state, or other sources of randomness, when determining your control.

Remark 3.1.8. 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 $\rho > 0$,

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

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

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

(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.

3.2 Infinite-horizon Discounted MDPs

A common extension 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 for MDPs, based on our finite-time results.

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 3.1.4, we now notice that $\max_{x\in\mathcal{X}, u\in\mathcal{U}} |g(x, u)| \leq \bar{g}$ 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 fairly easy to check that the dynamic programming principle still holds for this problem.[2]

Theorem 3.2.1. *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 $U_t = \mathbf{u}^*(X_t)$, where $\mathbf{u}^*(x)$ 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 $T \in \mathbb{T}$,

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

Therefore, for any $T > t$,

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

[2] A technicality that arises here is that the law of large numbers implies we can only usually assume that there is a measure $\mathbb{P}^{\text{ref}}$ such that different choices of U induce laws $\mathbb{P}^U$ which are absolutely continuous when restricted to $\mathcal{F}_t$ for all $t < \infty$, rather than on the full space. This does not cause significant difficulties when proving the dynamic programming principle here.

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

$$\begin{aligned}0 \le v^+_T(t,X_t) - v^-_T(t,X_t) &\le J^+_T(t,U^-) - v^-_T(t,X_t)\\ &\le 2\frac{e^{-\rho(T-t)}}{1-e^{-\rho}}\bar{g} \to 0 \quad \text{as} \quad T\to\infty.\end{aligned}$$

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 $v^\pm_T$ 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 $\tilde{\mathbf{u}}^*(t,x)$ 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 $\mathbf{u}$, by $p_{t,t+1}(x';x,\mathbf{u}) = p(x';x,\mathbf{u}(t,x))$ and

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

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

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

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

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

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

Remark 3.2.2. 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 $T \to \infty$.

3.2A Ergodic control

Consider an agent who faces a discrete time MDP over an infinite horizon, with a time-homogenous cost $g(x, u)$ and transition probabilities $p(x'; x, u)$. As usual, our actions take values in a compact set $\mathcal{U}$, and g and p are both continuous with respect to u.

However, this agent wants to minimize the long-run average (or Cesàro sum) cost

$$\bar{J}(X_0, U) := \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^U \Big[\sum_{t=0}^{T} g(X_t, U_t) \Big]. \tag{3.1}$$

We suppose the following geometric ergodicity property holds:

Assumption 3.2.3. *There exist constants $R, \gamma > 0$ such that*

$$\max_{U \in \mathbb{U}} \max_{x, \tilde{x}, x' \in \mathcal{X}} \Big| \mathbb{P}^U[X_t = x' | X_0 = x] - \mathbb{P}^U[X_t = x' | X_0 = \tilde{x}] \Big| \leq R e^{-\gamma t} \quad \text{for all } t.$$

This assumption guarantees that, for every feedback control $U_t = u(X_t)$, the state X is a Markov chain under $\mathbb{P}^U$ with a unique stationary distribution π^U, and the distribution of X_t converges to this stationary distribution (geometrically quickly, in total variation norm, with a rate which is uniform in the choice of control).

Lemma 3.2.4. *If there is an optimal control which is of feedback form, the long-run average cost criterion is the same as minimizing the expected cost*

$$\sum_x \pi^{\mathbf{u}}(x) g(x, \mathbf{u}(x))$$

where $\pi^{\mathbf{u}}$ is the stationary distribution of X when using the feedback control $U_t = \mathbf{u}(X_t)$.

Proof. Observe that if $\mathbf{u}$ is a feedback control, then

$$\mathbb{E}^{\mathbf{u}}[g(X_t, \mathbf{u}(X_t))] \to \sum_t \pi^{\mathbf{u}}(x) g(x, \mathbf{u}(x))$$

as $t \to \infty$ (because of the convergence of the law of X_t). A standard argument shows that the Cesàro sum of a convergent sequence is equal to its limit, that is,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}^{\mathbf{u}}[g(X_t, \mathbf{u}(X_t))] = \lim_{t\to\infty} \mathbb{E}^{\mathbf{u}}[g(X_t, \mathbf{u}(X_t))]$$

giving the result. □

We study this problem through an approximate infinite-horizon discounted problem, where the agent seeks to minimize

$$J^\rho(X_0, U) = \mathbb{E}^U\Big[\sum_{t=0}^{T} e^{-\rho t} g(X_t, U_t)\Big]. \tag{3.2}$$

We write v^ρ for the value function for the discounted problem, that is

$$v^\rho(X_0) = \min_{U\in\mathbb{U}} J^\rho(X_0, U)\,.$$

We aim to show that taking $\rho \to 0$ gives a problem which converges, in an appropriate sense, to the long-run average cost problem.

Remark 3.2.5. Typically we have that $v^\rho(x)$ diverges as $\rho \to 0$. For example, consider $g(x, u) \equiv 1$ and observe that

$$v^\rho(x) = \sum_{t=0}^{\infty} e^{-\rho t} = \frac{1}{1 - e^{-\rho}}.$$

Taking $\rho \to 0$ we see this diverges. This is the typical behaviour, as the total undiscounted cost becomes infinite when we consider it over an infinite horizon, as we will face the same costs infinitely many times.

Theorem 3.2.6. *The minimal average cost is the unique value $\lambda \in \mathbb{R}$ such that there is a bounded*[3] *solution $\bar{v}$ to the ergodic Bellman equation*

$$\bar{v}(x) = \min_{u \in \mathcal{U}} \Big\{ g(x,u) - \lambda + \sum_{x'} p(x';x,u)\bar{v}(x') \Big\}.$$

Furthermore, we have the expressions

$$\begin{aligned} \lambda &= \lim_{\rho \to 0} \big\{ (1-e^{-\rho})v^\rho(0) \big\} = \lim_{\rho \to 0} \big\{ \rho\, v^\rho(0) \big\}, \\ \bar{v}(x) &= \lim_{\rho \to 0} \big\{ v^\rho(x) - v^\rho(0) \big\}, \end{aligned}$$

where v^ρ is the value function for the discounted problem with discount rate $\rho > 0$.

Proof. For simplicity, we write

$$C = \max_{x,u} |g(x,u)|.$$

Next, with $\mathcal{T}^\rho$ the usual Bellman operator with discount rate ρ, we have

$$\mathcal{T}^\rho(\tilde{v}) = \mathcal{T}^\rho(v^\rho) - e^{-\rho}v^\rho(0) = v^\rho - e^{-\rho}v^\rho(0) = \tilde{v} + (1-e^{-\rho})v^\rho(0).$$

Rearrangement gives that $\tilde{v}^\rho(x) := v^\rho(x) - v^\rho(0)$ satisfies

$$\tilde{v}^\rho(x) = \min_u \Big\{ g(x,u) - (1-e^{-\rho})v^\rho(0) + e^{-\rho}\sum_{x'} p(x';x,u)\tilde{v}(x') \Big\}$$

and taking the arg min in this equation gives the same (optimal) controls as v^ρ. That the optimal control remains unchanged is simply because adding a constant doesn't ever change your optimal controls as we saw in Exercise 2.4.2.

If $\mathbf{u}^\rho$ is an optimal feedback control for the problem with discount rate ρ, we can write our value function as

$$v^\rho(x) = \sum_t \sum_{x'} p_{0,t}(x';x,\mathbf{u}^\rho)e^{-\rho t}g(x',\mathbf{u}^\rho(x')).$$

Subtracting, and using Assumption 3.2.3 we see that

$$\begin{aligned} |v^\rho(x) - v^\rho(0)| &= \Big| \sum_t \sum_{x'} \Big(p_{0,t}(x';x,\mathbf{u}^\rho) - p_{0,t}(x';0,\mathbf{u}^\rho) \Big) e^{-\rho t} g(x',\mathbf{u}^\rho(x')) \Big| \\ &\leq \sum_t \sum_{x'} \Big| p_{0,t}(x';x,\mathbf{u}^\rho) - p_{0,t}(x';0,\mathbf{u}^\rho) \Big| e^{-\rho t} |g(x',\mathbf{u}^\rho(x'))| \\ &\leq \sum_t \sum_{x'} Re^{-\gamma t}e^{-\rho t}C, \end{aligned}$$

[3] It is easy to check that if $\bar{v}$ is a solution to the ergodic Bellman equation, then so is $\bar{v} + \alpha$ for any $\alpha > 0$, so this bounded solution is generally not unique.

thus

$$|\tilde{v}^\rho(x)| = |v^\rho(x) - v^\rho(0)| \leq \frac{R|\mathcal{X}|}{1-e^{-(\gamma+\rho)}}C.$$

Using the same expansion as above, we see

$$|v^\rho(x)| \leq \sum_t \sum_{x'} p_{0,t}(x'; x, \mathbf{u}^\rho) e^{-\rho t} \max_{x,u} |g(x,u)| = \frac{1}{1-e^{-\rho}}C.$$

We notice from the above inequalities that $\tilde{v}(x)$ and $(1-e^{-\rho})v^\rho(0)$ live in compact sets. Therefore, we can take any sequence $\rho \to 0$, and find a subsequence for which these terms all converge. Taking limits in the Bellman equation, with $\lambda = \lim_{\rho\to 0}(1-e^{-\rho})v^\rho(0)$ and $\bar{v}(x) = \lim_{\rho\to 0}\tilde{v}^\rho(x)$, we have

$$\bar{v}(x) = \min_u \Big\{ g(x,u) - \lambda + \sum_{x'} p(x'; x, u)\bar{v}(x') \Big\}.$$

We also note that, as $\tilde{v}^\rho(0)$ is of order $O(\rho)$, we know

$$\lambda = \lim_{\rho\to 0}(1-e^{-\rho})v^\rho(0) = \lim_{\rho\to 0} \rho\, v^\rho(0)$$

From the ergodicity assumption, we know that any time-homogeneous feedback control $\mathbf{u}$ makes X into a Markov chain with stationary distribution $\pi^{\mathbf{u}}$, which satisfies $\pi^{\mathbf{u}}(x') = \sum_x p(x'; x, \mathbf{u}(x))\pi^{\mathbf{u}}(x)$. Therefore, using the fact that $(\lambda, \bar{v})$ satisfy the ergodic Bellman equation,

$$\begin{aligned}
\sum_x \pi^{\mathbf{u}}(x)\bar{v}(x) &= \sum_x \pi^{\mathbf{u}}(x) \min_u \Big\{ g(x,u) - \lambda + \sum_{x'} p(x'; x, u)\bar{v}(x') \Big\} \\
&\leq \sum_x \Big(\pi^{\mathbf{u}}(x) g(x, \mathbf{u}(x)) \Big) - \lambda + \sum_{x,x'} \Big(\pi^{\mathbf{u}}(x) p(x'; x, \mathbf{u}(x)) \bar{v}(x') \Big) \\
&= \sum_x \Big(\pi^{\mathbf{u}}(x) g(x, \mathbf{u}(x)) \Big) - \lambda + \sum_{x'} \Big(\pi^{\mathbf{u}}(x')\bar{v}(x') \Big),
\end{aligned}$$

and hence

$$\lambda \leq \sum_x \Big(\pi^{\mathbf{u}}(x) g(x, \mathbf{u}(x)) \Big).$$

This shows that λ is less than or equal to the long-run average cost under the arbitrary feedback control $\mathbf{u}$. We also see that this is an equality if and only if $\mathbf{u}$ achieves the minimum in the ergodic Bellman equation, that is, $\mathbf{u}$ is an optimal policy, that is,

$$\lambda = \inf_{\mathbf{u}} \sum_x \Big(\pi^{\mathbf{u}}(x) g(x, \mathbf{u}(x)) \Big).$$

As this infimum is unique, we conclude that λ is also uniquely defined by the solution of the ergodic Bellman equation. □

Corollary 3.2.7. *With $\bar{v}$ and λ as above, a time-homogenous feedback control u is optimal if and only if*

$$\mathbf{u}(x) \in \arg\min_{u\in\mathcal{U}} \Big\{ g(x,u) - \lambda + \sum_{x'} p(x'; x, u)\bar{v}(x') \Big\}.$$

Finally, observe that one can change the control at any finite number of times (arbitrarily) without changing the long-run average cost. Therefore, modifying an optimal control in this way will always yield another optimal control, and we conclude that the problem generally admits other (time dependent) optimal controls.

3.3 Aside: Entropy-regularized control

When we seek to approximate MDPs numerically, one challenge is that the Bellman equation gives a representation of the optimal strategy in terms of an arg min. Unfortunately, the arg min is generally discontinuous, which means that a small variation in the value function v can result in a large change in the estimate of the optimal control. It is also not generally clear, if our controls take only discrete values, what it would mean to vary the choice of control continuously.

In order to address these potential difficulties, a common strategy is to regularize the selection of the control. A 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

$$g_0(x, \pi) := \sum_{i=1}^{m} \pi_i g(x, u_i)$$

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 initial problem (as any control can be represented as a trivial distribution). However, this leads to a problem where our control is not continuous with respect to the value function – a small change in value can change what control is optimal, leading to instability.

There are a couple of ways to make the probabilities nontrivial. One simple way is to use an ϵ-greedy method, which for a given policy π, defines π^ϵ to be

$$\pi^\epsilon(u|x) = \frac{\epsilon}{|\mathcal{U}|} + (1 - \epsilon)\, \pi(u|x)\,.$$

Clearly, $\sum_u \pi^\epsilon(u|x) = 1$ for $x \in \mathcal{X}$, and each state has at least $\epsilon/|\mathcal{U}| > 0$ probability of being chosen.

A 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 3.3.1. *Take any $c \in \mathbb{R}^m$ and consider minimizing $\big(\langle c,\pi\rangle - \lambda H(\pi)\big)$ 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*

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

Proof. As log has infinite slope near zero, and is concave, 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 ones), 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

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

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. □

Proposition 3.3.2. *Let v be the value function of the unregularized problem (with cost g_0), and v^λ the value of the regularized problem (with cost g_λ). 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

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

where

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

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 3.3.1 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 3.3.3. We will see the Q function, as described above, appears in many approaches to reinforcement learning. At this point the key observation is that π^λ varies continuously with Q^λ, and hence continously with v^λ.

3.4 Exercises

Exercise 3.4.1. *Consider an infinite horizon discounted MDP, with finitely many actions and states. Show that if there are $|\mathcal{U}| = m$ actions, and $|\mathcal{X}| = n$ states, then the value function can be obtained by solving m^n systems of n linear equations and taking the component-wise minimum of these solutions.*

Exercise 3.4.2. *Write down a Markov decision problem with two states, for which there exists an optimal non-Markovian solution.*

Exercise 3.4.3. *Consider a sequence $\lambda_n \to 0$, and the corresponding entropy-regularized relaxed control problems. Show that the values of these problems v^{λ_n} converge to the value of the unregularized problem, and describe the possible limiting behaviour of the corresponding randomized controls.*

Exercise 3.4.4. *Consider the Markov Decision problem where an agent wishes to minimize*

$$J(U) = \mathbb{E}^U\Big[\sum_{t=0}^{\tau} g(X_t, U_t)\Big],$$

where τ is a geometric random variable independent of the control and the state process (and the other terms are as we usually define them). Show that this is equivalent to an infinite-horizon discounted control problem.

Exercise 3.4.5. (This problem is based on ideas due to Harrington [27].)

Consider a model of a firm considering whether to comply with regulations, for example, deciding whether to emit pollution into a river system. At each time, the firm may be inspected, and will be penalized if they are found to be polluting. The inspections are random, but the regulator does not treat all firms the same – they split firms into two classes $\mathcal{P}$ (polluters), and $\mathcal{E}$ (environmental), and can choose the probabilities of inspection, and the penalties for pollution, separately for each set. The firms know what group they are in, and seek to optimize their expected discounted reward.

Suppose that a firm in class $i \in \{\mathcal{E}, \mathcal{P}\}$ is inspected with probability ϕ_i. If found to be violating it will be placed into $\mathcal{P}$ for the next period. If found not to be violating it will be placed into $\mathcal{E}$ for the next period. If a firm is inspected and is found to be violating, they will have to pay a fine F_i (depending on group). If a firm chooses not to violate, they pay a fixed cost c (in both groups). If a firm is not inspected, it remains in its current group.

A firm facing this system must choose, at each time, the probability $u \in [0, 1]$ that it violates (it can randomize its choice, to avoid detection). Assume it wishes to minimize its cost, without considering other (environmental) impacts.

(i) Write down the dynamic programming equation describing the value function of the firm.

(ii) Show that there is an optimal strategy for the firm which will simply violate, or not, in each state (that is, randomization of the strategy is unnecessary).

(iii) Assuming the firm acts over an infinite horizon, with discount rate $\rho > 0$, and $\phi_{\mathcal{P}}, \phi_{\mathcal{E}} \in (0, 1)$, calculate the cost-to-go function in each state for each unrandomized policy, and hence the value function.

(iv) Show that, if $\phi_{\mathcal{E}} F_{\mathcal{E}} \leq \phi_{\mathcal{P}} F_{\mathcal{P}}$ (that is, past polluters are fined more heavily than past environmental firms), then it is never optimal for the firm to follow the rule 'violate when in $\mathcal{P}$ and comply when in $\mathcal{E}$'.

Hint: for each policy U and state i, write $(1 - e^{-\rho})J(i, U)$ as a weighted average of payoffs, and then compare different policies.

(v) Explain why, if

$$\phi_{\mathcal{E}} F_{\mathcal{E}} < c < \phi_{\mathcal{P}} F_{\mathcal{P}},$$

but $\phi_{\mathcal{P}} \ll \phi_{\mathcal{E}}$, the regulator will only be seen to issue small fines $F_{\mathcal{E}}$, and yet expected-profit-maximizing firms will usually comply with the environmental regulations.

Exercise 3.4.6 (Inverse control in entropy-regularized MDPs)**.** *(This problem is based on ideas in [11].)*

Let $\mathcal{X}$ and $\mathcal{U}$ be finite state and action spaces. Fix $\rho > 0$ and consider the infinite-horizon discounted entropy-regularized Markov decision problem with transition probabilities $p(x'; x, u)$ and stage cost $g : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$.

For $\lambda > 0$, define the entropy-regularized cost

$$g_\lambda(x, \pi) = \sum_{u \in \mathcal{U}} \pi(u|x)\, g(x, u) - \lambda H(\pi(\cdot|x)).$$

where $H(\pi) = -\sum_{u\in\mathcal{U}} \pi(u) \log \pi(u)$ is the Shannon entropy. Let v^λ denote the corresponding value function. The entropy-regularized Bellman equation is

$$v^\lambda(x) = -\lambda \log \sum_{u\in\mathcal{U}} \exp\Big(\frac{-Q^\lambda(x,u)}{\lambda}\Big),$$

where

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

The optimal policy is given by the softmin rule

$$\pi^\lambda(u|x) = \frac{\exp\big(-Q^\lambda(x,u)/\lambda\big)}{\sum_{w\in\mathcal{U}} \exp\big(-Q^\lambda(x,w)/\lambda\big)}.$$

Suppose an observer knows $p(x'; x, u)$, ρ and λ, and observes an agent behaving optimally, so that the policy $\pi^\lambda(u|x)$ is known for all (x, u). However, the running cost g is unknown.

(i) Show that for each $x \in \mathcal{X}$ and $u \in \mathcal{U}$ there exists a function $C^\lambda : \mathcal{X} \to \mathbb{R}$ such that

$$Q^\lambda(x, u) = -\lambda \log \pi^\lambda(u|x) + C^\lambda(x).$$

Identify $C^\lambda(x)$ in terms of $v^\lambda(x)$.

(ii) Using the definition of Q^λ, show that

$$g(x, u) = -\lambda \log \pi^\lambda(u|x) + C^\lambda(x) - e^{-\rho} \sum_{x'} p(x'; x, u)\, v^\lambda(x').$$

Explain why this yields a linear system in the unknown running costs g and the unknown state offsets C^λ (or equivalently v^λ).

(iii) Show that g is not uniquely determined from a single observed policy. In particular, prove that if $h : \mathcal{X} \to \mathbb{R}$ is any function, then

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

induces the same entropy-regularized optimal policy as g.

(iv) *Now suppose that the observer sees optimal behaviour for the same stage cost g under two different discount parameters $\rho_1 \neq \rho_2$ (with the same λ), yielding policies π^{λ,ρ_1} and π^{λ,ρ_2}.*

Write down the two corresponding systems from part (ii), and show that by combining them one can eliminate the unknown state offsets and obtain a linear system for g alone (under an appropriate algebraic).

Exercise 3.4.7 (Time-periodic infinite-horizon control)**.** *Let $\mathcal{X}$ and $\mathcal{U}$ be finite state and action spaces. Fix a period $K \in \mathbb{N}$ and consider an infinite-horizon discounted control problem with discount rate $\rho > 0$.*

The transition probabilities and stage costs are periodic in time *with period K, that is,*

$$p_t(x'; x, u) = p_{t+K}(x'; x, u), \qquad g_t(x, u) = g_{t+K}(x, u)$$

for all $t \geq 0$, $x, x' \in \mathcal{X}$ and $u \in \mathcal{U}$.

For an admissible control $U = (U_t)_{t\geq 0}$, the performance criterion is

$$J(x_0, U) = \mathbb{E}^U\left[\sum_{t=0}^{\infty} e^{-\rho t} g_t(X_t, U_t) \;\middle|\; X_0 = x_0\right].$$

(i) *Define the* phase process $(\theta_t)_{t\geq 0}$ *by*

$$\theta_t := t \bmod K, \qquad \theta_t \in \{0, 1, \ldots, K-1\}.$$

Show that the pair (X_t, θ_t) is a time-homogeneous Markov process under any feedback control.

(ii) *Define an augmented state space $\widetilde{\mathcal{X}} = \mathcal{X} \times \{0, 1, \ldots, K-1\}$. Write down the transition kernel for the augmented process: $\widetilde{p}\big((x', \theta'); (x, \theta), u\big)$.*

(iii) *Show that the periodic control problem is equivalent to a time-homogeneous infinite-horizon discounted MDP on the state space $\widetilde{\mathcal{X}}$, with running cost*

$$\widetilde{g}\big((x, \theta), u\big) = g_\theta(x, u).$$

(iv) *Write down the Bellman equation satisfied by the value function $v(x, \theta)$.*

(v) *Show that there exists an optimal feedback control of the form $U_t = \mathbf{u}(X_t, \theta_t)$, and interpret this as a K*-periodic optimal policy *for the original problem.*

(vi) *Suppose $K = 2$ and interpret θ_t as representing "high season" and "low season". Explain why the optimal control may differ between the two seasons, even when the state X_t is the same.*

Exercise 3.4.8 (Optimal stopping of a Markov chain)**.** *Let $(X_t)_{t\geq 0}$ be a Markov chain on $\mathbb{Z}$ defined by*

$$\mathbb{P}(X_{t+1} = x + 1 \mid X_t = x) = p, \qquad \mathbb{P}(X_{t+1} = x - 1 \mid X_t = x) = 1 - p,$$

where $p \in (0,1)$ is fixed.

At each time t, the controller may either (a) continue (and incur a running cost), or (b) stop permanently. If the controller continues at time t, a running cost $g(X_t) = cX_t$ is incurred, where $c > 0$. If the controller stops at time t, a terminal cost $\Phi(X_t) = K$ is incurred, where $K > 0$ is fixed.

Let τ denote the stopping time chosen by the controller. The objective is to minimize the infinite-horizon discounted cost

$$J(x,\tau) = \mathbb{E}\left[\sum_{t=0}^{\tau-1} e^{-\rho t} cX_t + e^{-\rho\tau} K \;\middle|\; X_0 = x\right],$$

where $\rho > 0$.

(i) Show that, despite the state space being infinite, under any stopping strategy, the expected discounded costs in this problem are bounded for each value of $x \in \mathbb{Z}$.

(ii) Show that this problem can be formulated as a Markov decision problem with control space $\mathcal{U} = \{continue, stop\}$, and that the Bellman equation can be written as

$$v(x) = \min\Big\{K,\ cx + e^{-\rho}\Big(p\, v(x+1) + (1-p)\, v(x-1)\Big)\Big\}.$$

(iii) Explain informally why it is optimal to stop whenever x is sufficiently large.

(iv) Show that any optimal stopping rule is of threshold form: there exists $x^* \in \mathbb{Z}$ such that it is optimal to stop if and only if $x \geq x^*$.

(Hint: First show that the value function is increasing in x.)

(v) Assume $p = \frac{1}{2}$ (symmetric walk). On the continuation region $\{x < x^*\}$, show that v satisfies the linear difference equation

$$v(x) = cx + e^{-\rho}\frac{v(x+1) + v(x-1)}{2}.$$

Solve this equation explicitly in the form

$$v(x) = A + Br^x + \alpha x,$$

for suitable constants A, B, α, r and then determine the constants using the boundary conditions

$$v(x^*) = K, \qquad v(x^*-1) = c(x^*-1) + e^{-\rho}\frac{v(x^*) + v(x^*-2)}{2}.$$

(vi) Explain how the boundary x^* balances the immediate stopping cost K against the expected future running cost.

Exercise 3.4.9. *This exercise explores the* turnpike phenomenon *in discrete time, under a uniform ergodicity assumption. We will use the notation and results of Section 3.2A, and in addition assume:*

- *$\mathcal{X}$ is finite and $\mathcal{U}$ is a compact metric space, with metric* dist*;*
- *the uniform geometric ergodicity assumption (Assumption 3.2.3) holds.*

Let $(\lambda, \bar{v})$ be a solution of the ergodic Bellman equation

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

and denote the corresponding set of minimizing controls by

$$\mathcal{U}^*(x) := \operatorname*{arg\,min}_{u \in \mathcal{U}} \Big\{ g(x,u) - \lambda + \sum_{x'} p(x'; x, u)\bar{v}(x') \Big\}, \qquad x \in \mathcal{X}.$$

For each finite horizon $T \geq 1$ let $v_T(t,x)$ be the finite-horizon value with terminal cost $\Phi : \mathcal{X} \to \mathbb{R}$:

$$v_T(t,x) = \min_U \mathbb{E}^U \Big[\sum_{s=t}^{T-1} g(X_s, U_s) + \Phi(X_T) \;\Big|\; X_t = x \Big], \qquad t \in \{0, \dots, T\}.$$

(a) Let

$$A(x,u) := g(x,u) - \lambda + \sum_{x'} p(x'; x, u)\bar{v}(x') - \bar{v}(x), \qquad x \in \mathcal{X}, \;\; u \in \mathcal{U}.$$

Show that $A(x,u) \geq 0$ and $\mathcal{U}^*(x) = \{u \in \mathcal{U} : \; A(x,u) = 0\}$.

(b) Fix $\varepsilon > 0$. Show there exists $\eta > 0$ such that, for all $x \in \mathcal{X}$, any measurable map $\mathbf{u} : \mathcal{X} \to \mathcal{U}$ satisfying

$$\Big| A(x, \mathbf{u}(x)) \Big| < \eta$$

also satisfies

$$\operatorname{dist}\big(\mathbf{u}(x), \mathcal{U}^*(x)\big) < \varepsilon.$$

(c) Show that, if $\mathbf{u} : \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ then we have the representation

$$\sum_{t=0}^{T-1} \big(g(X_t, \mathbf{u}(t, X_t)) - \lambda \big) = \bar{v}(X_0) - \bar{v}(X_T) + \sum_{t=0}^{T-1} A(X_t, \mathbf{u}(t, X_t)),$$

(d) Using the geometric ergodicity assumption, show that

a. For every fixed $x \in \mathcal{X}$,

$$v_T(0,x) - T\lambda - \bar{v}(x) \text{ is uniformly bounded as } T \to \infty.$$

b. *There exists a constant $c \in \mathbb{R}$ such that*

$$\lim_{T\to\infty} \Big[v_T(0,x) - T\lambda - \bar{v}(x) - c \Big] = 0.$$

Hint: Consider two convergent subsequences (with different choices of x).

(e) Let $\mathbf{u}^T(t,x)$ be an optimal feedback control for the finite horizon problem with horizon T and terminal value Φ.

Prove the following turnpike property: for every $\varepsilon > 0$ and every $\delta \in (0, 1/2)$ there exists T_0 such that for all $T \geq T_0$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{\mathrm{dist}(\mathbf{u}^T(t,X_t),\mathcal{U}^*(X_t))\geq\varepsilon\}} \leq \delta, \quad \mathbb{P}^{\mathbf{u}^T}\text{-almost surely.}$$

That is, for large horizon T, the optimal finite-horizon policy $\mathbf{u}^T$ is within ε of the ergodic optimal set $\mathcal{U}^*(\cdot)$ for a fraction $1-\delta$ of the time indices.

Hint: First give a bound on $\mathbb{E}^{\mathbf{u}^T}\big[\sum_{t=1}^{T} \mathbf{1}_{\{A(X_t,\mathbf{u}^T(t,X_t))\geq\eta\}}\big]$ and then sum over the possible states.

Exercise 3.4.10. *Let $\mathcal{X} = \{1, \ldots, n\}$ and $\mathcal{U} = \{1, \ldots, m\}$ be finite state and action spaces. Consider a discounted MDP with discount rate $\rho > 0$, transition probabilities*

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

running cost $g(x,u)$ and terminal cost $\Phi(x)$. Assume we restrict to feedback randomized controls

$$\pi_t(\cdot|x) \in \mathcal{P}(\mathcal{U}), \qquad U_t \mid X_t = x \sim \pi_t(\cdot|x).$$

Let $\mu_t(x) = \mathbb{P}^U(X_t = x)$ denote the (controlled) law of X_t, and assume $X_0 \sim \mu_0$ is given.

(i) We write $\Delta^{n-1} = \{y \in \mathbb{R}^n : y_i \geq 0, \sum_i y_i = 1\}$ for the probability simplex, and note that Δ^{n-1} is equivalent to $\mathcal{P}(\mathcal{X})$.

a. Show that under a feedback policy π_t, the law μ_t evolves deterministically according to

$$\mu_{t+1}(j) = \sum_{i\in\mathcal{X}} \mu_t(i) \sum_{u\in\mathcal{U}} \pi_t(u|i)\, p(j; i, u), \qquad j \in \mathcal{X}.$$

b. Define the controlled transition matrix

$$P^{\pi_t}_{ij} := \sum_{u} \pi_t(u|i) p(j; i, u).$$

Show that the law dynamics can be written compactly as

$$\mu_{t+1}^\top = \mu_t^\top P^{\pi_t}.$$

c. *Show that the expected discounted cost becomes*

$$J(\mu_0, \pi) = \sum_{t=0}^{T} e^{-\rho t} \Big(\sum_{i,u} \mu_t(i)\, \pi_t(u|i)\, g(i,u) \Big) + e^{-\rho T} \Big(\sum_{i,u} \mu_T(i)\, \Phi(i) \Big).$$

Conclude that the MDP is equivalent to a deterministic *optimal control problem with state* $\mu_t \in \Delta^{n-1} \equiv \mathcal{P}(\mathcal{X})$ *and control* π_t.

(ii) *For the deterministic control problem in part (i), we now seek to apply the discrete-time Pontryagin principle (Section 1.4). Introduce adjoint variables* $q_t \in \mathbb{R}^n$ *and define the Hamiltonian*

$$\mathcal{H}_t(\mu_t, \pi_t, q_{t+1}) = e^{-\rho t} \sum_{i,u} \mu_t(i) \pi_t(u|i) g(i,u) + \big(\mu_t^\top P^{\pi_t}\big) q_{t+1}.$$

a. *Show that the adjoint recursion is*

$$q_t(i) = e^{-\rho t} \sum_{u} \pi_t(u|i) g(i,u) + \sum_{j} P_{ij}^{\pi_t}\, q_{t+1}(j).$$

b. *Show that optimality of* $\pi_t(\cdot|i)$ *requires, for each state* i,

$$\pi_t(\cdot|i) \in \arg\min_{\pi \in \mathcal{P}(\mathcal{U})} \sum_{u} \pi(u) \Big(e^{-\rho t} g(i,u) + \sum_{j} p(j; i,u) q_{t+1}(j) \Big).$$

c. *Let* $v(t,i)$ *denote the usual Bellman value function of the MDP. Show that if we set*

$$q_t(i) = v(t,i),$$

then the adjoint recursion coincides with the Bellman equation.

d. *Conclude that the Pontryagin costate variable* q_t*, for the problem of the optimal control of the law, is exactly the value function of the MDP, for the problem of the control of the process.*

Chapter 4

Classical numerical methods

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 consider in this chapter. This has the advantage of avoiding ti me dependence in our solution, while still displaying a wide range of technical challenges.

4.1 Value iteration

Value iteration is the most fundamental approximation for control problems, and focusses on approximating $v(x)$ directly. In what follows, we work with a state space $\mathcal{X}$ with finitely many elements. As $|\mathcal{X}| < \infty$, we can identify functions $w : \mathcal{X} \to \mathbb{R}$ with vectors in $\mathbb{R}^{|\mathcal{X}|} \equiv \mathbb{R}^{\mathcal{X}}$, with components $w_i := w(x_i)$ where x_i is the i-th element in $\mathcal{X}$. We will use this identification liberally, to make notation simpler. We will also write $g(x, \mathbf{u})$ for $g(x, \mathbf{u}(x))$ and similarly $p(x; x', \mathbf{u})$, provided this does not lead to confusion.

Assumption 4.1.1 (Assumptions for Value iteration)**.** *We know the transition law $p(x'; x, u)$, and costs $g(x, u)$ perfectly. Assumption 3.1.4 holds, in particular, the space of controls $\mathcal{U}$ is compact.*

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

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

Definition 4.1.2. *For a feedback control $\mathbf{u} : \mathcal{X} \to \mathcal{U}$, we define the Bellman valuation operator $\mathcal{T}_{\mathbf{u}}$ by*

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

We also define the Bellman optimality operator $\mathcal{T}$ by

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

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

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

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 4.1.3. *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 4.1.4. 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 4.1.5. *With discount rate $\rho > 0$, the Bellman operators $\mathcal{T}, \mathcal{T}_{\mathbf{u}}$ 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,*

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

[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.

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

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

That is, $\mathbf{u}^{(w')}$ would be optimal if the value function were w'. Therefore, with this $\mathbf{u}^{(w')}$,

$$\begin{aligned}
(\mathcal{T}w)(x) - (\mathcal{T}w')(x) &= \min_{u \in \mathcal{U}} \big\{ \big(\mathcal{T}_{\mathbf{u}} w\big)(x) \big\} - \big(\mathcal{T}_{\mathbf{u}^{(w')}} w'\big)(x) \\
&\leq \big(\mathcal{T}_{\mathbf{u}^{(w')}} w\big)(x) - \big(\mathcal{T}_{\mathbf{u}^{(w')}} w'\big)(x) \\
&\leq \Big\{ g(x, \mathbf{u}^{(w')}) + e^{-\rho} \sum_{x' \in \mathcal{X}} p(x'; x, \mathbf{u}^{(w')}) w(x') \Big\} \\
&\qquad - \Big\{ g(x, \mathbf{u}^{(w')}) + e^{-\rho} \sum_{x' \in \mathcal{X}} p(x'; x, \mathbf{u}^{(w')}) w'(x') \Big\} \\
&= e^{-\rho} \sum_{x' \in \mathcal{X}} p(x'; x, \mathbf{u}^{(w')}) \big(w(x') - w'(x') \big). \qquad (4.1)
\end{aligned}$$

As $p(\cdot; x, \mathbf{u}^{(w')})$ is a probability mass function, and averages never exceed the maximum value,

$$\begin{aligned}
(\mathcal{T}w)(x) - (\mathcal{T}w')(x) &\leq e^{-\rho} \sum_{x' \in \mathcal{X}} p(x'; x, \mathbf{u}^{(w')}) \|w - w'\|_\infty \\
&= e^{-\rho} \|w - w'\|_\infty.
\end{aligned}$$

Interchanging w and w' we see that

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

so $\mathcal{T}$ is an L^∞ contraction, with rate $e^{-\rho} < 1$. The argument for $\mathcal{T}_{\mathbf{u}}$ is analogous (and simpler). The convergence result then 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 for value iteration (as we've shown we have a contraction), it is also interesting to see that the Bellman operator always improves the value, and this will prove important in other algorithms.

Theorem 4.1.6. *The Bellman operators are pointwise monotone, in the sense that if $w \leq w'$, then $\mathcal{T}w \leq \mathcal{T}w'$ (both inequalities being pointwise). In particular, if v_0 is such that $\mathcal{T}v_0 \leq v_0$ (for example, if $v_0(x) = (1 - e^{-\rho})^{-1} \max_{x,u} g(x, u)$), then value iteration decreases monotonically to v.*

Proof. From the previous proof, starting with (4.1) we have

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

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

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

$$\tilde{\mathbf{u}}(x) \in \arg\min_{\mathbf{u}} \big\{(\mathcal{T}_{\mathbf{u}}v')(x)\big\}.$$

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

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

Proof. Write $\tilde{v}(x) = J(x, \tilde{\mathbf{u}})$ for the true value associated with the control $\tilde{\mathbf{u}}$, so we know $\tilde{v} = \mathcal{T}_{\tilde{\mathbf{u}}}\tilde{v}$. As v is optimal, we know

$$v \leq \tilde{v}.$$

Now write $\varepsilon = \|v - v'\|_\infty$. 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 $\mathcal{T}_{\mathbf{u}}$. As $\mathcal{T}$ is monotone, we know

$$v = \mathcal{T}v \geq \mathcal{T}(v' - \varepsilon) = (\mathcal{T}v') - e^{-\rho}\varepsilon.$$

Therefore,

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

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

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

Combining these inequalities, we see

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

and hence

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

The result follows by rearrangement. □

Corollary 4.1.8. *If v_n is the sequence obtained by value iteration*[2]*, and we define*

$$\mathbf{u}_n(x) \in \arg\min_{\mathbf{u}} \big\{(\mathcal{T}_{\mathbf{u}} v_n)(x)\big\},$$

then $J(x, \mathbf{u}_n) \to v(x)$ (with geometric convergence). The sequence of feedback policies $\mathbf{u}_n$ has convergent subsequences (when considered as a sequence in $\mathcal{U}^{\mathcal{X}}$) and for any subsequence of policies which converges, the limit $\mathbf{u}^*$ is an optimal policy.

Proof. From Theorem 4.1.7, we have the geometric convergence

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

As our policies lie in a compact set, there is a subsequence $\{\mathbf{u}_{n_m}\}_{m\in\mathbb{N}}$ of $\{\mathbf{u}_n\}_{n\in\mathbb{N}}$ which is convergent. Writing

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

the dominated convergence theorem (applied to the infinite sum with respect to t) implies that along any convergent subsequence we have

$$v(x) = \lim_{m\to\infty} J(x, \mathbf{u}_{n_m}) = J(x, \lim_{m\to\infty} \mathbf{u}_{n_m}),$$

so $\mathbf{u}^* = \lim_{m\to\infty} \mathbf{u}_{n_m}$ is an optimal policy. □

4.2 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 computing minimum in every step becomes expensive. Policy iteration aims to separate these two operations, by allowing us to compute (accurately) the value using multiple (easy) steps, and only infrequently computing the minimum.

Definition 4.2.1. *Policy iteration consists of two alternating steps.*

(i) Evaluation: where for a given policy $\mathbf{u}_n$ we compute its value function $v_n(\cdot) = J(\cdot, \mathbf{u}_n)$.

(ii) Improvement: where for a given value function v_n, we compute a new policy

$$\mathbf{u}_{n+1}(x) \in \arg\min_{u} \big\{\big(\mathcal{T}_{\mathbf{u}} v_n\big)(x)\big\}.$$

[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.

This process can be initialized by specifying either u_0 or v_0, and proceeding iteratively. For example, starting with control u_0, we have the sequence:

$$\mathbf{u}_0 \xrightarrow[\text{evaluation}]{\text{policy}} v_0 \xrightarrow[\text{improvement}]{\text{policy}} \mathbf{u}_1 \xrightarrow[\text{evaluation}]{\text{policy}} v_1 \xrightarrow[\text{improvement}]{\text{policy}} \mathbf{u}_2 \dots$$

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

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

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

$$\mathbf{v}_n = \mathbf{g}(\mathbf{u}_n) + e^{-\rho} P(\mathbf{u}_n)\mathbf{v}_n = \big(I - e^{-\rho} P(\mathbf{u}_n)\big)^{-1} \mathbf{g}(\mathbf{u}_n). \tag{4.2}$$

Given that 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.

Remark 4.2.2. From the perspective of linear algebra, we see that, in order to prove that the value $\mathbf{v}_n$ is well posed in (4.2), the key is to know that $e^{-\rho} P(\mathbf{u}_n)$ has all eigenvalues with real part strictly below 1. This follows from the Perron–Frobenius theorem as $P(\mathbf{u}_n)$ is a stochastic matrix. This gives an alternative (algebraic) approach, as opposed to the more probabilistic method of long-horizon approximation we considered in the previous chapter.

Lemma 4.2.3 (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 $\mathcal{T} v_n = \mathcal{T}_{\mathbf{u}_{n+1}} v_n$ (improvement) and $v_n = \mathcal{T}_{\mathbf{u}_n} v_n$ (evaluation). It follows that

$$v_n = \mathcal{T}_{\mathbf{u}_n} v_n \geq \mathcal{T} v_n = \mathcal{T}_{\mathbf{u}_{n+1}} v_n.$$

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

$$\mathcal{T}_{\mathbf{u}_{n+1}} v_n \geq \big(\mathcal{T}_{\mathbf{u}_{n+1}}\big)^2 v_n.$$

and by induction,

$$v_n \geq \mathcal{T}_{\mathbf{u}_{n+1}} v_n \geq \cdots \geq \big(\mathcal{T}_{\mathbf{u}_{n+1}}\big)^k v_n$$

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

Theorem 4.2.4. *The value functions constructed via policy iteration converge (geometrically quickly) to the true value function (and hence the policies converge, in the sense that Corollary 4.1.8 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 4.2.5. If our set $\mathcal{U}$ is finite, then the fact that 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 $\mathbf{u}_0$, 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 Ye [47].

Example: A controlled walk on a discrete torus

Example 4.2.6. *Consider the control of a process X which takes values in the discrete set $\mathcal{X} = \{(i,j); i,j \in \{0,1,...,K-1\}\}$. From a given timestep to the next, the state moves in a given cardinal direction of the form $(i, j \pm 1) (\operatorname{mod} K)$ corresponding to east/west or $(i \pm 1, j) (\operatorname{mod} K)$ corresponding to south/north. Figure 4.1 illustrates the cardinal moves from two given starting points when $K = 4$.*

At each time, the controller can choose an action from the set:

- *Do nothing, in which case X will take a step in a randomly chosen cardinal direction (north, south, east, or west), that is, $X = (i,j) \in \mathcal{X}$ moves to one of the four points of the form $(i, j \pm 1) (\operatorname{mod} K)$ or $(i \pm 1, j) (\operatorname{mod} K)$.*
- *Choose one direction, and modify the probability of walking in that direction to be a given constant $\alpha < 1$, while all other directions are chosen with equal probability $(1-\alpha)/3$. Figure 4.2 shows the difference between doing nothing (left panel) against choosing to move north (right panel) with probability $\alpha \in (0,1)$.*

If the controller chooses to intervene, they must pay a cost $c = 1$ for doing so. In addition, at each time in state (i,j), they face a random state-dependent cost taken from a Binomial $(N, (i+1)/(i+j+1))$ distribution. The controller chooses actions $u \in \mathcal{U} = \{\varnothing, \mathrm{N}, \mathrm{S}, \mathrm{E}, \mathrm{W}\}$, where $\varnothing$ denotes "no control".

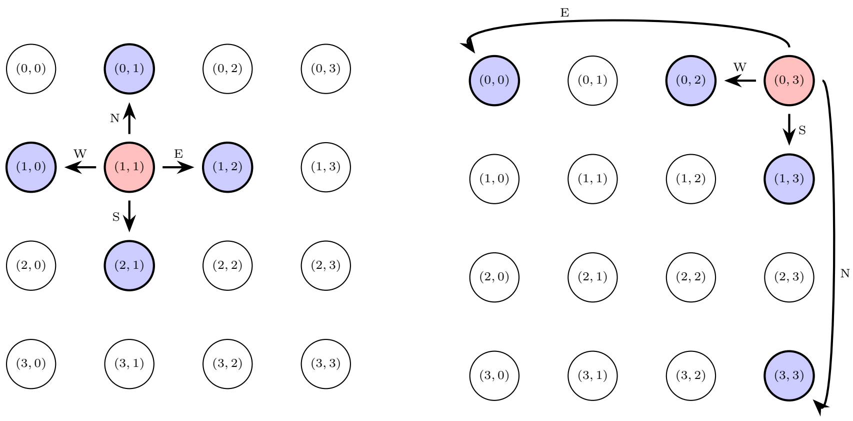


Figure 4.1: Possible cardinal moves with $\mathcal{G} = \{(i, j) : i, j \in \{0, 1, 2, 3\}\}$. Left panel: centre at $(1, 1)$ in red and its reachable neighbours $(0, 1), (1, 0), (1, 2), (2, 1)$ are blue (arrows show direction: N,S,E,W respectively). Right panel: centre at $(0, 3)$ with neighbours shown similar to the left panel.

Let's now write this problem in our familiar notation. For a given state $x = (i, j)$ define the state-dependent Binomial success probability

$$q(x) = \frac{i+1}{i+j+1},$$

and let N denote the Binomial size. The expected immediate cost at state x when action u is chosen is

$$g(x, u) = N\, q(x) + 1_{\{u \neq 0\}}\, c,$$

where c is the intervention cost.

Let $n = |\mathcal{X}| = K^2$. For each action u, let $P(u) \in \mathbb{R}^{n \times n}$ be the transition matrix. Each row is associated with a given state x, giving the $n = K^2$ probabilities of transitions to each alternative state.

For notational brevity, we identify $\{\mathrm{N}, \mathrm{S}, \mathrm{E}, \mathrm{W}\}$ with the vectors describing the corresponding transitions, and so can simply write $x + \mathrm{S} = x + (1, 0)$ (and similarly for other controls), leaving the $(\bmod K)$ implicit. Consequently, we can define the controlled transition probabilities:

$$p(x'; x, u) = \begin{cases} 1/4 & u = \varnothing; x' \in x + \{\mathrm{N}, \mathrm{S}, \mathrm{E}, \mathrm{W}\}, \\ \alpha & u \neq \varnothing; x' = x + u, \\ (1-\alpha)/3 & u \neq \varnothing; x' \in x + \{\mathrm{N}, \mathrm{S}, \mathrm{E}, \mathrm{W}\} \setminus \{u\}, \\ 0 & \textit{otherwise}. \end{cases}$$

Policy iteration seeks a stationary deterministic policy $\mathbf{u} : \mathcal{X} \to \mathcal{U}$ that minimizes expected discounted cost with discount factor $\rho > 0$. Given a policy

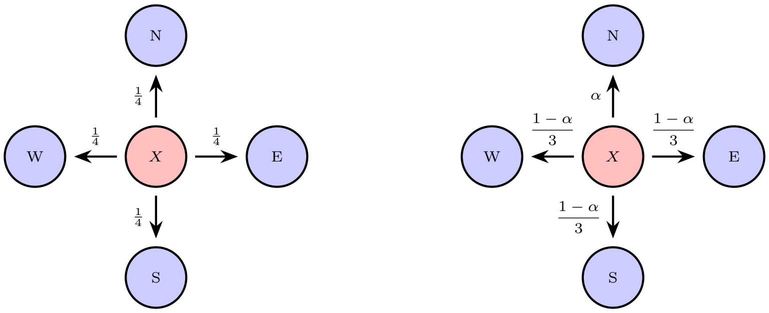


Figure 4.2: Left: when the controller does not intervene, transitions are uniform (1/4 each). Right: with intervention, the chosen direction (north in this example) receives probability α, while each other direction has probability $(1-\alpha)/3$. The current state is red and reachable neighbours are blue.

$\mathbf{u}$*, we can form the policy transition matrix* $P(\mathbf{u}) \in \mathbb{R}^{n\times n}$ *by taking row* x *of* $P(\mathbf{u})$ *to be the vector* $p(\cdot; x, \mathbf{u}(x))$*, and form the policy cost vector* $\mathbf{g}(\mathbf{u}) \in \mathbb{R}^n$ *with entries* $g(x, \mathbf{u}(x))$*.*

Policy iteration then creates a sequence of policies $\mathbf{u}_n$ *and approximate values* $\mathbf{v}_n$*, which approximate the optimal policy-value pair. The policy evaluation step solves the linear system*

$$\big(I - e^{-\rho} P(\mathbf{u}_n)\big)\mathbf{v}_n = \mathbf{g}(\mathbf{u}_n)$$

for the approximate value vector $\mathbf{v}_n \in \mathbb{R}^n$*. Policy improvement then computes, for every state* x*, an improved policy*

$$\mathbf{u}_{n+1}(x) \in \arg\min_{\mathbf{u}} \big\{\big(\mathcal{T}_{\mathbf{u}}\mathbf{v}_n\big)(x)\big\} \;=\; \arg\min_{\mathbf{u}} \big\{g(x, \mathbf{u}(x)) + e^{-\rho} \sum_{x'} P^{\mathbf{u}}(x, x')\, \mathbf{v}_n(x')\big\},$$

If $\mathbf{u}_n = \mathbf{u}_{n+1}$ *the algorithm terminates, as an optimal policy has been found. Otherwise, the sequence repeats, with improvement at every step. Given there are* K^2 *states and* $|\mathcal{U}| = 5$ *policies, we know that the sequence will certainly terminate, in at most* 5^{K^2} *steps. In practice, termination typically occurs far more quickly.*

Figure 4.3 shows the optimal policy when $K = 4$*,* $N = 10$*,* $\alpha = 0.5$*,* $\rho = 0.05$*, and* $c = 1$*.*

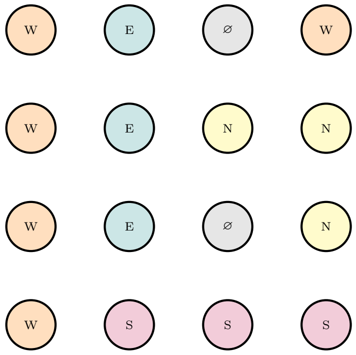


Figure 4.3: Optimal stationary policy for $K = 4, N = 10, \alpha = 0.5, \rho = 0.05$, and $c = 1$. Each circle shows the optimal action in that state (one of $\{\varnothing, N, S, E, W\}$).

We see that the optimal policy shows a preference to have the controlled state at $(0, 2)$ *or* $(2, 2)$*. To understand this, we look at the expected cost of each state, which is* $N\,\frac{i+1}{i+j+1}$, *which can be visualised as the matrix*

$$\begin{pmatrix} 10 & 5 & 3 & 2 \\ 10 & 6 & 5 & 4 \\ 10 & 7 & 6 & 5 \\ 10 & 8 & 6 & 5 \end{pmatrix},$$

The value function computed for each state can be represented by

$$\begin{pmatrix} 130.44 & 125.24 & 121.20 & 121.18 \\ 131.54 & 127.76 & 123.39 & 122.94 \\ 133.32 & 130.55 & 126.15 & 125.47 \\ 133.06 & 130.31 & 125.83 & 125.28 \end{pmatrix}.$$

It's clear from this that the strategy is principally trying to stay away from expensive regions (in particular the most westward column), accounting for the fact that the most eastward column can transition to the most westward column in a single step. At the same time, there is a state (2,2) where the cost of acting is sufficiently large that it is preferable to do nothing.

4.3 Exercises

Exercise 4.3.1. *Implement policy and value iteration for the environmental enforcement example of Exercise 3.4.5, and compare the number of iterations needed for convergence.*

Exercise 4.3.2. *Implement policy and value iteration for the random walk on the torus example of Example 4.2.6. Explore how the computational difficulty, and the rate of convergence, depends on the number of states* K^2 *of the problem.*

Exercise 4.3.3. *Consider a numerical method for solving an infinite-horizon discounted MDP where you have access to an approximation of the value function $\tilde{v}$, and $\|\tilde{v} - v\|_\infty < \eta$. This is used as an approximate terminal value at time T, and the problem is solved fully from time $t = T$ back to time $t = 0$. Give a bound on the error of this method.*

Exercise 4.3.4. *Consider a modification of the value iteration process, where instead of updating all states simultaneously, we iterate through the states and only update the value associated with each state in turn. Prove that this process gives a convergent approximation of the true value function.*

Exercise 4.3.5. *Consider the discounted MDP with state space $\mathcal{X} = \{1, 2, 3\}$ and discount factor $\gamma = 0.9$. We have an action space $\mathcal{U} = \{\text{Go}, \text{Stay}\}$; under action* Go*, the transitions and rewards are given in tabular form by*

State x	*Next state* $x' = \text{Go}(x)$	$r(x, \text{Go})$
1	2	5
2	3	2
3	1	0

Equivalently, the transition matrix and cost vector under $u = \text{Go}$ *are*

$$\mathbf{P}^{\text{Go}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{g}^{\text{Go}} = \begin{pmatrix} 5 \\ 2 \\ 0 \end{pmatrix}.$$

Under the action Stay*, we have*

$$\mathbf{P}^{\text{Stay}} = I, \qquad \mathbf{g}^{\text{Stay}} = 0.$$

(i) Compute the exact cost vector $J(\text{Go})$*, that is,* $J(u)$ *for the policy* $\mathrm{u}(x) = \text{Go}$ *for all* x*.*

With $\theta \in \mathbb{R}^2$*, we now approximate the cost function* $J(\text{Go})$ *using a linear architecture*

$$\widehat{J_\theta} = \Phi\theta,$$

with feature matrix

$$\Phi = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}.$$

Equivalently, we consider the approximation $\widehat{J_\theta}(x) = \theta_1 + \theta_2 x$*.*

(ii) Compute the least-squares projection of the true value $J(\text{Go})$ *onto* $\text{span}(\Phi)$*, that is,*

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^2} \|J(\text{Go}) - \Phi\theta\|_2^2.$$

Compute θ^* *and the approximate cost vector* $\widehat{J_\theta} = \Phi\theta^*$*. Evaluate the approximation errors*

$$\|J(\text{Go}) - \widehat{J_\theta}\|_\infty \quad \textit{and} \quad \|J(\text{Go}) - \widehat{J_\theta}\|_2.$$

(iii) For each state x, determine the greedy policy

$$\tilde{u}(x) \in \operatorname*{arg\,max}_{u\in\{\text{Go},\text{Stay}\}} \Big\{ g(x,u) + \gamma \sum_{s'} p(x';x,u)\widehat{J_\theta}(x'), \Big\}.$$

(iv) Using the policy error bound in Theorem 4.1.7, compute a numerical upper bound on the suboptimality of the greedy policy $\tilde{u}$ obtained in part (iii).

Exercise 4.3.6 (Policy iteration for the ergodic control problem)**.** *Consider a finite Markov decision process with state space $\mathcal{X}$ and finite action space $\mathcal{U}$. For each stationary deterministic policy $\mathbf{u} : \mathcal{X} \to \mathcal{U}$, let*

$$P^{\mathbf{u}}(x,x') = p(x';x,\mathbf{u}(x)), \qquad g^{\mathbf{u}}(x) = g(x,\mathbf{u}(x)).$$

Assume that for every stationary policy $\mathbf{u}$, the Markov chain with transition matrix $P^{\mathbf{u}}$ satisfies the geometric ergodicity of Assumption 3.2.3. We consider the infinite-horizon ergodic (average-cost) *control problem as in Section 3.2A*

$$\lambda^{\mathbf{u}} = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^{\pi}\Bigg[\sum_{t=0}^{T-1} g(X_t,\mathbf{u}(X_t)) \,\Big|\, X_0 = x\Bigg].$$

As $\lambda^{\mathbf{u}}$ is the long-run average cost, we know that it is independent of the initial state x.

For each policy $\mathbf{u}$, define the relative value function *$J^{\mathbf{u}}$ as the solution (unique up to an additive constant) of the Poisson equation*

$$\lambda^{\mathbf{u}} + J^{\mathbf{u}}(x) = \mathbf{g}^{\mathbf{u}}(x) + \sum_{x'\in\mathcal{X}} P^{\mathbf{u}}(x,x')\, J^{\mathbf{u}}(x'), \qquad x \in \mathcal{X},$$

with normalization condition $J^{\mathbf{u}}(x_0) = 0$ for some reference state $x_0 \in \mathcal{X}$.

Policy iteration for the ergodic problem proceeds as follows:

(i) Policy evaluation: *Given $\mathbf{u}_n$, compute (λ_n, J_n) solving*

$$\lambda_n + J_n(x) = \mathbf{g}^{\mathbf{u}_n}(x) + \sum_{x'} P^{\mathbf{u}_n}(x,x')\, J_n(x'), \qquad J_n(x_0) = 0.$$

(ii) Policy improvement: *Define a new policy $\mathbf{u}_{n+1}$ by*

$$\mathbf{u}_{n+1}(x) \in \operatorname*{arg\,min}_{u\in\mathcal{U}} \Big\{ g(x,u) + \sum_{x'} p(x';x,u)\, J_n(x') \Big\}.$$

(i) Prove that for every stationary policy $\mathbf{u}$, the Poisson equation admits a solution $(\lambda^{\mathbf{u}}, J^{\mathbf{u}})$ with $\lambda^{\mathbf{u}}$ equal to the average cost and $J^{\mathbf{u}}$ unique up to an additive constant.

(ii) *Show that if $\mathbf{u}_{n+1}$ is defined by the improvement step above, then*

$$\lambda_{n+1} \leq \lambda_n.$$

Hint: Use the defining equation for (λ_n, J_n) and compare it with the Bellman operator applied to J_n under $\mathbf{u}_{n+1}$.

(iii) *Show that if $\lambda_{n+1} = \lambda_n$, then $\mathbf{u}_n$ is already optimal.*

(iv) *Conclude that the sequence $\{\lambda_n\}$ produced by policy iteration is monotonically nonincreasing and, since there are only finitely many stationary deterministic policies when $\mathcal{U}$ is finite, the algorithm terminates in finitely many steps at an optimal policy.*

(v) *Explain why the contraction argument used in the discounted case cannot be applied directly here.*

Exercise 4.3.7 (Policy iteration for finite-horizon control)**.** *Consider a finite-horizon Markov decision process with state space $\mathcal{X}$, compact action space $\mathcal{U}$, horizon $T \in \mathbb{N}$, and terminal cost $\Phi : \mathcal{X} \to \mathbb{R}$.*

For $t = 0, 1, \ldots, T-1$, let $g_t(x,u)$ denote the running cost and $p_t(x'; x, u)$ the transition probabilities. A (feedback) policy is a sequence

$$\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_{T-1}), \qquad \pi_t : \mathcal{X} \to \mathcal{U}.$$

For a given policy $\boldsymbol{\pi}$ define the value functions

$$J_t^{\boldsymbol{\pi}}(x) = \mathbb{E}^{\boldsymbol{\pi}}\Big[\sum_{s=t}^{T-1} g_s(X_s, \pi_s(X_s)) + \Phi(X_T) \,\Big|\, X_t = x\Big],$$

which satisfy the backward recursion

$$J_t^{\boldsymbol{\pi}}(x) = g_t(x, \pi_t(x)) + \sum_{x'} p_t(x'; x, \pi_t(x))\, J_{t+1}^{\boldsymbol{\pi}}(x'), \qquad t = T-1, \ldots, 0,$$

with $J_T^{\boldsymbol{\pi}}(x) = \Phi(x)$. Policy iteration for the finite-horizon problem proceeds using the following two steps:

1. Policy evaluation: *Given a policy $\boldsymbol{\pi}^{(n)}$, compute $J_t^{(n)} := J_t^{\boldsymbol{\pi}^{(n)}}$ for $t = T, \ldots, 0$ by backward recursion.*

2. Policy improvement: *For each $t = T-1, \ldots, 0$ and each state $x \in \mathcal{X}$, define*

$$\pi_t^{(n+1)}(x) \in \arg\min_{u \in \mathcal{U}} \Big\{ g_t(x,u) + \sum_{x'} p_t(x'; x, u)\, J_{t+1}^{(n)}(x') \Big\}.$$

As usual, these steps are alternated until convergence.

(i) Show that for any policy $\boldsymbol{\pi}$*, the evaluation step uniquely determines* $J_t^{\boldsymbol{\pi}}$ *by backward induction.*

(ii) (Improvement lemma) By induction or otherwise, show that the improved policy satisfies

$$J_t^{(n+1)}(x) \leq J_t^{(n)}(x) \qquad \text{for all } t = 0, \dots, T \text{ and } x \in \mathcal{X}.$$

(iii) Show that if

$$\pi_t^{(n+1)}(x) = \pi_t^{(n)}(x) \quad \text{for all } t, x,$$

then $\boldsymbol{\pi}^{(n)}$ *is an optimal policy.*

(iv) Conclude that if $\mathcal{U}$ *is finite, then policy iteration terminates after finitely many iterations at an optimal policy.*

(v) Explain why, in contrast to the infinite-horizon discounted case, no contraction argument is needed for convergence of policy iteration in the finite-horizon setting.

Exercise 4.3.8 (Policy iteration with inexact policy improvement). *Consider a finite discounted MDP with state space* $\mathcal{X}$*, finite action space* $\mathcal{U}$*, and discount factor* $e^{-\rho} = \gamma \in (0,1)$*. Let* v^* *denote the optimal value function and* π^* *an optimal policy.*

In this exercise we consider policy iteration with an inexact *policy improvement step. With the usual Bellman optimality operator*

$$(\mathcal{T}v)(x) = \min_{u \in \mathcal{U}} \Big\{ g(x,u) + \gamma \sum_{x'} p(x'|x,u)\, v(x') \Big\}$$

and $v_n = J(\pi_n)$*, suppose that instead of choosing an exact minimiser, at each policy improvement step a policy* π_{n+1} *is found which satisfies the following condition: for some constant* $\mu > 0$*,*

$$\|v_n - \mathcal{T}v_n\|_\infty \ \geq \ \mu \, \|v_n - J(\pi_{n+1})\|_\infty. \tag{4.3}$$

That is, the decrease achieved by switching to π_{n+1} *is proportional to the Bellman residual of* v_n*.*

(a) Show that the Bellman residual controls suboptimality:

$$\|v_n - v^*\|_\infty \ \leq \ \frac{1}{1-\gamma} \|v_n - \mathcal{T}v_n\|_\infty.$$

(b) Using the contraction property of $\mathcal{T}$*, prove that*

$$\|J(\pi_{n+1}) - v^*\|_\infty \ \leq \ \gamma \|v_n - v^*\|_\infty + \|v_n - J(\pi_{n+1})\|_\infty.$$

(c) *Using* (4.3)*, show that there exists a constant* $q \in \mathbb{R}$ *(depending on* γ *and* μ*) such that*

$$\|J(\pi_{n+1}) - v^*\|_\infty \;\le\; q\,\|v_n - v^*\|_\infty.$$

Give an explicit expression for q*.*

(d) *Conclude that if* μ *is sufficiently large (so that* $q < 1$*), the sequence of policies produced by this inexact policy iteration scheme converges geometrically to the optimal policy in value, i.e.*

$$\|v_n - v^*\|_\infty \le q^n \|v_0 - v^*\|_\infty.$$

(e) *Discuss briefly how* (4.3) *relates to requiring the improvement step to make "sufficient progress" relative to the Bellman residual, and explain why an arbitrary approximate minimiser would not in general guarantee convergence.*

Chapter 5

Reinforcement 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...).

We assume that we have an agent who seeks to minimize the long-run discounted cost

$$J(U) = \mathbb{E}^U\Big[\sum_{t=0}^{\infty} e^{-\rho t} G(X_t, X_{t+1}, U_t)\Big].$$

We extend the control framework described in the previous chapters by now considering random costs G, which in general can depend on the state before and after the transition, the control chosen, and on independent randomness. To relate this to our earlier notation, we simply write

$$g(x,u) = \mathbb{E}^U\big[G(X_t, X_{t+1}, U_t)\big|X_t = x, U_t = u\big].$$

As before, the process X is assumed to be a controlled Markov chain (with finitely many states), however we do not assume we know the transition probabilities $p(\,\cdot\,; x, u)$.

We will attempt to learn an optimal policy based on observing examples of X and G. Suppose we have access to a collection of random variables $\{X_n, X_{n+}, U_n, G_n\}_{n\in\mathbb{N}}$, where

(i) $X_n \in \mathcal{X}$ and $U_n \in \mathcal{U}$ (sampled with some distribution),

(ii) $X_{n+} \sim p(\,\cdot\,; X_n, U_n)$ and $G_n \sim G(X_n, X_{n+}, U_n)$ (with an abuse of notation) and

(iii) (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, and do not need to sample from U which we believe should be close to the optimal action.

These distinctions lead to a variety of terms which get used to describe these methods:

(i) We say the method is *offline* if $\{X_n, X_{n+}, U_n, G_n\}_{n\in\mathbb{N}}$ is a fixed collection of observations that was computed following a given policy. In particular, the learning of the optimal policy happens “after” these samples were collected (so the choice of samples cannot be affected by the policy which is learned!).

(ii) We say the method is *online* if U_n is chosen based on previous (before timestep n) observations. For example, we could focus our estimation on actions we believe could be optimal, or where we have high uncertainty about the costs.

(iii) We say the method is *trajectory-based* if $X_{n+} = X_{n+1}$, that is, we observe a trajectory $X_1, X_2, ...$ obtained by following the controls $U_1, U_2, ...$ sequentially.

(iv) We say an online method is *on-policy* if U_n is chosen from our approximation of the optimal strategy. More precisely, suppose $\pi_n : \mathcal{X} \to \mathcal{P}(\mathcal{U})$ is a (randomized) strategy, and $U_n \sim \pi_n(\cdot|X_n)$; our method is online if π_n is also our approximation of the optimal strategy.

(v) We say an online method is *off-policy* if U_n is chosen from a distribution which does not need to approximate the optimal strategy. For example, if U_n is chosen uniformly at random (to ensure all actions are tried), even when this is clearly not the optimal thing to do.

See Chapters 5–6 in [42] more details on the standard terminology.

5.1 Q-learning

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 we saw briefly when we considered entropy regularized control. In general, the Q function is defined as the map $Q : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ which is a fixed point of the iteration

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

We observe that, for any $\alpha \in \mathbb{R}$, Q satisfies the identity

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

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. In particular, if we know Q, then a feedback control $\mathbf{u}$ is optimal if and only if $\mathbf{u}(x) \in \arg\min_u Q(x,u)$.

Definition 5.1.1. *With $\alpha_n \in [0,1)$ an $\{\mathcal{F}_n\}_{n\geq 0}$-adapted learning rate process, the Q-learning iteration is defined by*

$$Q_{n+1}(x,u) = Q_n(x,u)+\alpha_n\Big[G_n+e^{-\rho}\min_{u'} Q_n(X_{n+},u')-Q_n(x,u)\Big]\,\mathbf{1}_{\{X_n=x,U_n=u\}}\,.$$

Next, we state (and prove) one of the most powerful results in Q-learning [44], which establishes the following convergence result.

Theorem 5.1.2. *Consider a control problem with finitely many states and 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

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

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

In order to prove this theorem, we will make use of the following stochastic approximation result, the (rather fiddly) proof of which is in Appendix A.1B

Lemma 5.1.3. *Consider adapted random processes $\{Y_n\}_{n\in\mathbb{N}}$, $\{Z_n\}_{n\in\mathbb{N}}$ and $\{\alpha_n\}_{n\in\mathbb{N}}$ with values in $\mathbb{R}^m$, where Y_n has dynamics (for each i an index in $\{1,...,m\}$)*

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

and where α_i and Z_i are such that for all $i \in \{1,\dots,m\}$ and $n \in \mathbb{N}$, *(i)* $\alpha_n(i) \in [0,1]$, $\sum_{n\in\mathbb{N}} \alpha_n(i) = \infty$, $\sum_{n\in\mathbb{N}} \alpha_n^2(i) < \infty$,

(ii) $|\mathbb{E}[Z_{n+1}(i)|\mathcal{F}_n]| \leq \gamma \|Y_n\|_\infty$, *with* $\gamma < 1$,

(iii) $\mathbb{V}[Z_{n+1}(i)|\mathcal{F}_n] \leq c(1+\|Y_n\|_\infty^2)$ *for* $c > 0$.

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

Proof of Theorem 5.1.2. Take Q_n to be the sequence generated by Q-learning, and Q to be the true Q-function. Define

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

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

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

and thus

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

Taking an absolute value, we have the bound

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

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

$$\begin{aligned} &(Z_{n+1}(x,u))^2 \\ &\quad \leq 2\Big(G_n - Q(x,u) + e^{-\rho}Q(X_{n+},u^*)\Big)^2 + 2e^{-\rho}\Big(Q_n(X_{n+},u^*) - Q(X_{n+},u^*)\Big)^2, \end{aligned}$$

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

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

for some $c > 0$. Combining with our previous bound, we get the desired growth bound on the variance. As $e^{-\rho} < 1$, applying Lemma 5.1.3 we conclude that $\|Y_n\|_\infty \to 0$ a.s., as desired. □

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 above 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 5.1.4. 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^1 or L^2.

Remark 5.1.5. Particularly in offline settings, where we use Q-learning to calculate controls based on a fixed dataset, there is a concern that arises due to some state-action pairs not being frequently observed in training data: If there are statistical errors resulting in an overly optimistic estimate of the Q function, then the learned policy will tend to encourage these actions, resulting in poor out-of-sample performance. This leads to modifications of the Q-learning algorithm which try and penalize the value associated with states with high uncertainty, such as conservative Q-learning [34]. This closely links with ideas in robust and risk-averse control theory (see, for example, [26, 21]).

5.1A SARSA and Actor–Critic methods

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, $S_t, A_t, R_t, S_{t+}, A_{t+}$ reads $X_t, U_t, G_t, X_{t+}, U_{t+}$, and the (optimized) Q-learning iteration is replaced by the SARSA iteration

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

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

$$\begin{aligned} Q^\pi(x,u) &= \mathbb{E}\bigg[\sum_{t=0}^\infty e^{-\rho t} G(X_t, X_{t+1}, U_t)\,\bigg|\, X_0 = x, U_0 = u, \text{ and} \qquad (5.1)\\ &\qquad\qquad \text{for } t\geq 0, U_{t+1}\sim \pi(X_{t+1}), X_{t+1}\sim p(\cdot; X_t, U_t)\bigg]\\ &= \mathbb{E}\bigg[G(X_t, X_{t+1}, U_t) + e^{-\rho} J(X_{t+1}, \pi)\,\bigg|\, X_0 = x, U_0 = u, \text{ and}\\ &\qquad\qquad \text{for } t\geq 0, U_{t+1}\sim \pi(X_{t+1}), X_{t+1}\sim p(\cdot; X_t, U_t)\bigg]. \qquad (5.2)\end{aligned}$$

From this, we see that the cost function $J(x,\pi) = \mathbb{E}[Q^\pi(x,u)|u\sim\pi]$.

This iteration can be used as the basis for policy iteration methods, 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. In particular, we can define an "actor–critic" method, where we have two approximations involving in tandem:

(i) a 'critic', who seeks to evaluate the current policy π, that is, who computes Q^π;

(ii) an 'actor', who seeks to choose an optimal policy π given the current evaluation Q^π, that is, who computes $\pi(u|x)\approx \arg\min_u Q^\pi(x,u)$, for example by taking $\pi_{n+1}(u|x) = \frac{\exp(-\eta Q^{\pi_n}(x,u))}{\sum_{u'}\exp(-\eta Q^{\pi_n}(x,u'))}$, for some $\eta>0$.

By choosing how frequently to update the actor and critic, and how to choose η, we obtain a method that corresponds to variations of policy iteration, but now based on our noisy observations.

5.1B TD learning

Particularly given the comparison with policy iteration, it's worth considering whether there are methods which have better convergence properties than the simple SARSA iteration. The SARSA update can be viewed as a *temporal-difference (TD)* method, in which one observes a one-step target and nudges our current estimate of the Q^π function towards it. Define the TD error by

$$\delta_n := G_n + e^{-\rho}\,\mathcal{B}_n(Q_n; X_{n+}) \;-\; Q_n(X_n, U_n), \qquad (5.3)$$

where $\mathcal{B}_n(Q_n; X_{n+})$ is a quantity built from Q_n at the next state X_{n+}. Then most TD rules can be written in the generic form

$$Q_{n+1}(x,u) = Q_n(x,u) + \alpha_n \, \delta_n \, \mathbf{1}_{\{(X_n,U_n)=(x,u)\}}. \tag{5.4}$$

The difference between algorithms is the choice of the bootstrap $\mathcal{B}_n$ and how the samples $(X_n, U_n, G_n, X_{n+}, U_{n+})$ are generated.

The following is a far-from-exhaustive list of variants:

$[\mathrm{SC}]_3$:add references for these

(a) On-policy SARSA to learn Q^π: we use π to generate $U_n \sim \pi(\cdot \mid X_n)$ and $U_{n+} \sim \pi(\cdot \mid X_{n+})$, then $\mathcal{B}_n$ becomes

$$\mathcal{B}_n(Q_n^\pi; X_{n+}) := Q_n^\pi(X_{n+}, U_{n+}). \tag{5.5}$$

(b) Expected SARSA: to reduce variance we may replace the sampled next action by its conditional expectation, that is,

$$\mathcal{B}_n(Q_n; X_{n+}) := \sum_{u' \in \mathcal{U}} \pi(u' \mid X_{n+}) \, Q_n(X_{n+}, u'). \tag{5.6}$$

(c) m-step TD: Suppose we have a trajectory of observations, with associated controls. A classical extension of SARSA is to consider m-step TD (instead of just one-step). This corresponds to using the discounted costs over the future m steps when updating the value associated to the initial state. This allows us to average out the randomness of costs, and the discount term reduces our dependence on the (noisy) prior approximation of the Q function in our iteration. Using the truncated return

$$G_n^{(m)} := \sum_{k=0}^{m-1} e^{-\rho k} \, G_{n+k}. \tag{5.7}$$

Then one replaces (5.3) by

$$\delta_n^{(m)} := G_n^{(m)} + e^{-\rho m} \, \mathcal{B}_n^{(m)}(Q_{n+m}; X_{(n+m)+}) - Q_n(X_n, U_n), \tag{5.8}$$

where $\mathcal{B}_n^{(m)}$ is chosen as in SARSA or expected SARSA, and updates are as in (5.4) with $\delta_n^{(m)}$.

(d) TD(λ): When we use m-step TD, we observe costs over multiple steps and use them only to update the Q function at the initial value for each block of m steps. This seems inefficient, as we also have information about the interim states that were observed. An alternative is to update the Q function in all states, depending on how recently each state has been seen in the data.

The key idea is to maintain a trace $e_n(x,u)$ with recursion

$$e_n(x,u) := e^{-\rho} \lambda \, e_{n-1}(x,u) + \mathbf{1}_{\{(X_n,U_n)=(x,u)\}}, \qquad \lambda \in [0,1], \tag{5.9}$$

and update all state-action pairs as

$$Q_{n+1}(x,u) = Q_n(x,u) + \alpha_n\, \delta_n\, e_n(x,u), \tag{5.10}$$

where δ_n is the one-step TD error for the chosen bootstrap. This method is known as TD(λ).

The appeal here is that $e(x,u)$ keeps track of the states and actions that have recently contributed to the current nudge, where "recently" is modulated by the value of the parameter λ. The smaller the value of λ, the more we ignore the trajectory. In the limit when $\lambda = 0$ we recover SARSA.

5.2 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).

Let $\mathcal{U}_0$ be a set of finitely many actions. We suppose that we have an agent who is using a randomized policy $\pi_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{U}_0) = \mathcal{U}$, parametrized by θ. 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 in Section 3.3, is to take $\pi(u|x) \propto e^{\theta(x,u)}$, which is valid for all $\theta \in \mathbb{R}^{|\mathcal{U}_0|\times|\mathcal{X}|}$. More precisely, for each $u \in \mathcal{U}$ and $x \in \mathcal{X}$ (both finite sets), there is a real number associated with the (u,x) pair which we denote $\theta(u,x) \in \mathbb{R}$ and let

$$\pi(u|x) = \frac{e^{\theta(u,x)}}{\sum_{u'} e^{\theta(u',x)}}\,. \tag{5.11}$$

It follows that

$$\log \pi(u|x) = \exp\big(\theta(u,x)\big) - \log\Big(\sum_{u'} \exp\big(\theta(u',x)\big)\Big),$$

and the partial with respect to $\theta(x,u)$ is

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

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 or 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. First, we abuse our previous notation and write $g(x_t, x_{t+1}, u_t)$ for the expected value of $G(X_t, X_{t+1}, U_t)$ given $X_t = x_t$, $X_{t+1} = x_{t+1}$, and $U_t = u_t$. If we use the randomized control π_θ, then the probability of a sequence $\{x_0, u_0, \dots, x_T, u_T, x_{T+1}\}$ is given by

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

In particular, we can write

$$\begin{aligned} J(x_0, \pi_\theta) &= \sum_T \mathbb{E}\Big[e^{-\rho T} G_T \Big| X_0 = x_0, U_t \sim \pi_\theta(X_t)\Big] \\ &= \sum_T \sum_{\{x_t, u_t\}_{t \le T}} \Big[\prod_{t=0}^{T} p(x_{t+1}; x_t, u_t)\pi_\theta(u_t|x_t) e^{-\rho T} g(x_T, x_{T+1}, u_T)\Big]. \end{aligned} \tag{5.12}$$

This leads us to the following result.

Theorem 5.2.1 (Policy Gradient Theorem)**.** *The gradient of J with respect to θ is*

$$\nabla_\theta J(x_0, \pi_\theta) = \sum_T \mathbb{E}\Big[e^{-\rho T}\Big(\sum_{s \le T} \nabla_\theta \log \pi_\theta(U_s|X_s)\Big) G_T \Big| X_0 = x, U_t \sim \pi_\theta(X_t)\Big].$$

Proof. First, use (5.12) to obtain

$$\nabla_\theta J(x_0, \pi_\theta) = \sum_T \sum_{\{x_t, u_t\}_{t \le T}} \nabla_\theta \Big[\prod_{t=0}^{T} p(x_{t+1}; x_t, u_t)\pi_\theta(u_t|x_t) e^{-\rho T} g(x_T, x_{T+1}, u_T)\Big].$$

Next, we have 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,

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

Therefore, using this result together with

$$\nabla_\theta \log \Big(p(x_{t+1}; x_t, u_t)\pi_\theta(u_t|x_t)e^{-\rho T} g(x_T, x_{T+1}, u_T)\Big) = \nabla_\theta \log \pi_\theta(u_t|x_t)\,,$$

we obtain

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

□

This is sometimes called the 'fundamental lemma of policy gradients' because 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 $U_t \sim \pi_\theta(X_t)$)

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

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:

(i) 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.

(ii) Estimate the corresponding average gradients through

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

(iii) 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.

5.2A Convergence of policy gradients

The details under which the policy gradient method converges to the optimal solution can be somewhat subtle, see for example Bhandari and Russo [9] for recent results in this area. Here, we follow [1] to give a positive convergence result in a specific case, when we use the 'natural policy gradient', corresponding to a specific choice of preconditioning in our iteration.

Let $\pi_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{U}_0)$, with $\mathcal{U} = \mathcal{P}(\mathcal{U}_0)$, be a policy parametrized as in (5.11). For notational simplicity, let $J^\pi : \mathcal{X} \to \mathbb{R}$ be defined by

$$J^\pi(x) = J(x,\pi) = \mathbb{E}\Big[\sum_{t=0}^{\infty} e^{-\rho t} G(X_t, X_{t+1}, U_t)\,\Big|\, X_0 = x, \text{ and } \\ \text{for } t \geq 0, U_t \sim \pi(X_t), X_{t+1} \sim p(\cdot; X_t, U_t)\Big]$$

and $Q^\pi(x,u)$ as in (5.1). If instead of x we evaluate J^π at an initial state distribution $\Xi \in \mathcal{P}(\mathcal{X})$, we let

$$J^\pi(\Xi) := \mathbb{E}_{x\sim\Xi}[J^\pi(x)]\,. \tag{5.13}$$

The *advantage* function $A^\pi : \mathcal{X} \times \mathcal{U}_0 \to \mathbb{R}$ is defined as

$$A^\pi(x,u) = Q^\pi(x,u) - J^\pi(x)\,.$$

Finally, for $x_0 \in \mathcal{X}$ and $\pi \in \mathcal{U}$, we let $d^\pi_{x_0}$ be the *discounted state occupation distribution*, defined as

$$d^\pi_{x_0}(x) = (1 - e^{-\rho}) \sum_{t=0}^{\infty} e^{-\rho\, t} \mathbb{P}\Big(X_t = x \,\Big|\, X_0 = x_0, \text{ and for } r \geq 0, \\ U_r \sim \pi(X_r), X_{r+1} \sim p(\cdot; X_r, U_r)\Big).$$

Similar to (5.13), for $\Xi \in \mathcal{P}(\mathcal{X})$, we let

$$d^{\pi}_{\Xi}(x') := \mathbb{E}_{x\sim\Xi}[d^{\pi}_{x}(x')]\,.$$

With this definition, and using Theorem 5.2.1, we have the following helpful lemma, which shows that the gap between the values of two policies $\pi, \tilde{\pi}$ can be written as an expectation of the $\tilde{\pi}$-advantage, but evaluated along trajectories generated by π.

Lemma 5.2.2 (Performance difference lemma [31])**.** *Let $\pi, \tilde{\pi} \in \mathcal{U} = \mathcal{P}(\mathcal{U}_0)$. We have that for all $x_0 \in \mathcal{X}$*

$$J^{\pi}(x_0) - J^{\tilde{\pi}}(x_0) = \frac{1}{1-\gamma}\mathbb{E}_{x\sim d^{\pi}_{x_0}}\Big[\mathbb{E}_{u\sim\pi(\cdot|x)}\Big[A^{\tilde{\pi}}(x,u)\Big]\Big]\,,$$

with $\gamma := e^{-\rho} \in (0,1)$.

Proof. Fix $x_0 \in \mathcal{X}$ and two policies $\pi, \tilde{\pi} \in \mathcal{U}$. Consider a trajectory $(X_t, U_t)_{t\geq 0}$ generated by π from $X_0 = x_0$ (i.e. $U_t \sim \pi(X_t)$ and $X_{t+1} \sim p(\cdot; X_t, U_t)$). By the definition of $Q^{\tilde{\pi}}$, for every (x,u),

$$Q^{\tilde{\pi}}(x,u) = \mathbb{E}\left[\,G(X_0, X_1, U_0) + \gamma\, J^{\tilde{\pi}}(X_1)\,\Big|\, X_0 = x,\ U_0 = u\right], \tag{5.14}$$

where the conditional law of X_1 is $p(\cdot; x, u)$. Hence, conditioning on (X_t, U_t) and using (5.14),

$$\begin{aligned}\mathbb{E}\left[A^{\tilde{\pi}}(X_t,U_t)\,\middle|\, X_t, U_t\right] &= \mathbb{E}\left[Q^{\tilde{\pi}}(X_t,U_t) - J^{\tilde{\pi}}(X_t)\,\middle|\, X_t, U_t\right]\\ &= \mathbb{E}\left[G(X_t, X_{t+1}, U_t) + \gamma J^{\tilde{\pi}}(X_{t+1}) - J^{\tilde{\pi}}(X_t)\,\middle|\, X_t, U_t\right].\end{aligned}$$

Taking expectations and summing (with weights $\gamma^t = e^{-\rho t}$) we obtain

$$\begin{aligned}&\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}}(X_t,U_t)\Big]\\ &\quad= \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t G(X_t,X_{t+1},U_t)\Big] + \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^{t+1} J^{\tilde{\pi}}(X_{t+1})\Big] - \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t J^{\tilde{\pi}}(X_t)\Big]\\ &\quad= J^{\pi}(x_0) + \mathbb{E}\Big[\sum_{t=1}^{\infty}\gamma^t J^{\tilde{\pi}}(X_t)\Big] - \Big(J^{\tilde{\pi}}(x_0) + \mathbb{E}\Big[\sum_{t=1}^{\infty}\gamma^t J^{\tilde{\pi}}(X_t)\Big]\Big)\\ &\quad= J^{\pi}(x_0) - J^{\tilde{\pi}}(x_0).\end{aligned}$$

Therefore,

$$J^{\pi}(x_0) - J^{\tilde{\pi}}(x_0) = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}}(X_t,U_t)\Big]. \tag{5.15}$$

Next, rewrite the right-hand side to obtain

$$\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t A^{\tilde{\pi}}(X_t,U_t)\Big] = \sum_{t=0}^{\infty}\gamma^t \mathbb{E}\Big[\mathbb{E}_{u\sim\pi(\cdot|X_t)}\left[A^{\tilde{\pi}}(X_t,u)\right]\Big].$$

Using the definition of the discounted occupation measure $d^{\pi}_{x_0}$, we have that for any function $f : \mathcal{X} \to \mathbb{R}$

$$\sum_{t=0}^{\infty} \gamma^t \mathbb{E}[f(X_t)] = \frac{1}{1-\gamma}\, \mathbb{E}_{x \sim d^{\pi}_{x_0}}[f(x)].$$

Applying this with $f(x) = \mathbb{E}_{u \sim \pi(\cdot|x)}[A^{\tilde{\pi}}(x,u)]$ and combining with (5.15) yields

$$\begin{aligned} J^{\pi}(x_0) - J^{\tilde{\pi}}(x_0) &= \frac{1}{1-\gamma}\, \mathbb{E}_{x \sim d^{\pi}_{x_0}}\Big[\mathbb{E}_{u \sim \pi(\cdot|x)}\Big[A^{\tilde{\pi}}(x,u)\Big]\Big] \\ &= \frac{1}{1-e^{-\rho}}\, \mathbb{E}_{x \sim d^{\pi}_{x_0}}\Big[\mathbb{E}_{u \sim \pi(\cdot|x)}\Big[A^{\tilde{\pi}}(x,u)\Big]\Big], \end{aligned}$$

which is the desired identity. □

Next, we focus on the *natural policy gradient* (NPG) algorithm. The key distinction is that the gradient updates are modulated by the Fisher information matrix induced by π_θ. More precisely, let

$$F_{\Xi}(\theta) = \mathbb{E}_{x \sim d^{\pi}_{\Xi}}\Big[\mathbb{E}_{u \sim \pi_\theta(u|x)}\Big[\Big(\nabla_\theta \log\big(\pi_\theta(u|x)\big)\Big)\Big(\nabla_\theta \log\big(\pi_\theta(u|x)\big)\Big)^{\intercal}\Big]\Big], \quad (5.16)$$

and let the t-th update of parameters θ be given by

$$\theta^{(t+1)} = \theta^{(t)} - \eta\, F_{\Xi}\big(\theta^{(t)}\big)^{\dagger}\, \nabla_\theta V^{(t)}(\Xi)\,, \quad (5.17)$$

where the superscript † denotes the Moore–Penrose pseudoinverse.[1] In what follows, we use '(t)' in the superscript to denote that we are using the parameters from the t-th iteration.

Lemma 5.2.3. *For the softmax policy* (5.11)*, the NPG update in* (5.17) *becomes*

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta}{1-e^{-\rho}}\, A^{(t)},$$

and

$$\pi^{(t+1)}(u|x) = \pi^{(t)}(u|x) \frac{\exp\Big(-\eta\, A^{(t)}(x,u)/\big(1-\exp(-\rho)\big)\Big)}{Z_t(x)}, \quad (5.18)$$

with

$$Z_t(x) = \sum_u \pi^{(t)}(u|x)\, \exp\Big(-\eta\, A^{(t)}(x,u)/\big(1-\exp(-\rho)\big)\Big).$$

[1] This is essentially the inverse of the Fisher information matrix, but accounts for the fact that in the softmax formulation, the parameters $\theta(u,x) + \psi(x)$ result in the same strategies as $\theta(u,x)$ (for any function $\psi : \mathcal{X} \to \mathbb{R}$), which results in the Fisher information matrix not being invertible in these directions.

Proof. Let $t \geq 0$; for notational simplicity let us write $\pi^{(t)} = \pi_{\theta^{(t)}}$, $J^{(t)} = J^{\pi^{(t)}}$, $G_T := G(X_T, X_{T+1}, U_T),$, $Q^{(t)} = Q^{\pi^{(t)}}$, $A^{(t)}(x,u) = Q^{(t)}(x,u) - J^{(t)}(x)$, and $d^{(t)}(x) := d_{\Xi}^{\pi^{(t)}}(x)$. For the softmax/logit parametrization (5.11) we have that for each coordinate $\theta(u', x')$

$$\partial_{\theta(u',x')} \log \pi_\theta(u|x) = \mathbf{1}_{\{x=x'\}} \Big(\mathbf{1}_{\{u=u'\}} - \pi_\theta(u'|x) \Big). \tag{5.19}$$

Apply Theorem 5.2.1 and take the $\theta(u', x')$-coordinate. We obtain

$$\begin{aligned} &\partial_{\theta(u',x')} J(x_0, \pi_\theta) && (5.20)\\ &\quad = \sum_{T=0}^{\infty} \mathbb{E}\left[\gamma^T \Big(\sum_{s=0}^{T} \partial_{\theta(u',x')} \log \pi_\theta(U_s|X_s) \Big) G_T \,\middle|\, X_0 = x_0,\ U_t \sim \pi_\theta(X_t) \right] \\ &\quad = \sum_{T=0}^{\infty} \sum_{s=0}^{T} \mathbb{E}\Big[\gamma^T \, \partial_{\theta(u',x')} \log \pi_\theta(U_s|X_s) \, G_T \,\Big|\, X_0 = x_0,\ U_t \sim \pi_\theta(X_t) \Big]. \\ & && (5.21) \end{aligned}$$

Swap the order of summation (justified by discounting and the boundedness assumptions) to obtain that $\partial_{\theta(u',x')} J(x_0, \pi_\theta)$ becomes

$$\sum_{s=0}^{\infty} \mathbb{E}\left[\Big(\sum_{T=s}^{\infty} \gamma^T G_T \Big) \partial_{\theta(u',x')} \log \pi_\theta(U_s|X_s) \,\middle|\, X_0 = x_0,\ U_t \sim \pi_\theta(X_t) \right]. \tag{5.22}$$

Next, condition on (X_s, U_s) and use the Markov property to see that

$$\begin{aligned} \mathbb{E}\Big[\sum_{T=s}^{\infty} \gamma^T G_T \,\Big|\, X_s = x,\ U_s = u \Big] &= \gamma^s \, \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k G_{s+k} \,\Big|\, X_s = x,\ U_s = u \Big] \\ &= \gamma^s \, Q^{\pi_\theta}(x,u), \end{aligned}$$

where Q^{π_θ} is the state–action value function as in (5.1). Plugging this into (5.22) yields

$$\partial_{\theta(u',x')} J(x_0, \pi_\theta) = \sum_{s=0}^{\infty} \mathbb{E}\Big[\gamma^s \, Q^{\pi_\theta}(X_s, U_s) \, \partial_{\theta(u',x')} \log \pi_\theta(U_s|X_s) \Big]. \tag{5.23}$$

Next, for each s, condition on X_s and use (5.19) to get

$$\begin{aligned} &\mathbb{E}\Big[Q^{\pi_\theta}(X_s, U_s) \, \partial_{\theta(u',x')} \log \pi_\theta(U_s|X_s) \,\Big|\, X_s \Big] && (5.24)\\ &\quad = \mathbf{1}_{\{X_s = x'\}} \, \mathbb{E}_{u \sim \pi_\theta(\cdot|x')} \Big[Q^{\pi_\theta}(x', u) \big(1_{\{u=u'\}} - \pi_\theta(u'|x') \big) \Big] \\ &\quad = \mathbf{1}_{\{X_s = x'\}} \, \pi_\theta(u'|x') \Big(Q^{\pi_\theta}(x', u') - J^{\pi_\theta}(x') \Big) \\ &\quad = \mathbf{1}_{\{X_s = x'\}} \, \pi_\theta(u'|x') \, A^{\pi_\theta}(x', u'). && (5.25) \end{aligned}$$

In the second line we used $\mathbb{E}_{u\sim\pi_\theta(\cdot|x')}[Q^{\pi_\theta}(x',u)] = J^{\pi_\theta}(x')$. Taking expectations in (5.25) and substituting into (5.23) gives

$$\partial_{\theta(u',x')} J(x_0,\pi_\theta) = \pi_\theta(u'|x')\, A^{\pi_\theta}(x',u') \sum_{s=0}^{\infty} \gamma^s\, \mathbb{P}(X_s = x' \mid X_0 = x_0, \pi_\theta). \tag{5.26}$$

Finally, using the definition of the discounted occupation distribution,

$$d^{\pi_\theta}_{x_0}(x') = (1-\gamma) \sum_{s=0}^{\infty} \gamma^s\, \mathbb{P}(X_s = x' \mid X_0 = x_0, \pi_\theta),$$

we rewrite (5.26) as

$$\partial_{\theta(u',x')} J(x_0,\pi_\theta) = \frac{1}{1-\gamma}\, d^{\pi_\theta}_{x_0}(x')\, \pi_\theta(u'|x')\, A^{\pi_\theta}(x',u'). \tag{5.27}$$

Next, fix a state x and let $\pi_x \in \mathbb{R}^{|\mathcal{U}_0|}$ denote the vector $\pi_x(u) := \pi_\theta(u|x)$. We observe that the coordinates associated with different states do not affect π_x, which simplifies our analysis a little. From (5.19), the score vector (restricted to the coordinates $\theta(\cdot,x)$) is

$$\nabla_{\theta(\cdot,x)} \log \pi_\theta(U|x) \;=\; e_U - \pi_x,$$

where e_U is the canonical basis vector for the sampled action $U \sim \pi_x$ (that is, e_U has zero entries except for the U-th entry, which is a one). Therefore the conditional Fisher block at x is the covariance matrix

[SC]$_4$:fix from here

$$\mathbb{E}_{U\sim\pi_x}\Big[(e_U - \pi_x)(e_U - \pi_x)^\top\Big] \;=\; \mathrm{Diag}(\pi_x) - \pi_x\pi_x^\top \;=:\; C(\pi_x). \tag{5.28}$$

Since $F_\Xi(\theta)$ averages over $x \sim d^{\pi_\theta}_\Xi$, the full Fisher matrix is block-diagonal across states

$$F_\Xi(\theta) = \mathbb{E}_{x\sim d^{\pi_\theta}_\Xi}\big[\, C(\pi_\theta(\cdot|x))\,\big], \qquad \big(F_\Xi(\theta)\big)_{x\text{-block}} = d^{\pi_\theta}_\Xi(x)\, C(\pi_\theta(\cdot|x)). \tag{5.29}$$

Moreover, $C(\pi_x)$ is always positive semidefinite; it has a one–dimensional null space $\{c\mathbf{1}\}_{c\in\mathbb{R}}$, corresponding to the previously observed fact that adding the same constant to all logits at a fixed state does not change the softmax probabilities.

Fix x and write $C_x := C(\pi^{(t)}(\cdot|x))$ and $g_x := \nabla_{\theta(\cdot,x)} J^{(t)}(\Xi)$. By (5.29)

$$g_x = \frac{d^{(t)}(x)}{1-\gamma}\big(\pi^{(t)}(\cdot|x) \odot A^{(t)}(x,\cdot)\big), \qquad (F_\Xi(\theta^{(t)}))_{x\text{-block}} = d^{(t)}(x)\, C_x.$$

Hence, using $(cM)^\dagger = \frac{1}{c}M^\dagger$ for $c > 0$,

$$\big(F_\Xi(\theta^{(t)})^\dagger \nabla_\theta J^{(t)}(\Xi)\big)_{x\text{-block}} = \frac{1}{d^{(t)}(x)}\, C_x^\dagger\, g_x = \frac{1}{1-\gamma}\, C_x^\dagger\big(\pi^{(t)}(\cdot|x) \odot A^{(t)}(x,\cdot)\big). \tag{5.30}$$

We now use the identity

$$C(\pi_x)^\dagger\big(\pi_x \odot a\big) = a \quad \text{whenever} \quad \langle \pi_x, a\rangle := \sum_u \pi_x(u)a(u) = 0. \tag{5.31}$$

Here we simply pick a convenient representative of a by removing any irrelevant additive constant: since adding a multiple of $\mathbf{1}$ does not affect the quantities of interest, we may (and will) work with the centered version of a, characterized by $\langle \pi_x, a\rangle = 0$. To verify (5.31), note that if $\langle \pi_x, a\rangle = 0$, then

$$C(\pi_x)a = \big(\mathrm{Diag}(\pi_x) - \pi_x\pi_x^\top\big)a = \pi_x \odot a - \pi_x\langle \pi_x, a\rangle = \pi_x \odot a.$$

Thus a is a solution to $C(\pi_x)z = \pi_x \odot a$, and all solutions are of the form $a + c\,\mathbf{1}$. The Moore–Penrose pseudoinverse selects the minimum-norm solution in $\mathrm{Range}(C(\pi_x)) = \mathbf{1}^\perp$, which is precisely a (since $a \perp \mathbf{1}$ under the constraint $\langle \pi_x, a\rangle = 0$, and Pythagoras' identity shows this achieves the minimum norm). This proves (5.31).

Since advantages are centered, $\langle \pi^{(t)}(\cdot|x), A^{(t)}(x,\cdot)\rangle = 0$, we may apply (5.31) to (5.30) and obtain, for every (x,u),

$$\big(F_\Xi(\theta^{(t)})^\dagger \nabla_\theta J^{(t)}(\Xi)\big)(u,x) = \frac{1}{1-\gamma}\,A^{(t)}(x,u).$$

Therefore, the natural-gradient step (for minimizing J) becomes

$$\theta^{(t+1)} = \theta^{(t)} - \eta\, F_\Xi(\theta^{(t)})^\dagger \nabla_\theta J^{(t)}(\Xi) = \theta^{(t)} - \frac{\eta}{1-\gamma}\,A^{(t)} = \theta^{(t)} - \frac{\eta}{1-e^{-\rho}}\,A^{(t)},$$

which is the first claimed identity.

By the softmax formula,

$$\begin{aligned}\pi^{(t+1)}(u|x) = \frac{\exp(\theta^{(t+1)}(u,x))}{\sum_v \exp(\theta^{(t+1)}(v,x))} &= \frac{\exp(\theta^{(t)}(u,x))\exp\Big(-\frac{\eta}{1-\gamma}A^{(t)}(x,u)\Big)}{\sum_v \exp(\theta^{(t)}(v,x))\exp\Big(-\frac{\eta}{1-\gamma}A^{(t)}(x,v)\Big)}\\ &= \pi^{(t)}(u|x)\;\frac{\exp\Big(-\frac{\eta}{1-\gamma}A^{(t)}(x,u)\Big)}{\sum_v \pi^{(t)}(v|x)\exp\Big(-\frac{\eta}{1-\gamma}A^{(t)}(x,v)\Big)}.\end{aligned}$$

Thus the update is a multiplicative-weights (exponentiated-gradient) step with normalizer

$$Z_t(x) := \sum_v \pi^{(t)}(v|x)\exp\Big(-\frac{\eta}{1-\gamma}A^{(t)}(x,v)\Big).$$

□

With the above, we have the following convergence result.

Theorem 5.2.4 (Global convergence for NPG (c.f. Theorem 5.3 in [1]))**.** *Fix $\eta > 0$ and $\Xi \in \mathcal{P}(\mathcal{X})$. Let $\theta^{(0)} = 0$, and for $t \geq 0$ let $\theta^{(t+1)}$ satisfy* (5.17)*. For all $T > 0$ we have that*

$$V^{(T)}(\Xi) \leq V^*(\Xi) + \frac{\log(|\mathcal{U}_0|)}{\eta T} + \frac{1}{\big(1 - \exp(-\rho)\big)^2 T}\,.$$

In particular, if $\eta \geq (1-\exp(-\rho))^2 \log(|\mathcal{U}_0|)$, we see that NPG finds an ϵ-optimal policy in a number of iterations T bounded by

$$T \leq \frac{2}{(1 - \exp(-\rho))^2\, \epsilon}\,.$$

Proof. First, we have that for $x \in \mathcal{X}$ and using Jensen's inequality,

$$\begin{aligned}\log(Z_t(x)) &= \log\Big(\sum_u \pi^{(t)}(u|x)\, \exp\big(-\eta\, A^{(t)}(x,u)\big)/(1-\gamma)\Big)\\ &\geq \sum_u \pi^{(t)}(u|x)\, \log\exp\big(-\eta\, A^{(t)}(x,u)\big)/(1-\gamma)\,.\end{aligned}$$

From here, we see that

$$\log(Z_t(x)) \geq -\frac{\eta}{1-\gamma} \sum_u \pi^{(t)}(u|x)\, A^{(t)}(x,u) = 0\,.$$

This allows us to obtain an improvement lower bound for the gradient updates. We proceed as follows. Rewrite (5.18) as

$$A^{(t)}(x,u) = -\frac{1-\gamma}{\eta}\, \log\Big(\frac{\pi^{(t+1)}(u|x)\, Z_t(x)}{\pi^{(t)}(u|x)}\Big). \tag{5.32}$$

Plugging (5.32) into the performance difference lemma (with $\pi = \pi^{(t+1)}$ and $\tilde{\pi} = \pi^{(t)}$) yields

$$\begin{aligned}V^{(t+1)}(\Xi) - V^{(t)}(\Xi) &= \frac{1}{1-\gamma}\, \mathbb{E}_{x\sim d_\Xi^{\pi^{(t+1)}}}\Big[\sum_u \pi^{(t+1)}(u|x)\, A^{(t)}(x,u)\Big]\\ &= -\frac{1}{\eta}\, \mathbb{E}_{x\sim d_\Xi^{\pi^{(t+1)}}}\Big[\sum_u \pi^{(t+1)}(u|x) \log\frac{\pi^{(t+1)}(u|x)}{\pi^{(t)}(u|x)} + \log Z_t(x)\Big]\\ &= -\frac{1}{\eta}\, \mathbb{E}_{x\sim d_\Xi^{\pi^{(t+1)}}}\Big[\mathrm{KL}\big(\pi^{(t+1)}(\cdot|x)\,\|\,\pi^{(t)}(\cdot|x)\big)\Big] - \frac{1}{\eta}\, \mathbb{E}_{x\sim d_\Xi^{\pi^{(t+1)}}}\big[\log Z_t(x)\big]\\ &\leq -\frac{1}{\eta}\, \mathbb{E}_{x\sim d_\Xi^{\pi^{(t+1)}}}\big[\log Z_t(x)\big]\ \leq\ 0.\end{aligned} \tag{5.33}$$

In particular, $V^{(t)}(\Xi)$ is non-increasing in t. Moreover, since for any starting distribution μ we have the componentwise inequality $d_\mu^\pi \geq (1-\gamma)\mu$ (by the definition of discounted occupation measure), and since $\log Z_t(x) \ge 0$, it follows that (5.33) implies

$$\begin{aligned}\frac{1}{\eta}\,\mathbb{E}_{x\sim\mu}\big[\log Z_t(x)\big] &\le \frac{1}{(1-\gamma)\eta}\,\mathbb{E}_{x\sim d_\mu^{\pi^{(t+1)}}}\big[\log Z_t(x)\big]\\ &\le \frac{1}{(1-\gamma)^2}\big(V^{(t)}(\mu) - V^{(t+1)}(\mu)\big),\end{aligned}\tag{5.34}$$

where the last inequality is (5.33) applied with $\Xi = \mu$ and then rearranged.

Let π^* be an optimal policy for the objective $V^\pi(\Xi)$, and write $d^* := d_\Xi^{\pi^*}$ for its discounted state-occupation distribution. By the performance difference lemma (with $\pi = \pi^*$ and $\tilde{\pi} = \pi^{(t)}$),

$$V^{(t)}(\Xi) - V^*(\Xi) \le \frac{1}{1-\gamma}\,\mathbb{E}_{x\sim d^*}\Big[\sum_u \pi^*(u|x)\,A^{(t)}(x,u)\Big].\tag{5.35}$$

Using (5.32) inside (5.35), we obtain

$$\begin{aligned}V^{(t)}(\Xi) - V^*(\Xi) &\le -\frac{1}{\eta}\,\mathbb{E}_{x\sim d^*}\Big[\sum_u \pi^*(u|x)\log\frac{\pi^{(t+1)}(u|x)Z_t(x)}{\pi^{(t)}(u|x)}\Big]\\ &= \frac{1}{\eta}\,\mathbb{E}_{x\sim d^*}\Big[\mathrm{KL}\big(\pi^*(\cdot|x)\,\|\,\pi^{(t)}(\cdot|x)\big) \qquad (5.36)\\ &\qquad - \mathrm{KL}\big(\pi^*(\cdot|x)\,\|\,\pi^{(t+1)}(\cdot|x)\big) + \log Z_t(x)\Big].\end{aligned}\tag{5.37}$$

Average (5.37) over $t = 0, 1, \ldots, T-1$ to obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\big(V^{(t)}(\Xi) - V^*(\Xi)\big) \le \frac{1}{\eta T}\,&\mathbb{E}_{x\sim d^*}\Big[\mathrm{KL}\big(\pi^*(\cdot|x)\,\|\,\pi^{(0)}(\cdot|x)\big)\Big]\\ &+ \frac{1}{\eta T}\sum_{t=0}^{T-1}\mathbb{E}_{x\sim d^*}\big[\log Z_t(x)\big].\end{aligned}\tag{5.38}$$

Since $V^{(t)}(\Xi)$ is non-increasing in t, the left-hand side satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(V^{(t)}(\Xi) - V^*(\Xi)\big) \ge V^{(T)}(\Xi) - V^*(\Xi).$$

Hence, from (5.38),

$$V^{(T)}(\Xi) - V^*(\Xi) \le \frac{1}{\eta T}\,\mathbb{E}_{x\sim d^*}\Big[\mathrm{KL}\big(\pi^*(\cdot|x)\,\|\,\pi^{(0)}(\cdot|x)\big)\Big] + \frac{1}{\eta T}\sum_{t=0}^{T-1}\mathbb{E}_{x\sim d^*}\big[\log Z_t(x)\big].\tag{5.39}$$

With $\theta^{(0)} = 0$, the softmax policy is uniform: $\pi^{(0)}(u|x) = 1/|\mathcal{U}_0|$ for all x, u. Therefore, for every x,

$$\mathrm{KL}\big(\pi^*(\cdot|x)\,\|\,\pi^{(0)}(\cdot|x)\big) \le \log|\mathcal{U}_0|,$$

and so

$$\mathbb{E}_{x\sim d^*}\Big[\mathrm{KL}\big(\pi^*(\cdot|x)\,\|\,\pi^{(0)}(\cdot|x)\big)\Big] \le \log|\mathcal{U}_0|. \tag{5.40}$$

Apply (5.34) with $\mu = d^*$:

$$\frac{1}{\eta}\,\mathbb{E}_{x\sim d^*}\big[\log Z_t(x)\big] \le \frac{1}{(1-\gamma)^2}\big(V^{(t)}(d^*) - V^{(t+1)}(d^*)\big).$$

Summing over $t = 0, \ldots, T-1$ telescopes to

$$\frac{1}{\eta}\sum_{t=0}^{T-1}\mathbb{E}_{x\sim d^*}\big[\log Z_t(x)\big] \le \frac{1}{(1-\gamma)^2}\big(V^{(0)}(d^*) - V^{(T)}(d^*)\big) \le \frac{1}{(1-\gamma)^2}, \tag{5.41}$$

where in the last step we use the standard bounded-cost normalization $0 \le G(\cdot) \le 1$, which implies $0 \le V^{\pi}(\cdot) \le (1-\gamma)^{-1}$ for all π, hence $V^{(0)}(d^*) - V^{(T)}(d^*) \le (1-\gamma)^{-1}$.

Combining (5.39), (5.40), and (5.41) yields

$$V^{(T)}(\Xi) - V^*(\Xi) \le \frac{\log|\mathcal{U}_0|}{\eta T} + \frac{1}{(1-\gamma)^2 T}.$$

This proves the stated bound. Lastly, if $\eta \ge (1-\gamma)^2\log|\mathcal{U}_0|$, then

$$V^{(T)}(\Xi) - V^*(\Xi) \le \frac{\log|\mathcal{U}_0|}{\eta T} + \frac{1}{(1-\gamma)^2 T} \le \frac{1}{(1-\gamma)^2 T} + \frac{1}{(1-\gamma)^2 T} = \frac{2}{(1-\gamma)^2 T}.$$

Thus, to ensure $V^{(T)}(\Xi) - V^*(\Xi) \le \epsilon$, it suffices that

$$T \;\ge\; \frac{2}{(1-\gamma)^2\,\epsilon}.$$

□

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, etc...

5.3 Exercises

Exercise 5.3.1. *Prove that the SARSA iteration in Section 5.1A converges.*

Exercise 5.3.2 (Convergence of entropy-regularized Q-learning)**.** *Let $\mathcal{X}$ and $\mathcal{U}$ be finite state and action spaces, and let $\gamma \in (0,1)$ be the discount factor. Fix a regularization parameter $\lambda > 0$.*

Define the entropy-regularized Bellman operator $\mathcal{T}_\lambda$ by

$$(\mathcal{T}_\lambda Q)(x,u) = g(x,u) + \gamma\sum_{x'} p(x'|x,u)\Big(-\lambda\log\sum_{u'} e^{-Q(x',u')/\lambda}\Big),$$

which acts on functions $Q : \mathcal{X}\times\mathcal{U} \to \mathbb{R}$.

(i) *Show that the soft Bellman operator $\mathcal{T}_\lambda$ is a γ-contraction in the sup-norm:*

$$\|\mathcal{T}_\lambda Q - \mathcal{T}_\lambda \tilde{Q}\|_\infty \leq \gamma \|Q - \tilde{Q}\|_\infty .$$

Hint: *Prove that the function*

$$F(q) = -\lambda \log \sum_u e^{-q(u)/\lambda}$$

is 1-Lipschitz in $\|\cdot\|_\infty$.

(ii) *Deduce from part (i) that there exists a unique fixed point Q^λ satisfying*

$$Q^\lambda = \mathcal{T}_\lambda Q^\lambda ,$$

and that the value iteration sequence $Q_{k+1} = \mathcal{T}_\lambda Q_k$ converges to Q^λ for any initial Q_0.

(iii) *Consider the stochastic update*

$$Q_{n+1}(x,u) = Q_n(x,u) + \alpha_n \Big[G_n + \gamma \Big(-\lambda \log \sum_{u'} e^{-Q_n(X_{n+},u')/\lambda} \Big) \\ - Q_n(x,u) \Big] \mathbf{1}_{\{(X_n,U_n)=(x,u)\}} .$$

Rewrite the iteration in the form

$$Q_{n+1} = Q_n + \alpha_n \Big(\mathcal{T}_\lambda Q_n - Q_n + \varepsilon_{n+1} \Big),$$

where $\{\varepsilon_{n+1}\}$ is a martingale-difference noise sequence. Explicitly identify the noise term.

(iv) *Let $Y_n = Q_n - Q^\lambda$. Show that*

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] = Y_n + \alpha_n \big(\mathcal{T}_\lambda Q_n - \mathcal{T}_\lambda Q^\lambda - Y_n \big).$$

Using the contraction property from part (i), prove

$$\|\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n]\|_\infty \leq (1 - \alpha_n(1-\gamma)) \|Y_n\|_\infty .$$

(v) *Assume that every state–action pair is visited infinitely often, the learning rates satisfy the Robbins–Monro conditions,*

$$\sum_n \alpha_n(x,u) = \infty, \qquad \sum_n \alpha_n(x,u)^2 < \infty,$$

and the one-step costs have bounded second moment. Verify that the assumptions of Lemma 5.1.3 are satisfied and conclude that

$$Q_n \to Q^\lambda \quad \textit{almost surely.}$$

Exercise 5.3.3 (Actor–Critic with SARSA for a controlled random walk on the torus)**.** *Consider a controlled random walk on the two-dimensional discrete torus*

$$\mathbb{T}_N := \{0, 1, \dots, N-1\}^2,$$

with periodic boundary conditions (all arithmetic is modulo N). Each state is of the form $x = (i, j) \in \mathbb{T}_N$. The action space is

$$\mathcal{U} = \{\text{up}, \text{down}, \text{left}, \text{right}\}.$$

As for the dynamics, if $X_t = (i, j)$ and the action is U_t, the next state is

$$X_{t+1} = X_t + e(U_t) + \xi_t \quad mod\ N,$$

where $e(U_t)$ is the unit vector corresponding to the chosen direction, and ξ_t is a random noise taking values in $\{(0,0), (1,0), (-1,0), (0,1), (0,-1)\}$ with small probability $\varepsilon > 0$ of moving in each direction (and probability $1 - 4\varepsilon$ of staying put).

The costs are as follows. Fix a target state $x^\star \in \mathbb{T}_N$. The running cost is

$$g(X_t) := \|X_t - x^\star\|_2^2,$$

where the distance is computed using the minimal periodic distance on the torus.

The objective is to minimize the discounted cost

$$J(\pi) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t G_t\right], \qquad \gamma \in (0,1).$$

*We consider a tabular actor–critic algorithm in which the **critic** estimates $Q^\pi(x, u)$ using SARSA, and the **actor** uses a softmax policy*

$$\pi_\theta(u|x) = \frac{\exp(-\theta(x,u))}{\sum_{v \in \mathcal{U}} \exp(-\theta(x,v))}.$$

Consider the following algorithm. At time t:

1. *a) Sample $U_t \sim \pi_\theta(\cdot|X_t)$,*

   b) Sample ξ_t as described, and compute $X_{t+1} = X_t + e(U_t) + \xi_t$.

   c) Sample $U_{t+1} \sim \pi_\theta(\cdot|X_{t+1})$.

   d) Compute $G_t = g(X_t)$.

2. *Compute TD error $\delta_t = g(X_t) + \gamma Q(X_{t+1}, U_{t+1}) - Q(X_t, U_t)$.*

3. *Update the critic:*
$$Q(X_t, U_t) \leftarrow Q(X_t, U_t) + \alpha_t \delta_t.$$

4. *Update the actor:*
$$\theta(x,u) \leftarrow \theta(x,u) - \eta_t \delta_t \, \mathbf{1}_{\{x = X_t\}} \big(\mathbf{1}_{\{u = U_t\}} - \pi_\theta(u|X_t)\big).$$

Your task is to analyse the behaviour of this algorithm.

(i) *Show that the actor update corresponds to a stochastic gradient descent step for minimizing $J(\pi_\theta)$ using the policy-gradient identity with the SARSA estimate of the advantage*

$$A^\pi(X_t, U_t) \approx \delta_t.$$

(ii) *Implement the above algorithm for $N = 10$, $\gamma = 0.95$, and a small noise level $\varepsilon = 0.05$.*

(iii) *Initialize $Q \equiv 0$ and $\theta \equiv 0$. Run the algorithm for a sufficiently long horizon (e.g. 10^5 steps). Plot the average discounted cost over time, the policy $\pi_\theta(\cdot|x)$ at selected states, and the value function $V(x) = \sum_u \pi_\theta(u|x)Q(x,u)$.*

(iv) *Compare the learned policy to the intuitive optimal policy (which moves toward $x^\star$ along shortest periodic paths). Explain qualitatively how the noise level ε influences the learned policy.*

(v) *Experiment with different learning rates α_t and η_t and discuss:*

  a. *What happens if the actor learns much faster than the critic?*

  b. *What happens if the critic learns much faster than the actor?*

(vi) *Modify the critic to use TD(λ) with eligibility traces and compare learning speed and stability.*

Exercise 5.3.4 (Policy gradient in the multi-armed bandit problem)**.** *Consider the special case of a reinforcement learning problem where there is a* single *state $x \in \mathcal{X}$ (so $X_t \equiv x$ for all t) and a finite action set $\mathcal{U}_0 = \{1, \dots, m\}$. Given that the state is always the same, there is no added value in modulating the strategies by x. The problem is that we do not know, initially, the payoff value from each possible action.*

This is effectively the "multi-armed bandit problem"[2]. The (discounted) cost is

$$J(\pi_\theta) := \mathbb{E}^{\pi_\theta}\Big[\sum_{t=0}^{\infty} e^{-\rho t}\, g(U_t)\Big], \qquad \rho > 0,$$

where $g(u) \in \mathbb{R}$ is deterministic and, under π_θ, the actions $\{U_t\}_{t\geq 0}$ are independent with law $\pi_\theta(\cdot)$. Assume a tabular softmax parameterisation

$$\pi_\theta(u) = \frac{e^{\theta_u}}{\sum_{k=1}^m e^{\theta_k}}, \qquad \theta_1, \dots, \theta_m \in \mathbb{R}.$$

[2] In the setting considered here, it turns out that there is an almost-closed-form Bayesian optimal strategy for this problem, given by Gittins [22]. For an extensive study of related problems, see also the textbook by Lattimore and Szepesvári [35].

(i) *Show that $J(\pi_\theta)$ can be written in the form*

$$J(\pi_\theta) = \frac{1}{1-e^{-\rho}} \sum_{u=1}^{m} \pi_\theta(u)\, g(u).$$

(ii) *Prove that for each $v \in \{1, \dots, m\}$,*

$$\partial_{\theta_v} J(\pi_\theta) = \frac{1}{1-e^{-\rho}}\, \pi_\theta(v) \Big(g(v) - \sum_{u=1}^{m} \pi_\theta(u) g(u) \Big).$$

(In particular, the gradient is proportional to a baseline-subtracted *cost.)*

(iii) *Show that the same gradient can be written as*

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-e^{-\rho}}\, \mathbb{E}^{U_0 \sim \pi_\theta} \Big[g(U_0)\, \nabla_\theta \log \pi_\theta(U_0) \Big],$$

and explain why subtracting a constant baseline $b \in \mathbb{R}$ from $g(U_0)$ inside the expectation does not change the value of the gradient.

Exercise 5.3.5 (Q-learning as a stochastic gradient method)**.** *Consider a finite Markov decision process with discount factor $\gamma = e^{-\rho} \in (0, 1)$ and running costs $g(x, u)$. The objective is to* minimize *the discounted cost.*

Let $Q_\theta(x, u)$ be a parameterized action–value function with parameter vector $\theta \in \mathbb{R}^m$. Define the Bellman residual at a sample transition (X_t, U_t, G_t, X_{t+1}) by

$$\delta_t(\theta) = G_t + \gamma \min_{u'} Q_\theta(X_{t+1}, u') - Q_\theta(X_t, U_t).$$

Consider an algorithm which seeks to minimize the loss function

$$\mathcal{L}(\theta) = \frac{1}{2} \left(\delta_t(\theta) \right)^2.$$

(i) *Show that the gradient of the loss is*

$$\nabla_\theta \mathcal{L}(\theta) = \delta_t(\theta)\, \nabla_\theta \delta_t(\theta).$$

(ii) *Compute $\nabla_\theta \delta_t(\theta)$. Show that, ignoring the derivative of the* min *operator (i.e. treating the greedy action at X_{t+1} as fixed), we obtain*

$$\nabla_\theta \delta_t(\theta) = -\nabla_\theta Q_\theta(X_t, U_t).$$

(iii) *Deduce that a single gradient descent step $\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t)$ can be written as the iteration*

$$\theta_{t+1} = \theta_t + \eta\, \delta_t(\theta_t)\, \nabla_\theta Q_\theta(X_t, U_t).$$

(iv) *Assume now the tabular parameterization*

$$Q_\theta(x,u) = \theta(x,u).$$

Show that

$$\nabla_{\theta_{x,u}} Q_\theta(X_t,U_t) = \mathbf{1}_{\{(x,u)=(X_t,U_t)\}}.$$

Hence conclude that, with $Q_{\theta_t} =: Q_t$ the stochastic gradient step reduces to the Q-learning update

$$Q_{t+1}(x,u) = Q_t(x,u)+\eta\Big[G_t+\gamma\min_{u'} Q(X_{t+1},u')-Q(x,u)\Big]\mathbf{1}_{\{(x,u)=(X_t,U_t)\}}.$$

Exercise 5.3.6 (Comparison of Q-learning, TD(1) and TD($\frac{1}{2}$))**.** *Consider a finite Markov decision process with state space*

$$\mathcal{X} = \{1,2,3\}, \qquad \mathcal{U} = \{a,b\},$$

and discount factor $\gamma = 0.9$. The transitions and costs are deterministic and given as follows. From state 1:

$$p(2;1,a) = p(3;1,b) = 1, \qquad g(1,a) = 0, \qquad g(1,b) = 1.$$

From state 2:

$$p(1;2,a) = p(3;2,b) = 1, \qquad g(2,a) = 0.5, \qquad g(2,b) = 0.$$

State 3 is absorbing ($p(3;3,\cdot) = 1$) and $g(3,u) = 0$ for both actions. We observe the following short trajectory (generated by some behaviour policy):

$$(X_0 = 1,\ U_0 = a,\ G_0 = 0),\ (X_1 = 2,\ U_1 = b,\ G_1 = 0),\ X_2 = 3.$$

Assume a tabular parameterization with initial estimate

$$Q_0(x,u) = 0.5 \quad \textit{for all } (x,u),$$

and constant learning rate $\alpha = 0.1$.

(i) Q-learning: *Write the classical Q-learning update*

$$Q_{n+1}(x,u) = Q_n(x,u)+\alpha\Big[G_n+\gamma\min_{u'} Q_n(X_{n+},u')-Q_n(x,u)\Big]\mathbf{1}_{\{(X_n,U_n)=(x,u)\}}.$$

Compute explicitly the updates at times $n = 0$ and $n = 1$, and determine the resulting values of

$$Q(1,a) \quad \textit{and} \quad Q(2,b)$$

after processing the trajectory.

(ii) TD(1) (Monte–Carlo):*For each visited state–action pair, compute the full discounted return*

$$R_n = G_n + \gamma G_{n+1} + \gamma^2 G_{n+2} + \cdots ,$$

and perform the update

$$Q(x,u) \leftarrow Q(x,u) + \alpha\big(R_n - Q(x,u)\big)$$

at the corresponding time step.

Compute the updated values of $Q(1,a)$ and $Q(2,b)$.

(iii) TD($\frac{1}{2}$) with eligibility traces: *Let $\lambda = \frac{1}{2}$ and define the eligibility traces*

$$e_n(x,u) = \gamma\lambda\, e_{n-1}(x,u) + \mathbf{1}_{\{(X_n,U_n)=(x,u)\}}, \qquad e_{-1} \equiv 0,$$

and the one-step TD error

$$\delta_n = G_n + \gamma Q_n(X_{n+}, U_{n+}) - Q_n(X_n, U_n).$$

The update is

$$Q_{n+1}(x,u) = Q_n(x,u) + \alpha\, \delta_n\, e_n(x,u).$$

Compute explicitly the traces e_0, e_1 and the corresponding updates to $Q(1,a)$ and $Q(2,b)$.

(iv) Compare the numerical updates obtained in parts (i)–(iii). Which method produces the largest and smallest changes in $Q(1,a)$?

(v) Repeat the calculations assuming instead that

$$Q_0(1,a) = 2, \qquad Q_0(x,u) = 0.5 \text{ otherwise.}$$

Discuss qualitatively how optimistic initialization interacts differently with Q-learning and TD(λ) *in this example.*

Chapter 6

Continuous-time 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 Chapter 7.

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.

6.1 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:

$$\dot{X}_t^U := \frac{\mathrm{d}X_t^U}{\mathrm{d}t} = f(t, X_t^U, U_t), \tag{6.1}$$

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

[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.

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 $X_{t+\epsilon}$, 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 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 6.1.1. *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 $U_t = -\mathrm{sign}(X_t)$ 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 (6.1) 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:

$$X_{t'}^{t,x,U} = x + \int_t^{t'} f(s, X_s^{t,x,U}, U_s)\mathrm{d}s. \tag{6.2}$$

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

$$J(U) = \int_0^T e^{-\rho s} g(s, X_s^U, U_s)\mathrm{d}s + e^{-\rho T}\Phi(X_T^U) \tag{6.3}$$

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 6.1.2. *For our deterministic problem, we say a control $U : \mathbb{T} \to \mathcal{U}$ is open-loop*admissible*, and write $U \in \mathbb{U}$, if U is measurable and* (6.2) *admits a unique solution $X^{t,x,U}$, taking values in $\mathcal{X}$, for all $(t, x) \in \mathbb{T} \times \mathcal{X}$.*

We will similarly say that a feedback control $\mathbf{u} : \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ is closed-loop admissible, if $\mathbf{u}$ is Borel measurable (in both arguments) and the natural *extension of* (6.2)*, that is,*

$$X_{t'} = x + \int_t^{t'} f\big(s, X_s, \mathbf{u}(s, X_s)\big)\mathrm{d}s,$$

admits a unique solution $X_s^{t,x,\mathbf{u}}$ *for all* (t, x)*.*

We observe that these definitions are consistent, as if $\mathbf{u}$ *is an admissible closed-loop control, then* $U_s = \mathbf{u}(s, X_s^{t,x,\mathbf{u}})$ *defines an admissible open-loop control. For brevity, we will simply say 'admissible' for open-loop admissible controls, unless this leads to confusion.*

Remark 6.1.3. 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 $J(t, x, U) \geq \int_t^T g_*(s)\mathrm{d}s + g_*(T) > -\infty$.

Lemma 6.1.4. *With the above definitions, we have the following results:*

(i) the set of admissible controls is closed under pasting, that is, given two admissible controls U, U' *the control defined by* $U''_s = U_s\mathbf{1}_{s\leq t} + U'_s\mathbf{1}_{s>t}$ *is also admissible;*

(ii) for any admissible control, X *and* J *satisfy the flow properties, for* $t \leq t' \leq t''$*,*

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

$$J(t, x, U) = \int_t^{t'} e^{-\rho(s-t)} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + e^{-\rho(t'-t)} J(t', X_{t'}^{t,x,U}, U);$$

(iii) the cost-to-go does not depend on past actions, that is

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

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. □

6.2 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 6.2.1 (Dynamic programming)**.** *The value function*

$$v(t,x) := \inf_{U\in\mathbb{U}} J(t,x,U)$$

satisfies the dynamic programming equation

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

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

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

Fix $\epsilon > 0$, and write

$$\tilde{v}(t,x) = \inf_{U\in\mathbb{U}} \big\{ \int_t^{t'} e^{-\rho(s-t)} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + e^{-\rho(t'-t)} v(t', X_{t'}^{t,x,U})\big\}.$$

Then there exists $U \in \mathbb{U}$ such that

$$v(t,x) \leq J(t,x,U) \leq v(t,x) + \epsilon.$$

Therefore

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

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

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

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

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

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

and by pasting, we build a control $\tilde{U} = \mathbf{1}_{s\leq t'}U + \mathbf{1}_{s>t}U'$ 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

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

This implies that

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

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

Remark 6.2.2. 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

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

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

$$0 = \partial_t v(t,x) - \rho v(t,x) + \inf_{u\in\mathcal{U}} \Big\{ g(s,x,u) + \big\langle \nabla v(t,x), f(t,x,u)\big\rangle \Big\}. \tag{6.4}$$

(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 $|v| \le \|g\|_\infty/\rho$ (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 6.2.3. There are slightly different conventions regarding what is called the 'Hamiltonian'. In some literature, the Hamiltonian is defined to be

$$\tilde{H}(t, x, p, u) = g(t, x, u) + \big\langle p, f(t, x, u)\big\rangle \tag{6.5}$$

whereas here we will usually say that the Hamiltonian is the optimized version

$$H(t, x, p) = \inf_{u \in \mathcal{U}} \tilde{H}(t, x, p, u)$$

and refer to $\tilde{H}$ as the unoptimized Hamiltonian when we need clarity.

Remark 6.2.4. It's helpful to point out that the (unoptimized) Hamiltonian $\tilde{H}(\cdot, \cdot, \nabla v, u)$ is, in some sense, playing a role similar to what the advantage function $A(t, x, u) = Q(t, x, u) - v(t, x)$, where Q is the time-dependent Q function, played in discrete time. To see this, compare the Bellman iteration (for discrete deterministic systems, with $\rho = 0$) written in the form

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

with the Hamilton–Jacobi equation

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

The same connection also holds true in stochastic problems.

[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.

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 6.2.5 (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), we have that

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

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

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

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

(iii) g satisfies the growth bound, for every u,

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

It then follows that

(i) w is the unique solution to the Hamilton–Jacobi equation satisfying these continuity and growth bounds;

(ii) w is the value function of the control problem;

(iii) a control U is optimal if and only if

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

[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.

Proof. Recall 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

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

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

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

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

$$\big|\tilde{H}\big(t, X_t, \nabla w(X_t), u\big) - \tilde{H}\big(s, X_s, \nabla w(X_s), u\big)\big| \leq c|t-s|. \tag{6.6}$$

Taking $\epsilon > 0$ and u^ϵ such that

$$\tilde{H}\big(s, X_s, \nabla w(s, X_s), u^\epsilon\big) \leq H\big(s, X_s, \nabla w(s, X_s)\big) + \epsilon$$

we have

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

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

$$\big|H\big(t, X_t, \nabla w(t, X_t)\big) - H\big(s, X_s, \nabla w(s, X_s)\big)\big| \leq c|t-s|. \tag{6.7}$$

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 $\epsilon > 0$, $x_0 \in \mathcal{X}$ and $\delta = \epsilon/c$, where c is as in (6.6). Let $U^\epsilon_{t_0}$ be such that

$$\tilde{H}(t_0, x_0, \nabla w(t_0, x_0), U^\epsilon_0) \leq H(t_0, x_0, \nabla w(t_0, x_0)) + \epsilon.$$

[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.

Using this $U^\epsilon_{t_0}$, for $t \leq t_0 + \delta$, define $U^\epsilon_t = U^\epsilon_0$ and the ODE solution

$$X^\epsilon_t = x_0 + \int_{t_0}^t f(s, X^\epsilon_s, U^\epsilon_s)\mathrm{d}s.$$

As $\delta = \epsilon/c$, from (6.6) and (6.7) we know that

$$\begin{aligned}\tilde{H}(t, X^\epsilon_t, \nabla w(t, X^\epsilon_t), U^\epsilon_t) &\leq \tilde{H}(t_0, x_0, \nabla w(t_0, x_0), U^\epsilon_0) + \epsilon \\ &\leq H(t_0, x_0, \nabla w(t_0, x_0)) + 2\epsilon \\ &\leq H(t, X^\epsilon_t, \nabla w(t, X^\epsilon_t)) + 3\epsilon.\end{aligned} \tag{6.8}$$

We then repeat this construction started at $(t_0+\delta, X^\epsilon_\delta)$ instead of (t_0, x_0), which defines $U^\epsilon_t, X^\epsilon_t$ for $t \in (\delta, 2\delta]$. Iterating, we define U^ϵ, X^ϵ for all t, and know that $\frac{\mathrm{d}}{\mathrm{d}t}X^\epsilon = f(t, X^\epsilon_t, U^\epsilon_t)$ and $\tilde{H}(t, X^\epsilon_t, \nabla w(t, X^\epsilon_t), U^\epsilon_t) \leq H(t, X^\epsilon_t, \nabla w(t, X^\epsilon_t)) + 3\epsilon$ 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, $U^\epsilon \in \mathbb{U}$.

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

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

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

$$\begin{aligned}
&e^{-\rho t}w(t,X_t^U) = \tilde{w}(t)\\
&= \int_t^T e^{-\rho s}\Big[-\Big\langle \nabla w, f(s,X_s^U,U_s)\Big\rangle + H(s,X_s^U,\nabla w(s,X_s^U))\Big]\mathrm{d}s + e^{-\rho T}\Phi(X_T^U)\\
&= \int_t^T e^{-\rho s} g(s,X_s^U,U_s)\mathrm{d}s + e^{-\rho T}\Phi(X_T^U)\\
&\quad + \int_t^T e^{-\rho s}\Big[\underbrace{H(s,X_s^U,\nabla w(s,X_s^U)) - \tilde{H}(s,X_s^U,\nabla w(s,X_s^U),U)}_{\leq 0}\Big]\mathrm{d}s.
\end{aligned}$$

In particular, by rearranging, we see that

$$w(t,X_t^U) \leq J(t,X_t^U,U).$$

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

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

Taking $\epsilon \to 0$, 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 (satsifying the stated growth bounds) must be a value function for our control problem, hence the PDE must admit at most one solution $w = v$. □

Remark 6.2.6. As a corollary to this proof, we see that the infimal cost is approached by a sequence of piecewise constant controls.

[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

$$\begin{aligned}
w(t,X^U) = &\int_t^\infty e^{-\rho(s-t)} g(s,X_s^U,U_s)\mathrm{d}s\\
&+ \int_t^\infty e^{-\rho(s-t)}(H(s,X_s^U,\nabla w(s,X_s^U)) - \tilde{H}(s,X_s^U,\nabla w(s,X_s^U),U))\mathrm{d}s.
\end{aligned}$$

As (6.6) and (6.7) 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.

Corollary 6.2.7. *Suppose the assumptions of Theorem 6.2.5 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*

$$\mathbf{u}^*(t,x) \in \arg\min_{u\in\mathcal{U}} \tilde{H}\big(t,x,v(t,x),\nabla v(t,x),u\big).$$

It follows that $\frac{\mathrm{d}}{\mathrm{d}s}X_s^{t,x} = f(s, X_s^{t,x}, \mathbf{u}^*(s, X_s^{t,x}))$ admits a unique solution, and hence $U_s^* = \mathbf{u}^*(s, X_s^{0,x_0})$ is (open loop) admissible (so $\mathbf{u}^*$ is closed-loop admissible), and therefore is an optimal control.

Remark 6.2.8. 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 6.2.9. *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 6.2.10. *In the setting of Example 6.1.1, (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*

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

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 $|x| \leq T-t$, 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 6.2.11. *Consider a linear-quadratic problem with state $X \in \mathcal{X} = \mathbb{R}^n$ and control space $\mathcal{U} \subset \mathbb{R}^m$. We suppose X follows the the linear dynamics*

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

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 Hamiltonian is*

$$\begin{aligned}&H(t,x,p)\\ &= \min_u \Big\{x^\top Qx + u^\top Ru + 2x^\top Su + 2Wx + 2Yu + Z + p^\top(Ax + Bu + C)\Big\}.\end{aligned}$$

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 (with $\bar\Sigma_t = \frac{1}{2}(\Sigma_t + \Sigma_t^\top)$ *the symmetric part of* Σ_t*)*

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

Taking a first-order condition to find the optimizer in the Hamiltonian $H(t, x, \nabla v)$*, we find (assuming* $\mathcal{U}$ *is sufficiently large), we have*

$$0 = 2u^\top R + 2x^\top S + 2Y + 2(x^\top \bar\Sigma_t + \Psi_t)B$$

and hence the optimal strategy is of feedback form

$$\mathbf{u}^*(t,x) = -R^{-1}\Big(S^\top x + Y^\top + B^\top(\bar\Sigma_t x + \Psi_t^\top)\Big) =: K_t x + H_t.$$

Substituting into the Hamilton–Jacobi equation, we have

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

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

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

with terminal values Σ_T, Ψ_T *and* Γ_T *specified. This can be simplified further if desired. The optimal state dynamics can also be given in closed form,*

$$\frac{\mathrm{d}X_t^*}{\mathrm{d}t} = f\big(t, X_t^*, \mathbf{u}^*(t, X_t^*)\big) = (A + BK_t)X_t^* + (BH_t + C)$$

which is a linear ODE system (given the solutions $\Sigma_t, \Psi_t, \Gamma_t$*).*

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 also 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).*

6.3 Pontryagin's principle

Just as we saw in discrete time, 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, or through a change of sign where we maximize a Hamiltonian), 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.

A naïve approach to obtaining a first order condition, assuming the Hamilton–Jacobi equation admits a sufficiently smooth solution, would be to say that, with the notation $\tilde{H}$ as in (6.5) , our optimal control U^* should satisfy

$$\partial_u \tilde{H}\big(t, X_t^*, \nabla v(t, X_t^*), U_t^*\big) = 0. \tag{6.9}$$

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 $\nabla v(t, X_t^*)$ which we can use directly. Particularly when the dimension of X is high, it may be much more efficient to try and solve the resulting forward-backward system of ODEs, rather than computing the solution to the PDE.

Theorem 6.3.1 (Pontryagin's minimum principle)**.** *Consider a deterministic control problem as described in Section 6.1 and assume that $\mathcal{U}$ is an open subset of $\mathbb{R}^m$. Let $X_0 = x_0$ be the initial value of the controlled state process and $\mathbf{u}^* : \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ be a feedback control, with which we define the controlled process X^*. Suppose in addition that $\mathbf{u}^*$ is differentiable at (t, X_t^*) for all $t \in \mathbb{T}$, and that there is a $C^{1,2}$ solution to the Hamilton–Jacobi equation.*

If the control $U_t^* = \mathbf{u}^*(t, X_t^*)$ is optimal, then it yields a fixed point of the *system of equations*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}X_t^* &= f(t, X_t^*, U_t^*), \qquad \textit{with initial value } X_0^* = x_0;\\
\frac{\mathrm{d}}{\mathrm{d}t}p_t &= \rho p_t - \nabla\tilde{\mathcal{H}}\big(t, X_t^*, p_t, U_t^*\big)\\
&= \rho p_t - \nabla g(t, X_t^*, U_t^*) - \Big(D_x f(t, X_t^*, U_t^*)\Big)^\top p_t, \qquad (6.10)\\
&\quad \textit{with terminal value } p_T = \nabla\Phi(X_T^*);\\
U_t^* &\in \arg\min_{u\in\mathcal{U}}\Big\{g(t, X_t^*, u) + \Big\langle p_t, f(t, X_t^*, u)\Big\rangle\Big\}.
\end{aligned}$$

Proof. The first equation is the dynamics of X_t^*, and the third states the fact that the optimal control must minimize the Hamiltonian. We therefore focus on the middle equation. We define the *adjoint* or *costate* process $p : \mathbb{T} \to \mathbb{R}^n$ by

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

We recall, for notational clarity, that we think of ∇v as a column vector. Therefore, as $v : \mathbb{T} \times \mathbb{R}^n \to \mathbb{R}^1$, we can equivalently write this as $\nabla v = (D_x v)^\top$, as the Jacobian $D_x v : \mathbb{T} \times \mathbb{R}^n \to \mathbb{R}^{1\times n}$, should be thought of as a row vector. Differentiating in time, the multivariate chain rule for Jacobians shows that p satisfies

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}p_t^\top &= \partial_t\big(D_x v(t, X_t^*)\big) + \big(D^2_{xx}v(t, X_t^*)\big)\big(\frac{\mathrm{d}}{\mathrm{d}t}X_t^*\big)\\
&= D_x\big(\partial_t v(t, X_t^*)\big) + \big(D^2_{xx}v(t, X_t^*)\big)\big(f(t, X_t^*, U_t^*)\big).
\end{aligned} \qquad (6.11)$$

As v satisfies the Hamilton–Jacobi equation, and $\mathbf{u}^*(t, x)$ is optimal for all t, x, we know

$$-\partial_t v(t, x) = -\rho v(t, x) + g(t, x, \mathbf{u}^*(t, x)) + \Big(D_x v(t, x)\Big)\Big(f\big(t, x, \mathbf{u}^*(t, x)\big)\Big).$$

At any point where $\mathbf{u}^*$ is differentiable, in particular at (t, X_t^*), differentiating the Hamilton–Jacobi equation with respect to x we obtain,

$$\begin{aligned}
-D_x\big(\partial_t v\big) &= -\rho(D_x v) + (D_x g) + (D_u g)(D_x\mathbf{u}^*)\\
&\quad + (f)^\top\big(D^2_{xx}v\big) + \big(D_x v\big)\Big((D_x f) + \big(D_u f\big)\big(D_x\mathbf{u}^*\big)\Big)
\end{aligned} \qquad (6.12)$$

We pull out the terms depending on $D_x\mathbf{u}^*$, which simplify as

$$\begin{aligned}
&(D_u g)(D_x\mathbf{u}^*) + \big(D_x v\big)\big(D_u f\big)\big(D_x\mathbf{u}^*\big)\\
&= \Big((D_u g) + \big(D_x v\big)\big(D_u f\big)\Big)\big(D_x\mathbf{u}^*\big)\\
&= D_u\Big(g + \big(D_x v\big)\big(f\big)\Big)\big(D_x\mathbf{u}^*\big)
\end{aligned}$$

However, as we know that $\mathbf{u}^*(t, X_t^*)$ is an optimal control, the first-order optimality condition (6.9) on the Hamiltonian ensures that, at (t, X_t^*, U_t^*),

$$D_u\Big(g + (D_x v)f\Big)(t, X_t^*) = D_u\tilde{H}\Big(t, X_t^*, (D_x v(t, X_t^*))^\top, U_t^*\Big) = 0.$$

Therefore, we can omit these terms from (6.12), yielding

$$-D_x\big(\partial_t v\big) = -\rho(D_x v) + D_x g + f^\top\big(D^2_{xx}v\big) + \big(D_x v\big)^\top(D_x f)$$

We now substitute this formula into (6.13), to obtain the simplification (all evaluated at (t, X_t^*, U_t^*))

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}p_t^\top &= \rho(D_x v) - D_x g - \big(D_x v\big)(D_x f) \\ &= \rho p_t^\top - D_x g - p_t^\top(D_x f)\end{aligned} \tag{6.13}$$

Taking a transpose gives the desired dynamics. Clearly $p(T) = \nabla v(T, X_T^*) = \nabla\Phi(X_T^*)$, and we observe that this gives p_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^*. □

Remark 6.3.2. Numerous variations of this result exist, for different types of boundary conditions. In many cases, this involves computing an additional adjoint process, given by $\mu_t = \big(\partial_t H\big)(t, X_t^*)$, 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 (for example, if our problem stops whenever X hits a boundary).

Another variant is where (part of) the terminal state is required to be a fixed value – this can be considered as a limiting case of the version considered here, with a penalty $K(x - \bar{x})^2$, which leads to an additional requirement on the candidate X_T^*, and the terminal condition $p_T = 0$.

If you look in the literature you will often see these variants presented as Pontryagin's maximum principle.

Remark 6.3.3. It is slightly odd that our derivation depends on having a $C^{1,2}$ solution to the Hamilton–Jacobi PDE, and a differentiable optimal control, but we do not have the second spatial derivative 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.

Example 6.3.4. *Let* $\mathbb{T} = [0, 3]$ *and consider the problem of minimizing the cost*

$$\int_0^3 (U_t^4 - X_t^2)\mathrm{d}t + X_3^2,$$

where X *is a scalar process following the controlled dynamics*

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t = U_t - X_t + \sin(t)$$

and $X_0 = 1$, where $U \in \mathcal{U} = \mathbb{R}$. Observe that this gives a convex Hamiltonian, so we believe that the optimal strategy should be unique, and should be determined by Pontryagin's minimum principle. Using Theorem 6.3.1 the system becomes

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} X_t^* &= U_t^* - X_t^* + \sin(t), & X_0^* &= 1; \\ \frac{\mathrm{d}}{\mathrm{d}t} p_t &= 2\,X_t^* + p_t, & p_3 &= 2\,X_3^*; \\ U_t^* &\in \operatorname*{arg\,min}_{u\in\mathcal{U}} \Big\{ u^4 - \big(X_t^*\big)^2 + p_t\,(u - X_t^* + \sin(t) \Big\}. \end{aligned} \tag{6.14}$$

To solve this, we use a variant of the 'shooting method', where we approximate the dynamics and use numerical optimization to identify the solution:

Observe that, if we knew the value of $p_0 = \nabla v(x_0)$, we could implement a numerical solution to the system that is forward in time[6]*. More precisely, we could discretise $[0, 3]$ in N equally spaced timesteps of size $\Delta = T/N$, and let $0 = t_0 < t_1 < \cdots < t_N = 3$ and with the values of $p_{t_0} = p_0$ we compute*

$$U_{t_0}^* \in \operatorname*{arg\,min}_{u\in\mathcal{U}} \Big\{ u^4 - \big(X_{t_0}\big)^2 + p_{t_0}\,(u - X_{t_0} + \sin(t_0) \Big\}.$$

Then, using p_{t_0}, $X_{t_0}^*$, and $U_{t_0}^*$, we compute p_{t_1}, $X_{t_1}^*$ using an Euler discretisation of (6.14), that is

$$\begin{aligned} \frac{X_{t_1}^* - X_{t_0}^*}{t_1 - t_0} &\approx U_{t_0}^* - X_{t_0}^* + \sin(t_0), \quad &&\Rightarrow X_{t_1}^* = X_{t_0}^* + \Delta\left(U_{t_0}^* - X_{t_0}^* + \sin(t_0)\right) \\ \frac{p_{t_1} - p_{t_0}}{t_1 - t_0} &\approx 2\,X_{t_0}^* + p_{t_0} &&\Rightarrow p_{t_1} = p_{t_0} + \Delta\left(2\,X_{t_0}^* + p_{t_0}\right). \end{aligned}$$

Using the values of p_{t_1} and $X_{t_1}^*$ obtained from the Euler discretisation, we compute $U_{t_1}^*$ and iterate forward until time $t_N = 3$. Then, we could compare the final value p_3 with $2\,X_3^*$ and look for a starting point p_0 such that $p_3 = 2\,X_3^*$.

Figure 6.1 shows the discrepancy between p_3 and $2\,X_3^*$ when computing a forward-in-time Euler discretization of the equations in Pontryagin's minimum principle, assuming a known value of p_0 (x-axis).

[SC]$_5$:fix notation from q to p in Fig6.1

A simple bisection algorithm shows that for the problem above the desired value of p_0 that makes $p_3 = 2\,X_3^*$ is $p_0 \approx -2.11$.

Finally, in Figure 6.2 we compare the performance of the optimal strategy U_t^* (computed from the Pontryagin equations using $p_0 = -2.11$) against constant strategies, that is, $U_t = c$ for all $t \in [0, 3]$ and $c \in [0, 1]$. The left panel shows the trajectory of X_t under each of the considered policies (we include X_t^* in red). The right panel shows how the cost $\int_0^3 (U_t^4 - X_t^2)\mathrm{d}t + X_3^2$ accrues over time (we include that from the optimal strategy in red). More precisely, for $t \in [0, 3]$, we show $\int_0^t (U_s^4 - X_s^2)\mathrm{d}t + X_3^2\,\mathbf{1}_{\{t=3\}}$ when $U_t = c$ for $c \in [0, 1]$.

[6] An alternative approach is to start with a guess of a good policy U_t, use this to calculate X_T^U and then solve for p_0 backwards. This then yields a guess of the value of p_0, which can be used as the basis of the next iteration.

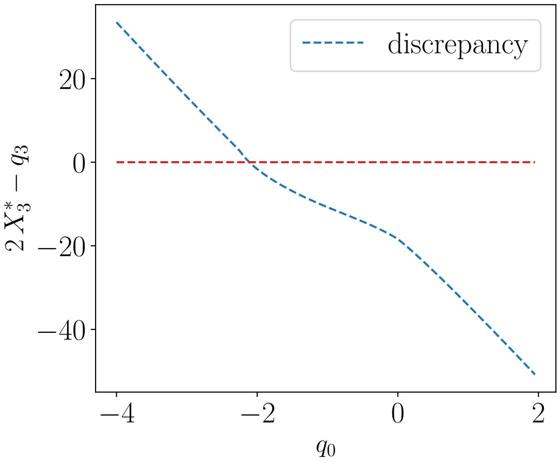


Figure 6.1: Discrepancy between p_3 and $2\,X_3^*$ for a range of starting values for p_0 (x-axis).

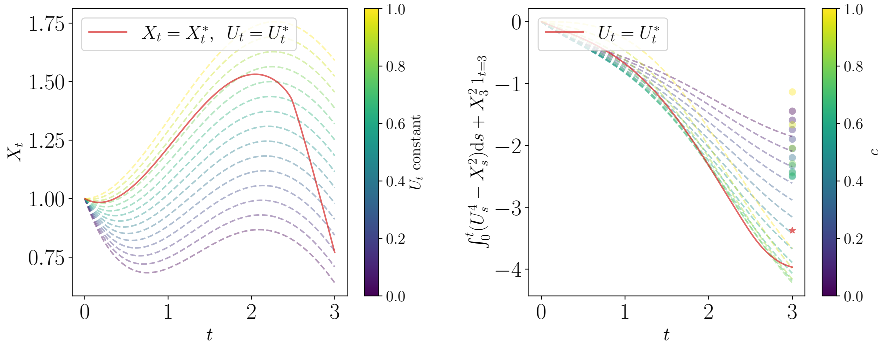


Figure 6.2: Left panel: trajectory of X_t when $U_t = c$ for a range of values of $c \in [0, 3]$. Right panel: accrued cost over time of each of the strategies considered. In both panels, the trajectory produced by the optimal strategy U_t^* is in red.

Interestingly, when t approaches the end of the time horizon, the accrued cost by some of the strategies goes below the red line (which corresponds to the optimal strategy). Of course, that happened at the expense of a higher terminal cost (see markers at $t = 3$). Figure 6.3 shows the total cost comparing the constant strategies to the optimal. Here, the superiority of the optimal strategy is clear.

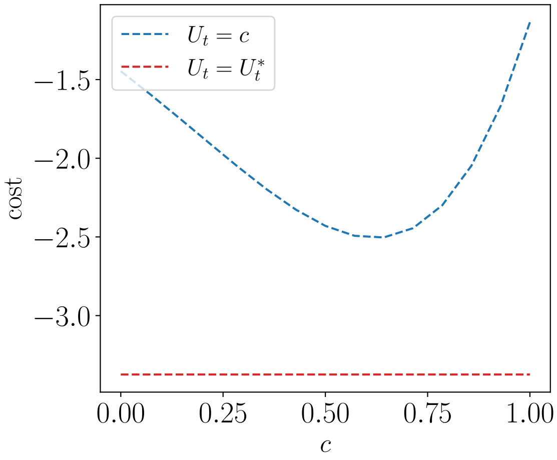


Figure 6.3: Comparison of the cost of following each of the constant strategies we consider against that of the optimal strategy U_t^*.

Remark 6.3.5. Pontryagin's principle 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 $X_t + \epsilon\eta$ 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 $[T-\epsilon, T]$), we see that $\lambda_T = \partial_x \Phi(X_T)$. Therefore, λ satisfies the same differential equation as p 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 (6.5),

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

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

Remark 6.3.6. One of the most useful interpretations of Pontryagin's minimum principle is that the costate variable represents a *shadow price*: the marginal value of relaxing a dynamic constraint. The next example develops that idea in a simple resource allocation problem.

Example 6.3.7. *Consider a firm which owns a finite stock of a non-renewable resource $X(t) \geq 0$ (for example, oil in a well). The firm chooses an extraction (or harvest) rate $U(t) \geq 0$. The stock dynamics are*

$$\mathrm{d}X(t) = -U(t)\mathrm{d}t, \qquad X(0) = x_0 > 0,$$

and the planner wishes to maximize a quantity

$$-J(U) = \int_0^T e^{-\rho t}\, \pi\big(U(t)\big)\, \mathrm{d}t, \qquad \rho > 0,$$

where the instantaneous profit function $\pi : \mathbb{R}_+ \to \mathbb{R}$ satisfies

$$\pi'(u) > 0, \qquad \pi''(u) < 0 \qquad \text{for } u > 0.$$

Note that here we are maximizing $-J(U)$, which of course is no different to minimizing $J(U)$, and simply changes the minimization to a maximization in the definition of the Hamiltonian and in Pontryagin's principle.

We now use Pontryagin's principle to characterise interior extrema. We write the (unoptimized) Hamiltonian

$$\tilde{H}\big(x, u, p\big) = \pi(u) + p(-u),$$

where p is the adjoint (costate) process. The Pontryagin necessary conditions are:

(i) state equation

$$\frac{\mathrm{d}}{\mathrm{d}t} X^*(t) = -U^*(t), \qquad X(0) = x_0;$$

(ii) costate equation

$$\frac{\mathrm{d}}{\mathrm{d}t} p(t) = \rho p(t) - \partial_x \tilde{H}(t, X^*(t), U^*(t), p(t)) = \rho p(t),$$

since $\tilde{H}$ has no explicit x-dependence;

(iii) $U^*(t)$ *is a maximizer of* $\tilde{H}\big(X^*(t), \cdot, p(t)\big)$*; assuming we are looking for an interior optimizer, this simplifies to*

$$0 = \partial_u \tilde{H} = \pi'(U^*(t)) - p(t),$$

or equivalently

$$\pi'\big(U^*(t)\big) = p(t).$$

From this setup, we can give an economic interpretation of the costate variable: The equality $\pi'(u) = p$ *equates the immediate marginal benefit of extracting one extra unit now,* $\pi'(u)$*, with the (implicit) marginal value of keeping one extra unit of resource in the ground,* p*. Thus the costate* $p(t)$ *is the* shadow price *of the resource: it measures the marginal value to the controller of leaving an additional unit in the ground. The first-order condition therefore reflects the usual economic rule*

marginal benefit of extracting now
= marginal cost (shadow price) of depleting stock.

We can do more with this relationship: solving the linear costate ODE gives $p(t) = p(0)e^{\rho t}$*. That is, the shadow price grows at the discount rate* ρ*. Intuitively, the opportunity cost of extracting today increases at the discount rate* ρ*. By substituting in the first order condition, we derive* Hotelling's rule*, which says that the price of a non-renewable resource should grow at the discount rate*

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\pi'\big(u(t)\big)\Big) = \frac{\mathrm{d}p}{\mathrm{d}t}(t) = \rho\, p(t) = \rho\, \pi'\big(u(t)\big).$$

Applying the chain rule, we obtain the following ODE for the extraction rate

$$\pi''(u(t))\,\frac{\mathrm{d}u}{\mathrm{d}t}(t) = \rho\,\pi'(u(t)) \quad \textit{or equivalently} \quad \frac{\mathrm{d}u}{\mathrm{d}t}(t) = \frac{\rho\,\pi'(u(t))}{\pi''(u(t))}.$$

Because the numerator is positive and $\pi'' < 0$*, we find that* $\frac{\mathrm{d}u}{\mathrm{d}t} < 0$*, that is, typically the optimal extraction rate decreases over time as the shadow price of remaining stock increases.*

6.4 Exercises

Exercise 6.4.1. *Write down a deterministic control problem where there are exactly two (distinct) trajectories for the optimally controlled state process* X*.*

Exercise 6.4.2 (Ramsey–Cass–Koopmans problem of optimal saving)**.** *The following exercise is based on [40, 15, 32]. Consider an economy in which the labour force* L *is assumed to grow at a constant rate* r*, that is*

$$\frac{\mathrm{d}L_t}{\mathrm{d}t} = r\,L_t\,, \qquad L_0 > 0\,.$$

Capital is denoted by K, aggregate production is Y, and aggregate consumption is C. We define the 'per labour' variables

$$c_t = \frac{C_t}{L_t}, \qquad k_t = \frac{K_t}{L_t}.$$

Denote total production by Y, and suppose that unconsumed production is invested in capital. Simultaneously, the current capital stock depreciates at rate δ. Thus, the change in capital is given by

$$\frac{\mathrm{d}K_t}{\mathrm{d}t} = \underbrace{Y_t}_{production} - \underbrace{c_t\, L_t}_{consumption\ C} - \underbrace{\delta\, K_t\,.}_{depreciation}$$

We model the relationship between aggregate production, capital stock, and labour using a 'Cobb–Douglas production function'

$$Y_t = F(K_t, L_t) = A\, K_t^{1-\alpha}\, L_t^{\alpha}\,, \tag{6.15}$$

with $\alpha \in (0,1)$ and $A > 0$. It is easy to check that

$$F(K_t, L_t) = L_t\, F(K_t/L_t, 1) = L_t\, f(k_t)$$

for $f(x) = F(x, 1)$.

(i) Show that

$$\frac{\mathrm{d}k_t}{\mathrm{d}t} = f(k_t) - (r + \delta)\, k_t - c_t\,.$$

(ii) Suppose a social planner wishes to maximise discounted utility of consumption, that is

$$\int_0^\infty e^{-\rho\, t}\, L_t\, \pi(c_t)\, \mathrm{d}t,\, ,$$

for $\rho > 0$. Here the initial capital is strictly positive, $\rho > r$, and $\pi(c)$ is strictly increasing, concave, and satisfies[7]

$$\lim_{c \downarrow 0} \pi'(c) = \infty.$$

By treating k as a state variable, c as a control, and applying Pontryagin's principle over an arbitrary horizon, show that the optimal rate of consumption satisfies

$$\frac{\mathrm{d}c_t^*}{\mathrm{d}t} = -\frac{\pi'(c_t^*)}{\pi''(c_t^*)}\big(f'(k_t^*) - \delta - \rho\big)\,.$$

(iii) Consider the resulting Ramsey–Cass–Koopmans dynamical system (which describes the locally optimal trajectory of the economy from a given initial state)

$$\begin{cases} \frac{\mathrm{d}k_t}{\mathrm{d}t} = f(k_t) - (r + \delta)\, k_t - c_t \\ \frac{\mathrm{d}c_t}{\mathrm{d}t} = -\frac{\pi'(c_t)}{\pi''(c_t)}\big(f'(k_t) - \delta - \rho\big)\,. \end{cases}$$

[7] These are related to the 'Inada conditions' commonly assumed in macroeconomic models.

Find the three stationary solutions, that is, values of (k_0, c_0) *such that*

$$\frac{\mathrm{d}k_t}{\mathrm{d}t} = \frac{\mathrm{d}c_t}{\mathrm{d}t} = 0\,.$$

Exercise 6.4.3 (Optimal stopping problem)**.** *Consider the optimization problem of finding* $\tau \in [0, T]$ *to minimize*

$$\int_0^\tau G(s, X_s)\mathrm{d}s + \Phi(X_\tau),$$

when

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t = F(t, X_t),$$

where $X \in \mathbb{R}^m$*,* F *is Lipschitz in* x*, and* G *is integrable.*

(i) By interpreting this as an optimal control problem with controlled drift $f(t, X_t, U_t) = F(t, X_t)U_t$ *and* $U_t \in \{0, 1\}$*, and similarly for* g*, write down the Hamilton–Jacobi equation describing the minimal value of this problem.*

(ii) Show that the Hamilton–Jacobi equation is equivalent to the linear complementarity problem

$$0 = \min\Big\{\Phi(x) - v(t, x), \partial_t v + G(t, x) + \langle \nabla v, F(t, x)\rangle\Big\}.$$

Exercise 6.4.4. *Suppose the (un-minimized) Hamiltonian* $\tilde{H}(t, x, p, u)$ *is convex with respect to* u*, where* u *takes values in* $\mathcal{U} = \mathbb{R}^m$*, and* $H(t, x, p, u) \to \infty$ *as* $\|u\| \to \infty$*.*

(i) Show that Pontryagin's minimum principle can be expressed in the form:

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t^* = \partial_p \tilde{H}, \qquad \frac{\mathrm{d}}{\mathrm{d}t}p_t^* = \rho p_t^* - \partial_x \tilde{H}, \qquad 0 = \partial_u \tilde{H}.$$

(ii) Show that, if $\rho = 0$ *and the Hamiltonian does not depend on time, and the optimal strategy is differentiable with respect to time, then Pontryagin's minimum principle shows that, along the optimal path,*

$$\tilde{H}(X_t^*, p_t^*, u_t^*) = \text{constant}.$$

*(iii) (*Trickier*) Suppose that the state dynamics* f *are Lipschitz continuous, the Hamilton–Jacobi equation admits a* C^1 *solution* v *with Lipschitz continuous derivatives, and* $\tilde{H}$ *is twice differentiable and strictly convex with respect to* u*, in particular* $\partial^2_{uu}\tilde{H}(t, u, x, p)$ *has all eigenvalues above* ε*, for all* t, u, x, p*.*

Show that there exists an optimal control, and that this control is unique.

(iv) Explain why, in the setting above, Pontryagin's minimum principle is both necessary and sufficient for optimality.

Exercise 6.4.5. *Suppose we have controlled state dynamics f which satisfy (for all x, u)*

$$\langle x, f(x,u)\rangle \leq C(1 + \|x\|^{3/2})$$

for some constant C.

(i) Using the comparison theorem for ODEs, show that for any control U, the controlled trajectory satisfies $\|X_t^U\| \leq C(1+\|x_0\|)(1+t^2)$, for some $C > 0$.

Hint: The process y_t with $y_0 = 1$ and dynamics $\frac{\mathrm{d}}{\mathrm{d}t}y_t = Cy_t^{1/2}$ may provide a useful upper bound on $\|X_t^U\|^2/(1 + \|x_0\|^2)$.

(ii) Consider controlling this process over the horizon $[0, T]$, where we assume the cost process satisfies $|g(t, x, u)| < C(1 + \|x\|)$ and we have a discount rate $\rho > 0$. Show that the total discounted cost of any control remains bounded as $T \to \infty$.

(iii) Suppose the time-homogenous Hamilton–Jacobi equation

$$0 = -\rho v + H(x, \nabla v)$$

*admits a bounded C^1 solution. Assuming the dynamic programming principle holds, show that this must be the value function of the infinite horizon discounted control problem. (*Hint: Compare the behaviour over finite horizons.*)*

Exercise 6.4.6. *Consider a deterministic control problem with underlying state dynamics satisfying the scalar second order controlled ODE*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}X_t = U_t - \frac{\mathrm{d}}{\mathrm{d}t}X_t$$

with initial values $(X_0, \dot{X}_0) = (1, 0)$. Suppose we seek to minimize the value of

$$\int_0^T (X_t^2 + U_t^2)\mathrm{d}t.$$

(i) Re-express this problem as a vector-valued first order equation, and hence derive an equation for the value function in terms of a system of Riccati equations, and for the optimal control in feedback form.

(ii) Using a backwards Euler scheme, solve the system of Riccati equations numerically, and hence state the value, and optimal control at time $t = 0$.

Exercise 6.4.7. *Consider a deterministic optimal control problem, with optimal state trajectory X^*, and costs $\int_0^T g(t, X_t^U, U_t)\mathrm{d}t + \Phi(X_T^U)$.*

(i) Show that X^* is still optimal for the problem with costs defined by $\tilde{g}(t, x, u) = g(t, x, u) + \alpha 1_{\{\|x - X_t^*\| > \varepsilon\}}$ and $\tilde{\Phi}(x) = \Phi(x) + \alpha 1_{\{\|x - X_T^*\| > \varepsilon\}}$, for any $\alpha, \varepsilon > 0$.

(ii) *Show that X^* may not be optimal for the problem with costs defined by $\tilde{g}(t,x,u) = g(t,x,u) - \alpha 1_{\{\|x - X_t^*\| > \varepsilon\}}$ and $\tilde{\Phi}(x) = \Phi(x) - \alpha 1_{\{\|x - X_T^*\| > \varepsilon\}}$, for some choices of $\alpha, \varepsilon > 0$.*

(iii) *Suppose Pontryagin's principle is satisfied for the original problem. Show that it also holds for both of the variations above. What conclusion do you draw?*

Exercise 6.4.8. *Consider a deterministic Mayer-form optimal control problem on a finite horizon $\mathbb{T} = [0,T]$, that is we seek to minimize a terminal cost*

$$J(U) = \Phi\big(X_T^U\big).$$

Suppose we have linear state dynamics

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = A(t)X_t + B(t)u,$$

where $X_t \in \mathbb{R}^n$, $U_t \in \mathcal{U} \subseteq \mathbb{R}^m$, and $A(\cdot), B(\cdot)$ are given continuous matrix-valued functions. Assume $\mathcal{U}$ is compact and suppose that the terminal cost $\Phi : \mathbb{R}^n \to \mathbb{R}$ is convex.

The value function is

$$v(t,x) = \inf_U \Phi\big(X_T^{t,x,U}\big),$$

where $X^{t,x,U}$ denotes the state trajectory starting from (t,x).

(i) *Show that for a fixed control $U : \mathbb{T} \to \mathcal{U}$, the mapping $x \mapsto X_T^{t,x,U}$ is affine in x.*

(ii) *Deduce that for fixed U, the mapping $x \mapsto \Phi\big(X_T^{t,x,U}\big)$ is convex.*

(iii) *Recall that the infimum of an arbitrary family of convex functions need* not *be convex. Explain why this observation prevents one from immediately concluding that $v(t,x)$ is convex in x.*

(iv) *Suppose now that $\mathcal{U}$ is convex and observe that the control enters linearly in the dynamics. Show that the set*

$$\mathcal{R}_{t,T}(x) = \{X_T^{t,x,U}\}_{U \in \mathbb{U}}$$

of reachable terminal states is convex in x. (Hint: use convex combinations of controls.)

(v) *Show that under these assumptions the value function satisfies the convexity statement*

$$v(t, \theta x_1 + (1-\theta)x_2) \le \theta v(t,x_1) + (1-\theta) v(t,x_2), \qquad \textit{for } \theta \in [0,1].$$

(vi) *Give an example showing that convexity may fail if either*

(a) the control set U is not convex, or

(b) the dynamics are nonlinear in x or u.

(vii) How does convexity of V relate to convexity of the Hamiltonian for this problem?

Exercise 6.4.9. *Let $S^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}$ be the unit sphere in $\mathbb{R}^3$. Write $P_X := I - XX^\top$ for the orthogonal projection onto the tangent plane $T_X S^2$ at X. Consider the controlled system in the ambient space*

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = P_{X_t} U_t, \qquad X_t \in S^2, \quad U_t \in \mathbb{R}^3,$$

i.e. the control U_t is an unconstrained vector in $\mathbb{R}^3$ and only its projection $P_{X_t} U_t$ impacts the motion on the sphere.

We seek to minimize the control energy

$$\mathcal{E}(\gamma, U) = \frac{1}{2} \int_0^T \|U_t\|^2 \, \mathrm{d}t$$

over all admissible pairs (γ, U) with $\gamma(0) = x_0$, $\gamma(T) = x_1$, and $\gamma = \{X_t\}_{t \in [0,T]}$ satisfying the dynamics above.

(i) Explain why the Hamiltonian for this problem is given by

$$\tilde{H}(x, p, u) \;=\; \tfrac{1}{2}\|u\|^2 + p \cdot (P_X u),$$

where $p(t) \in \mathbb{R}^3$ is the costate with terminal value $p(T) = 0$ (as discussed, as we will take the terminal value to be fixed).

(ii) Compute the first–order condition $\partial_u \tilde{H} = 0$ and solve for U^* in terms of p and X. Conclude that the optimal control is the (negative) projected costate:

$$U^* = -P_X p.$$

Hint: It may help to observe that $P_X^\top = P_X$ and $P_X^2 = P_X$.

(iii) Express the costate dynamics in terms of $X, p, \frac{\mathrm{d}}{\mathrm{d}t} X$.

(iv) Combine the expressions for $\frac{\mathrm{d}}{\mathrm{d}t} X$ and $\frac{\mathrm{d}}{\mathrm{d}t} p$ to show that an optimal trajectory satisfies the second–order geodesic ODE

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} X_t = -\left\| \frac{\mathrm{d}}{\mathrm{d}t} X_t \right\|^2 X_t.$$

(v) Consider the energy-minimizing trajectory from a fixed target point x_1. Define $v(x) = \arccos(x_1 \cdot x)$, which gives the geodesic distance between two points on the surface of a sphere. Show that v solves the static Hamilton–Jacobi equation for $x \in S^2 \setminus \{x_1, -x_1\}$:

$$|P_X \nabla v(x)| = 1.$$

(vi) *Let $x_0 = -x_1$ be the antipode of x_1. Show that every great circle through x_1 provides a minimizing trajectory starting at x_0, so there are infinitely many minimizing geodesics between antipodes.*

(vii) *Explain how the value function $v(x) = \arccos(x_1 \cdot x)$ exhibits a singularity at the antipode: although v is continuous and attains value π at $-x_1$, the gradient ∇v does not exist there. Comment on how this is reflected in the non-uniqueness of geodesic paths between antipodes.*

Exercise 6.4.10 (Deterministic optimal execution)**.** *An active field of research is that of optimal execution of large orders in financial markets [2]. Here, the goal is to unwind a large position in a given asset while minimising costs. The costs are mainly from the impact of trades on future prices (i.e. the more quickly you sell, the lower the price in the future all else being equal). One of the classical examples (see Chapter 6.5 in [14]) to study this problem can be reduced to a deterministic continuous time control problem.*

Consider the dynamics

$$\mathrm{d}X_t = -U_t \,\mathrm{d}t, \qquad X_0 = x_0\,,$$

where x_0 is the initial number of shares that the agent wishes to liquidate, positive values of U_t denote selling, and X_t is the number of shares left to liquidate at time t.

The classical temporary-and-permanent price impact control problem can be expressed as minimizing

$$\phi\, X_T^2 + \int_0^T \left(k\, U_t^2 + b\, U_t\, X_t + \alpha\, X_t^2\right) \mathrm{d}t\,.$$

Here $T > 0$ is the time horizon of the liquidation programme,

1. *$k > 0$ is the* temporary price impact*, interpreted as the cost due to not being able to trade arbitrarily large quantities at the current price, so having to accept a lower price in order to trade at a high rate;*

2. *$b \geq 0$ is the* permanent price impact*, interpreted as the cost due to your trades affecting the price expectations of other participants in the market, and hence reducing the notional value of your current share holdings;*

3. *$\phi \geq 0$ is the* terminal inventory penalty *and $\alpha \geq 0$ is a* running inventory penalty*, interpreted as an incentive to fully liquidate the position, and to do so quickly, so as to avoid the risk of future (random) price changes.*

We assume that $\phi - 0.5\, b > 0$.

(i) *Write down the Hamilton–Jacobi equation for this problem, and solve the PDE with a quadratic ansatz in X.*

(ii) *Compute the solution to this problem using a Pontryagin minimum principle approach.*

(iii) Discuss the behaviour of the optimal strategy as $\alpha \to \infty$.

Exercise 6.4.11. *Let $T > 0$ and fix a constant $K \in \mathbb{R}$ with $K < -T$. Consider the finite-horizon control problem on the interval $[0, T]$ with state dynamics*

$$\mathrm{d}X_t = U_t \mathrm{d}t, \qquad X_0 = x_0 > 0,$$

where controls U are taken from $[\eta^{-1}, \eta]$, for some η sufficiently large. The objective is to minimise *the cost*

$$J(t, x, U) = \int_t^T \mathrm{e}^{-\rho(s-t)} \big(-\log(U_s) \big)\, ds \;+\; \mathrm{e}^{-\rho(T-t)} \big(-K \log(X_T) \big),$$

with discount rate $\rho \geq 0$.

(i) Write down the Hamilton–Jacobi PDE for this problem.

(ii) We will seek a solution to the HJB equation of the separable logarithmic form

$$v(t, x) = A(t) + B(t) \log x, \qquad t \in [0, T],\ x > 0,$$

with terminal conditions $A(T) = 0,\ B(T) = -K$.

a. Show that, for a smooth v of the ansatz form, and assuming an interior solution, the pointwise minimisation

$$\inf_{u \in [\eta^{-1}, \eta]} \{ -\log u + \partial_x v(t, x)\, u \}$$

has first-order condition

$$-\frac{1}{\mathbf{u}^*(t, x)} + \partial_x v(t, x) = 0,$$

and hence the candidate optimal control is determined by the feedback function

$$\mathbf{u}^*(t, x) = \frac{x}{B(t)}.$$

b. Substitute the ansatz and the minimiser back into the HJB PDE, and hence derive ODEs determining A and B. Give a condition on ρ such that $B(0) > 0$.

c. Write X^* for the state process obtained by following the feedback control $U_t^* = \mathbf{u}^*(t, X_t^*)$. Show that $\frac{\mathrm{d}}{\mathrm{d}t} U_t^* = -\rho U_t^*$, and hence find a condition which ensures that the candidate control takes values in $\mathcal{U} = [\eta^{-1}, \eta]$.

(iii) Assume the function v defined by the formulas in (ii) is $C^{1,1}$ on $[0, T) \times (0, \infty)$ and satisfies the HJB PDE.

a. Apply the fundamental theorem of calculus to the function

$$s \mapsto \mathrm{e}^{-\rho(s-t)} v(s, X_s^{t,x,\mathbf{u}^*})$$

on $[t, T]$ and hence show that $U^* = \mathbf{u}^*(t, X_t^*)$ attains the minimal cost.

b. *Show that for any admissible control U and the corresponding state $X^{t,x,U}$,*

$$v(t,x) \leq \int_t^T \mathrm{e}^{-\rho(s-t)}\big(-\log(U_s)\big)\,ds \;+\; \mathrm{e}^{-\rho(T-t)}\big(-K\log(X_T^{t,x,U})\big),$$

and so v is indeed the value function and $U^* = \mathbf{u}^*(t, X_t^*)$ is optimal.

Exercise 6.4.12 (Optimal control of an SIR epidemic)**.** *Consider a population of fixed size N modelled by an SIR system, which is a standard basic model of an epidemic. The state variables represent numbers of people of each type, and are $S(t)$ (susceptible), $I(t)$ (infected), and $R(t)$ (removed/recovered), with*

$$S(t) + I(t) + R(t) = N.$$

We introduce a time-dependent control $U_t \in [0, u_{\max}]$ representing the intensity of government policies (e.g. social distancing or contact reduction). The control reduces the transmission rate multiplicatively.

The controlled state dynamics on a time horizon $[0,T]$ are

$$\begin{aligned}\dot{S}(t) &= -\beta(1-U_t)\frac{S(t)I(t)}{N},\\ \dot{I}(t) &= \beta(1-U_t)\frac{S(t)I(t)}{N} - \gamma I(t),\\ \dot{R}(t) &= \gamma I(t),\end{aligned}$$

with initial condition

$$(S(0), I(0), R(0)) = (S_0, I_0, R_0),$$

where $\beta > 0$ is the infection rate and $\gamma > 0$ is the recovery rate.

We penalize infections and control efforts. Consider the cost functional

$$J(u) \;=\; \int_0^T \Big(A\,I(t) + \tfrac{1}{2}B\,U_t^2\Big)\,dt + C\,I(T),$$

with constants $A, B > 0$, $C \geq 0$. Here A measures the running health cost of infections, B the quadratic economic and social costs of control, and C a terminal weight.

(i) *Explain why the control enters as $\beta(1-u)$ and why u is constrained to $[0, u_{\max}]$. What do the extremes $u = 0$ and $u = u_{\max}$ represent?*

(ii) *Write down the (unoptimized) Hamiltonian*

$$\tilde{H}\big(t, x, p, u\big)$$

for this problem, where $X = (S, I, R)$ is the state and $p = \nabla v(t,x) \in \mathbb{R}^3$ is the adjoint (costate) variable. Write down the adjoint ODEs and the terminal condition at $t = T$ from Pontryagin's principle.

(iii) Derive the first-order optimality condition for $u(t)$ from Pontryagin's principle. Compute the unconstrained minimizer $u^{\text{uncon}}(t)$ in terms of the states and costates, and write the admissible optimal control in projected form

$$u^*(t) = \operatorname{proj}_{[0,u_{\max}]}\big(u^{\text{uncon}}(t)\big).$$

Give the explicit projection formula.

(iv) Explain how the parameter B influences whether the optimal control is bang-bang (i.e. takes only values 0 and $u_{\max}$) or interior (takes intermediate values). Give an epidemiological interpretation.

(v) Implement a forward-backward search algorithm to approximate the optimal control for the parameter set

$$\beta = 0.6, \quad \gamma = 0.2, \quad N = 10^6, \quad S_0 = N - 100, \quad I_0 = 100, \quad R_0 = 0,$$

$$A = 1, \quad B = 10^4, \quad C = 1, \quad u_{\max} = 0.9, \quad T = 365.$$

Describe your numerical method, plot $S(t), I(t), R(t)$ and $u^*(t)$, and report the resulting value of $J(u^*)$.

Exercise 6.4.13 (Optimal control of an inverted pendulum)**.** *Consider a rigid pendulum of length ℓ and mass m that is free to rotate in the vertical plane. We control the pendulum by applying a torque $u(t)$ at the pivot. Let $\theta(t)$ denote the angle measured from the upright vertical position, so $\theta = 0$ is the unstable upright equilibrium; small positive θ means the pendulum has fallen slightly forward. The dynamics are*

[SC]$_6$:would be nice to have a picture here

$$m\ell^2 \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = mg\ell \sin\theta(t) + u(t),$$

or, written as a first-order system with state $X = (X_1, X_2)^\top = (\theta, \dot\theta)^\top$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}X_1(t) &= X_2(t), \\ \frac{\mathrm{d}}{\mathrm{d}t}X_2(t) &= \frac{g}{\ell}\sin X_1(t) + \frac{1}{m\ell^2}u. \end{aligned}$$

We consider controls in $\mathcal{U} = \mathbb{R}$ for the finite-horizon optimal control problem on $[0, T]$:

$$\min_{u(\cdot)} \ J(u) \ = \ \int_0^T \Big(\tfrac{1}{2}Q_1 x_1(t)^2 + \tfrac{1}{2}Q_2 x_2(t)^2 + \tfrac{1}{2}R\, u(t)^2\Big)\, dt + \tfrac{1}{2}S\, x_1(T)^2,$$

with $Q_1, Q_2, R, S \geq 0$ (weights). The objective penalizes deviations from the upright position and control effort; the terminal weight penalizes final angle error.

(i) Write down the unoptimized Hamiltonian for the nonlinear problem and derive Pontryagin's necessary conditions.

(ii) *Show that the first-order condition yields a candidate optimal control satisfying*

$$U(t) = -\frac{1}{Rm\ell^2}\, p_2(t).$$

where $p = [p_1, p_2]^\top$ *is the costate process.*

(iii) *Derive the two-point boundary-value problem that must be solved to obtain an extremal control. Outline a numerical method to compute a candidate optimal control.*

(iv) *For small-angle deviations, approximate* $\sin\theta \approx \theta$ *and derive the linearised dynamics*

$$\dot{X} = AX + Bu, \qquad A = \begin{pmatrix} 0 & 1 \\ g/\ell & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1/(m\ell^2) \end{pmatrix}.$$

Write the corresponding finite-horizon Linear-Quadratic Regulator (LQR) problem and show that the optimal state-feedback control has the form

$$u(t) = -K(t)\, x(t),$$

where $K(t) = R^{-1}B^\top P(t)$ *and* $P(t)$ *solves the matrix Riccati differential equation*

$$-\dot{P}(t) = A^\top P + PA - PBR^{-1}B^\top P + Q, \qquad P(T) = S,$$

with

$$Q = \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix}.$$

(v) *Discuss the relation between the nonlinear Pontryagin extremals and the LQR solution: under what circumstances is the LQR feedback a good approximation?*

Exercise 6.4.14 (Optimal control of a lunar lander)**.** *We consider a simplified one-dimensional model of the Apollo lunar lander during the final vertical descent to the Moon's surface.*

The dynamics are

$$\begin{aligned} \dot{h}(t) &= v(t), \\ \dot{v}(t) &= -g + \frac{u(t)}{m(t)}, \\ \dot{m}(t) &= -\alpha\, u(t), \end{aligned}$$

with initial conditions $h(0) = h_0 > 0$, $v(0) = v_0 < 0$, $m(0) = m_0 > 0$ *and control bounds* $0 \le u(t) \le u_{\max}$. *Here* h *is the height from the moon surface,* v *is the lander velocity, and* m *is the total mass of the lander including fuel. The parameter* $\alpha > 0$ *is the fuel consumption of our control, and* $g > 0$ *is lunar gravity.*

We minimise a cost that trades off fuel and terminal landing accuracy (both in height and velocity):

$$J(U) = \int_0^T U(t)\,\mathrm{d}t \;+\; \frac{K_h}{2}\,h(T)^2 \;+\; \frac{K_v}{2}\,v(T)^2,$$

with penalty weights $K_h, K_v \geq 0$.

(i) Write down the unoptimized Hamiltonian for this problem and derive the adjoint (costate) equations in Pontryagin's principle, together with their terminal values.

(ii) Compute $\partial\tilde{H}/\partial u$ and characterise the pointwise minimiser of the Hamiltonian. Explain why the control will typically be of bang–bang type (that is, will only take the values 0 and $u_{\max}$).

(iii) Discuss qualitatively how the penalisation parameters K_h, K_v influence the terminal values of the costates and the optimal strategy. What happens in the limits $K_h, K_v \to 0$ and $K_h, K_v \to \infty$?

(iv) (Numerical experiment) Using semi-realistic values for the physical parameters:

$$\begin{aligned} &g = 1.62\ m/s^2, \quad u_{\max} = 45\,000\ N, \quad m_0 = 15\,000\ kg,\\ &\alpha = 0.35 \times 10^{-3}\ kg/N \end{aligned}$$

and initial values $h_0 = 2000\ m$, $\ v_0 = -50\ m/s$, for and with K_h, K_v large, simulate the optimal policy and determine:

(a) the time at which the thrusters are used (i.e. when $U_t > 0$),

(b) the total fuel consumption,

(c) the touchdown time.

Chapter 7

Continuous-time Stochastic Control

In this chapter, 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 establish some continuity estimates for the value function directly, 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).

7.1 Notation and problem formulation

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

$$\mathrm{d}X_t = f(\omega, t, X_t, U_t)\mathrm{d}t + \sigma(\omega, t, X_t, U_t)\mathrm{d}W_t \tag{7.1}$$

where W is an $\mathbb{R}^m$-dimensional Brownian motion, $f : \Omega \times \mathbb{T} \times \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$, $\sigma : \Omega \times \mathbb{T} \times \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^{d\times m}$. As usual, we will often drop ω in our notation.

The process U is an $\{\mathcal{F}_t\}_{t\in\mathbb{T}}$-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. In order to guarantee our processes are well defined, we may need to impose additional integrability assumptions on U (which will naturally depend on the behaviour of f and σ); we therefore write $\mathbb{U}$ for the space of *admissible* controls (where the required assumptions are imposed). We say U is of feedback form if there is a (Borel measurable) function u such

[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.

that $U_t = \mathbf{u}(t, X_t)$ 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 $\mathbf{u}(t, X_t)$ 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[4] of controlled processes

$$X_{t'}^{t,x,U} = x + \int_t^{t'} f(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \int_t^{t'} \sigma(s, X_s^{t,x,U}, U_s)\mathrm{d}W_s \tag{7.2}$$

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

We hope to define a value function

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

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 in many cases 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 - at this stage, this doesn't even guarantee that v is measurable with respect to (t, x).

Example 7.1.1 (Merton portfolio problem). *A classic financial example is as follows. Consider a financial asset described by the SDE*

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

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

[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.

[4] It turns out to be remarkably not obvious to guarantee that $X_{t'}^{t,x,U}$ satisfies a flow property of the type $X_{t'}^{t,x,U} = X_{t'}^{s, X_s^{t,x,U}, U}$ for all $t < s < t'$, due to measurability issues. For this reason, we will avoid explicitly using such an assumption.

each time. Their wealth (assuming a zero interest rate) then is modelled by the SDE

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

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).

We have already seen a version of the martingale optimality property in discrete time (Theorem 2.1.16). We will prove a version of this result in Corollary 7.2.10, once we have constructed the value function carefully (which proves more difficult than one might expect!). Here, however, we will state a useful result which allows us to characterize optimal controls in cases where we are able to 'guess' what the value function is. In particular, this will apply with relatively few assumptions on the problem, which means we will have the option of using our results below to construct what we believe is the value function (even if we don't satisfy all the assumptions), and then check (using this result) that what we have constructed is, in fact, correct!

Theorem 7.1.2 (Martingale verification)**.** *Consider a random field*[5] *$v : \Omega \times \mathbb{T} \times \mathcal{X} \to \mathbb{R}$ which is adapted and right-continuous with respect to (ω, t), Borel measurable with respect to x, and satisfies*

$$v(\omega, T, x) = \Phi(\omega, x) \textit{ for all } x, \textit{ almost surely.}$$

Assume that, for each $(t, x) \in \mathbb{T} \times \mathcal{X}$ and each admissible control $U \in \mathbb{U}$, the controlled SDE 7.2 has a unique strong[6] *solution $X^{t,x,U}$. For each $(t, x, U) \in \mathbb{T} \times \mathcal{X} \times \mathbb{U}$, define the process*

$$M_{t'}^{t,x,U}(\omega) = \int_t^{t'} g(\omega, s, X_s^{t,x,U}, U_s)\mathrm{d}s + v(\omega, t', X_{t'}^{t,x,U}).$$

Suppose that, for every $(t, x, U) \in \mathbb{T} \times \mathcal{X} \times \mathbb{U}$, the process $M^{t,x,U}$ is a submartingale (when restricted to $[t, T]$). Suppose also that for each (t, x) there exists some $U^{(t,x)} \in \mathbb{U}$ such that $M^{t,x,U^{(t,x)}}$ is a martingale. Then

$$v(\omega, t, x) = \operatorname*{ess\,inf}_{U \in \mathbb{U}} \Big\{ \mathbb{E}\Big[\int_0^T g(t, X_t^{t,x,U}, U_t)\mathrm{d}t + \Phi(X_T^{t,x,U}) \Big| \mathcal{F}_t \Big] \Big\}$$

[5] Of course, a function $w : \mathbb{T} \times \mathcal{X} \to \mathbb{R}$ is a special case of a random field, so this case is immediately included in this result.

[6] If we restrict v to being a function (rather than a random field), and assume g and Φ do not have direct dependance on ω (cf. Assumption 7.3.1), then it's fairly easy to give a version of this result assuming only that $X^{t,x,U}$ has a weak solution which is unique in law.

(so v is a version of the value function), and $U^{(t,x)}$ is an optimal control for the problem started at (t,x).

Proof. The proof is almost identical to that in discrete time (Theorem 2.2.5). For any $U \in \mathbb{U}$, as $M^{t,x,U}$ is a submartingale, we know that

$$\begin{aligned} v(\omega,t,x) &= M_t^{t,x,U} \\ &\leq \mathbb{E}[M_T^{t,x,U}|\mathcal{F}_t] = \mathbb{E}\Big[\int_t^T g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + \Phi(X_T^{t,x,U})\Big|\mathcal{F}_t\Big]. \end{aligned}$$

On the other hand, we know that there exists $U^{(t,x)} \in \mathbb{U}$ such that

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

It follows immediately that v is a version of the desired essential infimum, and that $U^{(t,x)}$ achieves at least as low an expected cost (given $\mathcal{F}_t$, almost surely) as any other admissible policy, and hence is optimal for the problem started at (t,x). □

Remark 7.1.3. If we replace our terminal horizon T with an almost-surely finite stopping time, the above argument does not change.

7.2 Dynamic programming

Our next goal is to construct the value function, and show that it satisfies a dynamic programming equation. Essentially, this involves two closely related problems: finding a sufficiently regular (with respect to (t,x)) definition of the value that our problems started at different times interact well, and constructing a near-optimal control at each point in time, which allows us to determine dynamic properties of the value function from those of the cost-to-go.

7.2A A first set of assumptions

In order for all of this to be well posed, we will make the following (somewhat restrictive) assumptions, which allow us to give bounds on the controlled process.

Assumption 7.2.1. *There exists a constant $K < \infty$ such that*

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

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

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

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

(iii) g and Φ satisfy the bounds, for some $k \geq 1$ and all t, x, x', u,

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

(iv) The space of admissible controls $\mathbb{U}$ is equal to the whole space of all $\{\mathcal{F}_t\}_{t\geq 0}$-progressive processes taking values in $\mathcal{U}$.

Remark 7.2.2. 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 (with a power $\|x\|^{k+1}$ for 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 7.2.3. The assumption that controls have only a bounded impact on costs, and the strong Lipschitz assumptions on our dynamics, are somewhat restrictive, but we will lift them in what follows.

Remark 7.2.4. The key property of $\mathbb{U}$, guaranteed by (iv) above, 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 $\tilde{U}_s := U^0_s \mathbf{1}_{s\leq\tau} + \sum_{i\in\mathbb{N}} \mathbf{1}_{s>\tau}\mathbf{1}_{A_i} U^i_s \in \mathbb{U}$.

7.2B 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, in some sense similarly to what we did for the strong formulation in discrete time.

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 7.2.5. *Under Assumption 7.2.1, we have the following standard properties:*

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

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

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

(iii) 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 $t \leq t'$,

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

In particular, for each choice of U, there exists a version of $X_s^{t,x,U}$ which is continuous with respect to (t, x) (i.e. a function $\mathfrak{X}$ with $\mathfrak{X}(\omega, t, x, s) = X_s^{t,x,U}$ almost surely, for all t, x, s).

Proof. The core of these results is presented in Appendix A.2A, which gives the more general case of SDEs with stochastic dynamics – this can be applied here by setting $\mu(\omega, t, x) = f(\omega, t, x, U_t(\omega))$, and similarly for σ, from which Theorem A.2.1 yields the existence of the unique solution.

To see the stated bounds, set $\mu(\omega, r, \xi) = f(\omega, r, \xi, U_r(\omega))\mathbf{1}_{t\leq r\leq T}$ and $\tilde{\mu}(\omega, r, \xi) = f(\omega, r, \xi, U_r(\omega))\mathbf{1}_{t'\leq r\leq T}$, 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 $t \leq t'$,

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

which we further bound as

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

The main result follows as $\mathbb{E}\big[\|X_r^{t,x,U}\|^2|\mathcal{F}_t\big] \leq C(1+\|x\|^2)$. The existence of the continuous version $\mathfrak{X}$ follows from the Kolmogorov continuity criterion[7]. □

[7] 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 [41, 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.

Lemma 7.2.6. *Under Assumption 7.2.1, for any $K > 0$, the set of random variables*

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

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

Proof. Our assumptions guarantee that shown that Φ and g are of polynomial growth. Applying Jensen's inequality, we see that there exists $C, p > 0$ such that

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\Phi(X_T^{t,x,U})\Big)^2 + \Big(\int_t^{t'} g(s, X_T^{t,x,U}, U_s)\mathrm{d}s\Big)^2\Big]\\
&\leq \mathbb{E}\Big[\Big(\Phi(X_T^{t,x,U})\Big)^2 + T\int_t^{t'} \Big(g(s, X_T^{t,x,U}, U_s)\Big)^2\mathrm{d}s\Big]\\
&\leq C\mathbb{E}\Big[\sup_{s\in[t,T]} \|X_s^{t,x,U}\|^p\Big].
\end{aligned}$$

Our bounds in Lemma 7.2.5(ii) show that this quantity is uniformly bounded for $\|x\| < K$. Applying the de la Vallée Poussin criterion for uniform integrability, the result follows. □

Using these bounds, we can obtain a continuity estimate on the cost-to-go, and hence establish the existence of the value function simultaneously for all (t, x).

Theorem 7.2.7. *Under Assumption 7.2.1, there exists a (deterministic) constant $K < \infty$ such that, with $k \geq 1$ 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 $x, x' \in \mathcal{X}$,*

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

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 $t \leq t'$,

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

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

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

where we have repeatedly used the inequality $(x+y)^{1/2}\le x^{1/2}+y^{1/2}$ 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, 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 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.

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 7.2.8. *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 fixed deterministic time $t' > t$. We will first consider how much error is introduced if we were to re-start our process at time t', starting at an approximation of $X_{t'}^{t,x,U}$.

For arbitrary $\epsilon > 0$, and take a countable partition $\{A_n^\epsilon\}_{n\in\mathbb{N}}$ of $\mathcal{X}$ such that $\max_{x,x'\in A_n^\epsilon}\{\|x-x'\|^2\} \leq \epsilon$. Choose[8] a point $x_n \in A_n^\epsilon$ for each n. We now define an approximation process

$$\hat{X}_s^{t,x,U,\epsilon} = \sum_n \mathbf{1}_{\{X_{t'}^{t,x,U} \in A_n^\epsilon\}} X_s^{t',x_n,U}$$

which is well defined for all $s \geq t'$. As this involves only countably many points x_n, we know that this is well defined and measurable, and $\hat{X}_{t'}^{t,x,U,\epsilon} = x_n$ on A_n^ϵ.

We now consider the difference between $X_s^{t,x,U}$ and $\hat{X}_s^{t,x,U,\epsilon}$; by a variation of the bounds we showed above, we know that there exists K such that, almost

[8] 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!

surely,

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

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

$$\hat{X}_s^{t,x,U,\epsilon_k} \to X_s^{t,x,U} \qquad (\mathrm{d}s + \delta_{s=T}) \times \mathrm{d}\mathbb{P}\text{-a.e.}$$

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

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

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

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

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

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

From standard stochastic analysis results, for any integrable random variable, its conditional expectations define a uniformly integrable martingale, and under the usual conditions this martingale will admit a right continuous modification. Therefore, we define a (right continuous, uniformly integrable) martingale M by

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

Using our earlier representation of J, almost surely, for each t', we know

$$M_{t'} = \int_t^{t'} g(s, X_s^{t,x,U}, U_s)\mathrm{d}s + J(t', X_{t'}^{t,x,U}, U).$$

Now observe that both sides of this equality are right continuous in t', so this equality holds up to a null set independent of time (and the right hand side is continuous in t', which implies M is also continuous). The martingale optional stopping theorem implies that

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

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

7.2C Describing the value function

In order to obtain a dynamic programming equation, we need to consider the value function, rather than the cost-to-go. This will involve a similar approximation as in the previous lemma, but now including time, and approximating the controls, rather than just the value of X.

Theorem 7.2.9. *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 $\epsilon > 0$ 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

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

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

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

As this is based on a countable pasting, it is still admissible. Furthermore, from the continuity estimates above, writing $\tilde{K}_i = K(1 + 2\|X_\tau^{t,x,U}\|^{2k} + \|x_i\|^{2k} + \epsilon)$ for K as in Theorem 7.2.7,

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

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)),

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

Finally, taking $\epsilon \to 0$, 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. □

We can now give the Martingale Optimality principle, which forms a converse (in this restrictive setting) to our earlier verification result (Theorem 7.1.2).

Corollary 7.2.10 (Martingale Optimality Principle)**.** *For any control $U \in \mathbb{U}$, the process defined by*

$$M^{t,x,U}_{t'} = v(t', X^{t,x,U}_{t'}) + \int_0^{t'} g(s, X^{t,x,U}_s, U_s)\mathrm{d}s$$

is a submartingale (when restricted to $t' \in [t,T]$), and is a martingale if and only if U is optimal.

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

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

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

This result tells us various things about our problem. From a modelling perspective, this is enough to ensure that a variant of Theorem 2.1.17 holds, with the natural defintion of 'optimal at time t' (and indeed, at a stopping time). However, as we are working in a general filtration, and all our costs and dynamics are allowed to depend on ω, and the value is only given as a random field, we cannot generally conclude that we have a feedback control policy, as the optimizers we construct are only guaranteed to be adapted (or more precisely, progressively measurable). In order to prove this, we need to consider the interaction between the state at time t and the conditional probability, which is most naturally done using variations of the Fokker–Planck equation.

7.3 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. To do this, we need to avoid having additional direct dependence on ω in our costs or dynamics.

Assumption 7.3.1. *In addition to Assumption 7.2.1, the dynamics and costs are given by functions f, σ, g with domain $\mathbb{T} \times \mathcal{X} \times \mathcal{U}$ and Φ with domain $\mathcal{X}$. In particular, the dynamics and costs do not depend on ω.*

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,

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

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

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

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

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

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. For simplicity, we will do this under the strong Assumption 7.2.1. By comparing with the assumptions in the martingale verification result in 7.1.2, we see that the key challenge is to use our PDE to construct a (near) optimal control, for which the state process is well defined.

Theorem 7.3.2 (HJB Verification)**.** *Assume that Assumption 7.2.1 holds. 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*

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

(i) Suppose that

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

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

(ii) Suppose that

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

Then $w(t,x)\ge v(t,x)$ 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 of feedback type).

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

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

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 our costs are uniformly integrable, we can take $n \to \infty$, and thus

$$\begin{aligned} w(t,x) = &\,\mathbb{E}\Big[w(\tau, X^{t,x,U}_\tau) + \int_t^\tau g(s, X^{t,x,U}_s, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big] \\ &- \lim_{n\to\infty} \mathbb{E}\Big[\int_t^{\tau\wedge\tau_n} \partial_t w(s, X^{t,x,U}_s) + \tilde{H}(s, X^{t,x,U}_s, D_x w, D^2_{xx} w, U_s)\mathrm{d}s\Big|\mathcal{F}_t\Big]. \end{aligned} \tag{7.3}$$

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 (7.3) is negative. Therefore, with $\tau = T$,

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

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 $\epsilon > 0$,

$$\tilde{H}(s, X_s^{t,x,U}, D_x w, D_{xx}^2 w, U_s) \leq H(s, X_s^{s,x,U}, D_x w, D_{xx}^2 w) + \epsilon.$$

For notational simplicity, write $h(t,x,u) = \tilde{H}(t,x,D_x w, D_{xx}^2 w, u)$. 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 $h(t_i, x_i, u_i) \leq \inf_u h(t_i, x_i, u) + \epsilon/2$. We define

$$\mathbf{u}^\epsilon(t,x) = u_0 \mathbf{1}_{t=0} + \sum_i u_i \mathbf{1}_{(t,x)\in[t_i,t_{i+1})\times A_i},$$

and observe that $\mathbf{u}^\epsilon : \mathbb{T} \times \mathcal{X} \to \mathcal{U}$ 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 $U_t^\epsilon := \mathbf{u}^\epsilon(0, x_0)$, up to the stopping time

$$\rho_1 = \min\Big\{\inf\{t : \|h(t, X_t^\epsilon, U_t^\epsilon) - h(t, x_0, U_t^\epsilon)\| > \epsilon/2\},\, T\Big\}.$$

With this definition, we know

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

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

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

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),

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

As X^ϵ has non-explosive paths, we also know that $\rho_n \to T$ as $n \to \infty$, almost surely.

We now compare with (7.3), to observe that

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

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

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

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

Step 4. Describing an optimizer. Under condition (iii), we see from (7.3) 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 7.3.3. 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.

We will return to this in more detail in the coming chapters, where we will study the theory of viscosity solutions to these PDEs.

Remark 7.3.4. The 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 [33, Chapter 4] for details in this direction.

Remark 7.3.5. 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. See Exercise 7.4.13.

Remark 7.3.6. While we have derived all our existence results under the assumption that controls have a bounded impact, and Lipschitz continuity of our dynamics, we could now consider relaxing this condition. In particular, suppose we can find a smooth solution v to the HJB equation, together with a candidate optimal feedback control $\mathbf{u}^*$ which gives the minimum in the Hamiltonian. Fixing a starting point (t, x), and assuming the state process $X^{t,x,\mathbf{u}^*}$ is well defined, the HJB equation will guarantee that the corresponding process $M^{t,x,\mathbf{u}^*}$ in Theorem 7.1.2 is a local martingale (by applying Ito's lemma), and that the processes $M^{t,x,U}$ are local submartingales for all $U \in \mathbb{U}$. Hence, it remains to check v has enough growth bounds that our processes are a true martingale and submartingales respectively. For example, if we restrict the admissible controls to those where $\mathbb{E}\big[\sup_s |M_s^{t,x,U}|\big] < \infty$ (by imposing sufficent integrability and growth conditions on our controls and cost processes), then the optimality of U^* at (t, x) follows whenever $U^* \in \mathbb{U}$.

Example 7.3.7. *Consider the control problem with state dynamics*

$$\mathrm{d}X_t = (a + bU_t)\mathrm{d}t + (c + fU_t)\mathrm{d}W_t$$

where W is a one-dimensional Brownian motion, $a, b \in \mathbb{R}$ and $\mathcal{U} = \mathbb{R}$. Suppose that the costs are given by

$$g(t, x, u) = qx^2 + ru^2 + 2sxu; \qquad \Phi(x) = x^2$$

with $r > 0$, and $q, s \in \mathbb{R}$, and $\Sigma_T > 0$.

Given the quadratic structure of g and Φ, we guess that the value function is of the form $v(t, x) = \Sigma_t x^2 + 2\Psi_t x + \Gamma_t$, for some functions Σ, Ψ, and Γ. It follows that the spatial derivatives are

$$D_x v = 2\Sigma_t x + 2\Psi; \qquad D^2_{xx} v = 2\Sigma_t.$$

Substituting in the Hamiltonian, we have

$$\begin{aligned} &H(t, x, D_x v, D_{xx} v)\\ &\quad = \inf_u \left\{ qx^2 + ru^2 + 2sxu + (2\Sigma_t x + 2\Psi_t)(a + bu) + \frac{(c + fu)^2}{2}(2\Sigma_t) \right\}. \end{aligned}$$

Taking a first order condition (this gives a minimizer assuming $r + f^2\Sigma_t > 0$, which holds when t is close enough to T) we have

$$0 = 2ru + 2sx + b(2\Sigma_t x + 2\Psi_t) + 2f(c + fu)\Sigma_t$$

which simplifies to give the optimal control,

$$u = -\frac{1}{r + f^2\Sigma_t}\Big((s + b\Sigma_t)x + b\Psi_t + cf\Sigma_t\Big) = K_t x + H_t,$$

where

$$K_t = -\frac{s + b\Sigma_t}{r + f^2\Sigma_t}, \quad \textit{and} \quad H_t = -\frac{b\Psi_t + cf\Sigma_t}{r + f^2\Sigma_t}. \tag{7.4}$$

We substitute this back into the HJB equation to get

$$\begin{aligned}
&-\partial_t(\Sigma_t x^2 + 2\Psi_t x + \Gamma_t)\\
&\quad = H(t, x, D_x v, D_{xx} v)\\
&\quad = qx^2 + r(K_t x + H_t)^2 + 2sx(K_t x + H_t)\\
&\qquad + (2\Sigma_t x + 2\Psi_t)(a + b(K_t x + H_t)) + (c + f(K_t x + H_t))^2(\Sigma_t).
\end{aligned}$$

Matching coefficients, we get $\Sigma_T = 1, \Psi_T = \Gamma_T = 0,$ *and*

$$\begin{aligned}
-\partial_t\Sigma_t &= q + rK_t^2 + 2sK_t + 2b\Sigma_t K_t + \Sigma_t f^2 K_t^2,\\
-\partial_t\Psi_t &= rK_t H_t + sH_t + \Sigma_t(a + bH_t) + \Psi_t bK_t + \Sigma_t f K_t(fH_t + c),\\
-\partial_t\Gamma_t &= rH_t^2 + 2\Psi_t(a + bH_t) + (c + fH_t)^2\Sigma_t.
\end{aligned}$$

Figure 7.1 shows how the solution to the system changes as b *changes. The values of the parameters* b *and* f *modulate how much* U_t *controls the drift and the volatility of* X_t*. The remainder of model parameters are* $a = 0.1, c = 0.5, q = 1.0, r = 1.0, s = 0.0,$ $X_0 = 1,$ *and* $T = 1.$

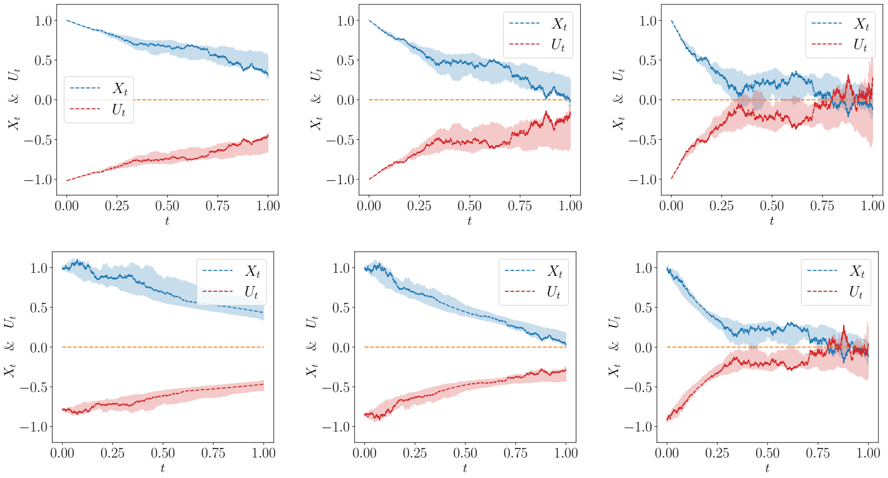


Figure 7.1: State X_t and optimal control U_t, when $b = 1$ (left panels), $b = 2$ (middle panels), and $b = 5$ (right panels). Top panels are for $f = 0.5$ and bottom panels are for $f = 1.0$. Both processes (X and U) are accompanied by the running quantiles (5% and 95%) across time.

The condition $r + f^2\,\Sigma_t > 0$ *is satisfied for* $t \in [0, T]$ *with the parameters we employ.*

7.4 Exercises

Exercise 7.4.1. *Prove the following extension of the martingale verification theorem (Theorem 7.1.2): Let all notation be as given in Theorem 7.1.2, and also assume the stated regularity properties of v and the existence of strong solutions. As before, write*

$$M_{t'}^{t,x,U}(\omega) = \int_t^{t'} g(\omega, s, X_s^{t,x,U}, U_s)\mathrm{d}s + v(\omega, t', X_{t'}^{t,x,U}).$$

Suppose that, for every $(t,x) \in \mathbb{T} \times \mathcal{X}$,

- *for every $U \in \mathbb{U}$, we know $M^{(t,x,U)}$ is a submartingale on $[t,T]$;*
- *for every $\epsilon > 0$, there exists a control $U^{(t,x,\epsilon)} \in \mathbb{U}$ such that the process defined by $(\omega, s) \mapsto M_s^{t,x,U^{(t,x,\epsilon)}}(\omega) - \epsilon s$ is a supermartingale on $[t,T]$.*

Show that, under these conditions, v is a version of the value function.

Exercise 7.4.2. *Consider a stochastic control problem on a fixed horizon where we discount costs at a rate $\rho > 0$.*

(i) Show that the martingale verification principle applies when we consider the process

$$M_{t'}^{\rho,t,x,U}(\omega) = \int_t^{t'} e^{-\rho(s-t)} g(\omega, s, X_s^{t,x,U}, U_s)\mathrm{d}s + v(\omega, t', X_{t'}^{t,x,U}).$$

(ii) Show that the result will also hold over infinite horizons, given v satisfies the transversality condition

$$e^{-\rho t'} v(t', X_{t'}^{t,x,U}) \to 0 \text{ almost surely and in } L^1 \text{ as } t' \to \infty, \text{ for all } U \in \mathbb{U}$$

replaces the terminal value condition for v, and we assume costs are bounded.

Exercise 7.4.3. *By considering the martingale verification theorem in the case $\sigma = 0$, give a direct verification result for deterministic control problems (your final answer should not mention any stochastic results).*

Exercise 7.4.4. *Consider the control problem with one-dimensional state dynamics given by*

$$\mathrm{d}X_t = \frac{1}{3}U_t\mathrm{d}t + \mathrm{d}W_t,$$

where W is a one-dimensional Brownian motion, U_t is a control process taking values in the set $[0,1]$, the horizon is $T = 1$ and the costs are given by

$$g(t,x,u) = \frac{2}{3}ux(1-t) - t + 3x + 1, \qquad \Phi(x) = -x^3.$$

Show that the value function is given by $v(t,x) = -x^2(x+1-t)$, and describe the optimal control.

Exercise 7.4.5. *Consider a controlled diffusion with drift f and volatility σ. For a given feedback control U, we define the linear differential operator*

$$\mathcal{L}^U v = g(t,x,U_t) + f(t,x,U_t)^\top (D_x v) + \frac{1}{2}\mathrm{Tr}\Big[(D^2_{xx}v)(\sigma\sigma^\top)(t,x,U_t)\Big]$$

(i) Assuming all relevant equations admit sufficiently smooth solutions, and all stochastic integrals with respect to martingales are martingales, show that the value of the control $J(\cdot,\cdot,U)$ satisfies the PDE

$$-\partial_t J = \mathcal{L}^U J.$$

(ii) The (classical) comparison principle states that if w and w' are $C^{2,1}$ and satisfy

$$-\partial_t w \geq \mathcal{L}^U w; \qquad -\partial_t w' \leq \mathcal{L}^U w'$$

and $w(T,\cdot) \geq w'(T,\cdot)$, then $w \geq w'$ for all (t,x).

Design a policy iteration scheme to solve the Hamilton–Jacobi–Bellman equation, assuming you are able to solve linear PDEs. For your scheme, assuming the comparison theorem holds, prove the policy improvement lemma.

Exercise 7.4.6. *Consider the Merton problem, where an investor has nonnegative wealth X, invests a fraction U^s of their wealth into a risky asset following a geometric Brownian motion, and consumes a fraction U^c of their wealth per unit time. This means that their wealth satisfies*

$$\mathrm{d}X_t = -U^c X_t \mathrm{d}t + U^s X_t(\mu \mathrm{d}t + \sigma \mathrm{d}W_t)$$

for $\mu, \sigma > 0$, and costs

$$g(t,x,u) = -\frac{(u^c x)^{1-\gamma}}{1-\gamma}; \qquad \Phi(x) = -\frac{x^{1-\gamma}}{1-\gamma}.$$

where $u^c, u^s > 0$ and $\gamma > 0$, $\gamma \neq 1$. Show that the optimal strategy is to invest a constant proportion of wealth U^s in the risky asset, and consume at a deterministic time-dependent rate U^c, determined by the solution to an ODE.

You may wish to use an ansatz of the form $v(t,x) = -w(t)\frac{x^{1-\gamma}}{1-\gamma}$, for some positive function w.

Exercise 7.4.7. *In this question, we consider finite difference approximations of the Hamilton–Jacobi–Bellman equation. We will consider the problem where the state variable is in one dimension, and the Hamiltonian is given by*

$$H(t,x,q,a) = \inf_{u\in[-1,1]} \Big\{uq + a\Big\} = \inf_{u\in[-1,1]} \tilde{H}(t,x,q,a,u).$$

For $N \in \mathbb{N}$, consider a discrete grid $\{0, \delta_t, 2\delta_t, ..., N\delta_t\}$ in time, and

$$\{-N\delta_x, (-N+1)\delta_x, ..., 0, ..., (N-1)\delta_x, N\delta_x\}$$

in space, where $\delta_t = T/N$ *and* $\delta_x = 1/\sqrt{N}$.

We assume the problem stops when we hit the numerical boundary (so $v(t,x) = \Phi(x)$ *for all* t *and all* $x \in \{\pm N\delta_x\}$*). We assign a terminal value* $\Phi(x) = 1 - \frac{x^4}{1+x^4}$.

(i) *If* v *is a* $C^{1,2}$ *function, show that (as* $N \to \infty$*), with* k *chosen such that* $t \in [(k-1)\delta t, k\delta_t]$,

$$\frac{v(k\delta_t, x) - v((k-1)\delta_t, x)}{\delta_t} = \partial_t v(t,x) + o(1)$$

and, with k *chosen such that* $x \in [(k-1)\delta_x, (k+1)\delta_x]$,

$$\frac{v(t,(k+1)\delta_x) - v(t, k\delta_x)}{\delta_x} = D_x v(t, k\delta_x) + o(1)$$

and

$$\frac{v(t,(k+1)\delta_x) - 2v(t,k\delta_x) + v(t,(k-1)\delta_x)}{\delta_x^2} = D^2_{xx} v(t,k\delta_x) + o(1).$$

(ii) *Using this finite difference scheme, write down an approximation of*

$$v_{k-1,j} := v((k-1)\delta_t, j\delta_x)$$

in terms of $v_{k,j}, v_{k,j+1}, v_{k,j-1}$*, when* v *satisfies the HJB equation.*

(iii) *Show that* $v_{k-1,j}$ *is a componentwise monotone increasing function of* $v_{k,\bullet}$ *(where* $v_{k,\bullet}$ *represents the vector* $[v_{k,j}]$ *for* $j = -N, -N+1, ..., N$*), provided* $3\delta_t/\delta_x^2 < 1$*. (This is known as being a* monotone scheme*.) How does this compare with the discrete-time discrete-state control problem?*

(iv) *Now consider the deterministic problem with Hamiltonian*

$$H(t,x,q) = \inf_{u\in[-1,1]} \Big\{ uq \Big\}.$$

Is the basic finite difference numerical scheme for this problem typically monotone?

(v) *For the deterministic problem in the previous part, consider the modified scheme*

$$\hat{H}(t,x,v_{k,\cdot}) = \begin{cases} \frac{v_{k,j+1} - v_{k,j}}{\delta_x} & \text{if } v_{k,j+1} < v_{k,j}, \\ -\frac{v_{k,j} - v_{k,j-1}}{\delta_x} & \text{if } v_{k,j+1} \geq v_{k,j}. \end{cases}$$

Show that using $\hat{H}(v)$ *in the place of* $H(t,x,q)$ *in our numerical approximation gives a monotone scheme, provided* $\delta_t/\delta_x < 1$*. (Approximations of this type are commonly known as upwind schemes, and are important in order to prove numerical stability).*

(vi) *Implement the three numerical schemes considered above, for various choices of* T *and* N*, and observe the behaviour of the numerical approximations.*

Exercise 7.4.8. *Let $T > 0$, $x_0 \in \mathbb{R}$, and let $b \in L^1([0,T];\mathbb{R})$ be a deterministic function. Consider the controlled SDE*

$$\mathrm{d}X_t = b_t \,\mathrm{d}t + U_t \,\mathrm{d}W_t, \qquad X_0 = x_0,$$

where W is a one–dimensional Brownian motion on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ satisfying the usual conditions. Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable convex function (i.e. $\Phi'' \geq 0$) with polynomial growth. Consider the Mayer control problem of minimizing

$$J(t, x, U) := \mathbb{E}\big[\Phi(X_T^{t,x,U})\big]$$

over $U \in \mathbb{U}$.

(i) Show, using Jensen's inequality, that if $0 \in \mathcal{U}$ then the control $U_t = 0$ is an optimal control, and is unique up to equality $\mathrm{d}\mathbb{P} \times \mathrm{d}t$-almost everywhere.

(ii) By applying Itô's lemma to compute the cost of a policy U, show that for any $\mathcal{U}$, any admissible control minimizing the accumulated variance $\int_0^T U_s^2 \,\mathrm{d}s$ almost surely is optimal.

Hint: treat the term depending on b as a deterministic constant

Exercise 7.4.9 (Cole–Hopf transformation)**.** *Let W be a one–dimensional Brownian motion and consider the controlled SDE*

$$\mathrm{d}X_t = U_t \,\mathrm{d}t + \sigma \,\mathrm{d}W_t, \qquad X_0 = x,$$

where $\sigma > 0$ is constant and U is progressively measurable. The objective is to minimize

$$J(U) = \mathbb{E}\left[\int_0^T \tfrac{1}{2} U_t^2 \,\mathrm{d}t + e^{\beta X_T}\right],$$

where $\beta > 0$ is a constant. Assume throughout that the value function v is sufficiently smooth, and that admissible controls are restricted to guarantee uniform integrability of costs.

(i) For an arbitrary admissible control U, define

$$M_t^U := v(t, X_t) + \int_0^t \tfrac{1}{2} U_s^2 \,\mathrm{d}s.$$

a. Using Itô's formula, write down the drift of M_t^U in terms of the derivatives of v.

b. By minimizing the drift pointwise in U, show formally that the optimal control is given by

$$U_t^* = -D_x v(t, X_t),$$

and deduce that the value function should satisfy the Hamilton–Jacobi–Bellman equation

$$-\partial_t v(t,x) = -\tfrac{1}{2}(D_x v(t,x))^2 + \tfrac{1}{2}\sigma^2 D_{xx}^2 v(t,x), \qquad v(T,x) = e^{\beta x}.$$

(ii) Consider the Cole–Hopf transformation

$$\phi(\tau, x) = \exp\Big(\frac{-v(T-\tau, x)}{\sigma^2}\Big).$$

Show that ϕ satisfies the linear heat equation

$$\partial_\tau \phi(\tau, x) = \tfrac{\sigma^2}{2} D^2_{xx}\phi(\tau, x), \qquad \phi(0, x) = \exp\big(-e^{\beta x}/\sigma^2\big).$$

(iii) Show that ϕ has the Green's function representation

$$\phi(\tau, x) = \int_{\mathbb{R}} G_{\sigma^2\tau}(x-y)\exp\big(-e^{\beta y}/\sigma^2\big)\,\mathrm{d}y.$$

where $G_s(z) = \frac{1}{\sqrt{2\pi s}}\exp\Big(-\frac{z^2}{2s}\Big)$. Hence write down the value function in closed form.

(iv) Show that the optimal control has the feedback form

$$U_t^* = \sigma^2\partial_x \log\phi(T-t, X_t).$$

Exercise 7.4.10. *Let $T > 0$. Define the annular domain*

$$\mathcal{D} := \{x \in \mathbb{R}^2 : 1 < \|x\| < 2\}.$$

Let W be a two-dimensional Brownian motion. Consider the controlled diffusion

$$\mathrm{d}X_t = b(t, X_t, U_t)\,\mathrm{d}t + \sigma(t, X_t, U_t)\,\mathrm{d}W_t, \qquad X_0 = x \in R^2,$$

where U_t takes values in a compact set $\mathcal{U}$, and the coefficients satisfy the standing regularity assumptions of the chapter. Define the exit time

$$\tau^U := \inf\{t \geq 0 : X_t^U \notin \mathcal{D}\},$$

and consider the objective of maximizing $\mathbb{P}(\tau^U > T)$.

(i) Show that the desired probability can be expressed in terms of a standard minimization control problem with $b = \sigma = 0$ for $x \notin \mathcal{D}$ and terminal value $\Phi(x) = -\mathbf{1}_{x\in\mathcal{D}}$).

(ii) Write down a Hamilton–Jacobi–Bellman equation describing the value function for this equivalent control problem. Explain why this can be seen either on the whole space $x \in \mathbb{R}^2$, or by considering only the behaviour on $\mathcal{D}$ together with an additional boundary condition.

(iii) Consider the special case where

$$\mathrm{d}X_t = -\frac{U_t}{2\|X_t\|^2}X_t\mathrm{d}t + \sqrt{U_t}\,\mathrm{d}W_t, \qquad U_t \in \mathcal{U} = [1, 2].$$

(a) *Write down the dynamics of* $Y_t = \|X_t\|$ *and hence give a one-dimensional problem equivalent to the control of* X *considered previously.*

(b) *Identify the optimal control, and hence show that the optimal probability of not hitting the boundary has Fourier series expansion*

$$\sum_{k=0}^{\infty} \frac{4}{(2k+1)\pi} \sin\Big((2k+1)(\|X_0\|-1)\pi\Big) \exp\Big(-\frac{\pi^2}{2}(2k+1)^2 T\Big).$$

Exercise 7.4.11. *Let* $T > 0$ *and fix parameters* $a, b, r > 0$ *and a control bound* $U_{\max} > 0$*. Consider the one–dimensional controlled diffusion on* $[0, T]$,

$$\mathrm{d}X_s = f(U_s)\,\mathrm{d}s + \sigma(U_s)\,\mathrm{d}W_s, \qquad X_t = x,$$

where the control U_s *takes values in the compact set* $\mathcal{U} := [0, U_{\max}]$*, the drift and volatility are given by*

$$f(u) = a\big(1 - e^{-u}\big) \quad (\textit{concave in } u), \qquad \sigma(u) = b\,u \quad (\textit{increasing in } u),$$

and W *is a one–dimensional Brownian motion.*

Define the cost functional with a quadratic control penalty

$$J(t, x; U) := \mathbb{E}\left[\int_t^T \Big(g(X_s^{t,x,U}) + \tfrac{r}{2}U_s^2\Big)\,\mathrm{d}s + \Phi(X_T^{t,x,U})\right],$$

and the value function

$$v(t, x) := \inf_{U \in \mathbb{U}} J(t, x; U).$$

Take the running and terminal state costs to be

$$g(x) := \frac{1}{1+x} \qquad \textit{and} \qquad \Phi(x) := \frac{1}{1+x},$$

both defined for $x > -1$ *(so that derivatives below are well-defined). Note that* g *and* Φ *are* decreasing *and* convex *functions on* $(-1, \infty)$.

(i) *Briefly explain why for any admissible* U *the SDE above admits a unique strong solution and* $J(t, x; U)$ *is finite for* $x > -1$.

(ii) a. *Fix an admissible control* U*. Show that the map* $x \mapsto J(t, x; U)$ *is* decreasing *and* convex *on* $x > -1$*. (Hint: use the affine dependence of* $X_s^{t,x,U}$ *on the initial condition:* $X_s^{t,x,U} = x + M_s$ *where* M_s *is independent of* x*; then use that expectation preserves convexity and monotonicity.)*

b. *Deduce that the value function* $v(t, x) = \inf_U J(t, x; U)$ *is also decreasing and convex in* x.

(iii) *Assume v is sufficiently smooth and write the (unoptimized) Hamiltonian*

$$\tilde{H}(t,x,p,q) := g(x) + p\, f(u) + \tfrac{1}{2} q\, \sigma(u)^2 + \tfrac{r}{2} u^2,$$

where $p = D_x v(t,x)$ and $q = D^2_{xx} v(t,x)$. Show that $\tilde{H}$ is strictly convex in u for every (t,x), given that $r > 0$. (Recall from part (ii) that $p \le 0$ and $q \ge 0$.) Conclude that for each (t,x) there is a unique *minimizer $u^*(t,x) \in \mathcal{U}$.*

(iv) a. *Show that the first–order condition for the minimizer can be written as*

$$(r + qb^2)u + pf'(u) = 0.$$

b. *Explain why, since $f'(u) = ae^{-u} > 0$ and $p \le 0$, this equation admits a unique solution $u \in (0, U_{\max})$ when $U_{\max}$ is large enough; and why, when the solution lies outside $[0, U_{\max}]$, the minimizer is the projected value onto the interval $\mathcal{U}$.*

(v) *Substitute the closed form of u^* into the HJB equation to show that*

$$-\partial_t v(t,x) = g(x) + \tfrac{1}{2} qb^2 (u^*)^2 + pf(u^*) + \tfrac{r}{2}(u^*)^2,$$

where $p = D_x v$, $q = D^2_{xx} v$ and $u^* = \mathbf{u}^*(t,x)$ is the unique minimizer determined implicitly above.

(vi) *Suppose a classical solution $v \in C^{1,2}([0,T] \times (-1,\infty))$ exists with the properties $D_x v \le 0$ and $D^2_{xx} v \ge 0$. Define*

$$M^U_s := v(s, X^{t,x,U}_s) + \int_t^s \Big(g(X^{t,x,U}_r) + \tfrac{r}{2} U_r^2 \Big)\, \mathrm{d}r.$$

Use Itô's formula to show that for any U,

$$\mathrm{d}M^U_s = \big(\textit{nonnegative drift}\big)\, \mathrm{d}s + \textit{martingale term},$$

and that the drift vanishes if and only if $U_s = U^*_s := \mathbf{u}^*(s, X^{t,x,U}_s)$ a.s.

Conclude that M^U is a submartingale and, assuming $\mathbf{u}^*$ is sufficiently regular that M^{U^*} is well defined, U^*_s is optimal.

Exercise 7.4.12. *Let W be a one–dimensional Brownian motion W. Fix $\sigma > 0$ and consider the controlled diffusion*

$$\mathrm{d}X_t = U_t\, \mathrm{d}t + \sigma\, \mathrm{d}W_t, \qquad X_0 = x \in (-1,1),$$

where the control takes values in the discrete set $U_t \in \{-1, 1\}$. Let

$$\tau := \inf\{t \ge 0 : \ |X_t| = 1\}$$

be the exit time from the interval $(-1,1)$. Consider the problem of minimizing the expected exit time, that is, we take

$$v^\sigma(x) := \inf_U \mathbb{E}_x[\tau].$$

You may assume that v is a Lipschitz continuous function.

(i) *By considering an appropriate control problem with* $\Phi = v^\sigma$ *over some horizon, show that* v^σ *should satisfy*

$$\inf_{u\in\{-1,1\}} \Big\{ 1 + u\, D_x v^\sigma(x) + \frac{\sigma^2}{2} D^2_{xx} v^\sigma(x) \Big\} = 0, \qquad x \in (-1,1),$$

with boundary conditions $v^\sigma(\pm 1) = 0$.

(ii) *Show that, at any point where* v^σ *is smooth, the minimizing feedback control is given by*

$$\mathbf{u}^*(x) = -\operatorname{sign}\big(D_x v^\sigma(x)\big).$$

(iii) *Using symmetry, argue that* v^σ *is even, and why we expect it to be concave, and hence the optimal feedback control should be* $\mathbf{u}^*(x) = \operatorname{sign}(x)$.

(iv) *Supposing that* v^σ *is also differentiable at* $x = 0$, *show that* v^σ *should also be represented by the linear ODE*

$$1 + D_x v^\sigma(x) + \frac{\sigma^2}{2} D^2_{xx} v^\sigma(x) = 0, \qquad x \in (0,1),$$

with the boundary conditions $v^\sigma(1) = 0$, $D_x v^\sigma(0) = 0$. *Hence derive the explicit formula*

$$v^\sigma(x) = \; 1 + \frac{\sigma^2}{2} e^{-2/\sigma^2} - \frac{\sigma^2}{2} e^{-2|x|/\sigma^2} - |x|.$$

(v) *Verify directly that the function above is* C^2 *on* $(-1,1)$ *and satisfies the Hamilton–Jacobi–Bellman equation.*

(vi) *Compute the pointwise limit of* $v^\sigma(x)$ *as* $\sigma \to 0$ *for* $x \in (-1,1)$. *Show that*

$$\lim_{\sigma \downarrow 0} v^\sigma(x) = 1 - |x|.$$

and compare with the result of Example 6.1.1.

Exercise 7.4.13. *Consider a stochastic control problem satisfying Assumption 7.3.1, where the control domain is a convex compact subset* $\mathcal{U} \subset \mathbb{R}^m$. *Let* $v \in C^{1,2}([0,T] \times \mathbb{R}^d)$ *be a classical solution of the HJB equation on* $[0,T) \times \mathbb{R}^d$. *Suppose that*

- *For each fixed* (t,x,p,q), *the map* $u \mapsto \tilde{H}(t,x,p,q;u)$ *is* C^2 *in* u.
- *(Uniform strong convexity) There exists* $\lambda > 0$ *such that for all* (t,x,p,q) *and all* $u \in \mathbb{R}^m$,
$$D^2_{uu} \tilde{H}(t,x,p,q;u) \geq \lambda I_m;$$
- *For all* (t,x,p,q), $\tilde{H}(t,x,p,q;u) \to \infty$ *as* $u \to \partial\mathcal{U}$;
- *The maps* $(t,x,p,q,u) \mapsto \partial_u \tilde{H}(t,x,p,q;u)$ *and* $(t,x,p,q,u) \mapsto D^2_{uu}\tilde{H}(t,x,p,q;u)$ *are locally Lipschitz.*

(i) *Show that for each (t, x) there exists a unique minimizer $u^*(t, x) \in \mathbb{R}^m$ of $u \mapsto \tilde{H}(t, x, D_x w(t, x), D^2_{xx} w(t, x); u)$. which satisfies the first–order condition*

$$\partial_u \tilde{H}\big(t, x, D_x w(t, x), D^2_{xx} w(t, x); u^*(t, x)\big) = 0.$$

(ii) *Using the implicit function theorem, or otherwise, show that the mapping*

$$(t, x) \mapsto u^*(t, x)$$

is locally Lipschitz.

(iii) *Assume that $f(t, x, u)$ and $\sigma(t, x, u)$ are globally Lipschitz in x uniformly in u. Show that the closed-loop SDE*

$$\mathrm{d}X_t = f\big(t, X_t, u^*(t, X_t)\big)\,\mathrm{d}t + \sigma\big(t, X_t, u^*(t, X_t)\big)\,\mathrm{d}W_t$$

admits a unique strong solution, and conclude that the feedback control

$$U_t = u^*(t, X_t)$$

is admissible.

Exercise 7.4.14 (Polynomial closure and linear–quadratic structure)**.** *Consider the controlled diffusion*

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

with value function $v(t, x)$ solving the Hamilton–Jacobi–Bellman equation

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

Assume all coefficients are smooth and polynomial in (x, u).

(i) *Suppose that for some fixed integer $n \geq 2$ the following 'polynomial closure' property holds:*

For every polynomial terminal cost $\Phi(x)$ of degree $\leq n$, the corresponding value function $v(t, x)$ remains a polynomial in x of degree $\leq n$ for all $t \in [0, T]$.

a. *Show that $D_x v$ has degree $\leq n - 1$ and $D^2_{xx} v$ has degree $\leq n - 2$.*

b. *Deduce that the HJB operator must map polynomials of degree $\leq n$ into polynomials of degree $\leq n$.*

(ii) *Assume $f(t, x, u)$ is polynomial in x. Show that if f contains terms nonlinear in x (degree ≥ 2), then $D_x v^\top f$ generically produces monomials of degree $\geq n + 1$. Deduce that polynomial closure forces*

$$f(t, x, u) = A(t)x + B(t, u),$$

that is, the drift must be at most linear in x.

(iii) *Assume $\sigma\sigma^\top(t,x,u)$ is polynomial in x. Show that if it contains terms of degree ≥ 2 in x, then*

$$\mathrm{Tr}\big(D^2_{xx}v\,\sigma\sigma^\top\big)$$

generically produces degree $\geq n+1$. Deduce that polynomial closure forces

$$\sigma\sigma^\top(t,x,u) \text{ to be at most quadratic in } x.$$

(iv) *Assume the Hamiltonian is polynomial in u. Show that if $g(t,x,u)$ is not quadratic in u, then either:*

- *the minimizer $u^*(t,x)$ is non-polynomial in D_xv, or*
- *substituting u^* into the Hamiltonian raises the degree in x beyond n.*

Deduce that polynomial closure forces

$$g(t,x,u) = \bar{g}(t,x) + u^\top h(t,x) + \tfrac{1}{2}u^\top R(t)u, \qquad R(t) \succ 0,$$

that is, the running cost must be quadratic in the control.

(v) *Conclude that if polynomial closure holds for arbitrary degree n, then the control problem must be of linear–quadratic type:*

$$\begin{aligned} f(t,x,u) &= A(t)x + B(t)u + a(t),\\ \sigma(t,x,u) &= \Sigma(t)x + \Gamma(t)u + \sigma_0(t),\\ g(t,x,u) &= \tfrac{1}{2}x^\top Q(t)x + x^\top S(t)u + \tfrac{1}{2}u^\top R(t)u + \ell(t)^\top x. \end{aligned}$$

In particular, polynomial families are closed under the HJB operator only in the linear–quadratic setting.

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, 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 [12].

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.

[SC]$_7$:some of this is too basic. I say we cut everything which is explained in Capinski and Kopp, and only highlight the slightly more unusual bits.

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.

Remark A.1.5. 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.1.6. *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.1.7 (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.1.24 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.1.8. *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

$$\int_{\mathcal{S}} \phi(x)\mu(\mathrm{d}x) = \sum_{n\in\mathbb{N}} a_n \mu(A_n)\,.$$

[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.

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.1.9. *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 $\|f\|_p < \infty$.*

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.1.19).

Definition A.1.10. *Let $(S,\mathcal{F})$ be a measurable space. Let μ,ν be measures on $(S,\mathcal{F})$. The measure ν is said to be absolutely continuous with respect to μ, written as $\nu \ll \mu$, if $\mu(A) = 0$ implies $\nu(A) = 0$ for all $A \in \mathcal{F}$. If $\nu \ll \mu$ and $\mu \ll \nu$, then μ and ν are equivalent measures.*

The terminology comes from the fact that $\nu \ll \mu$ if and only if for all $\epsilon > 0$ there exists $\delta > 0$ such that $\nu(A) < \epsilon$ whenever $\mu(A) < \delta$ $(A \in \mathcal{F})$.

Theorem A.1.11 (Radon–Nikodym Theorem)**.** *Let μ and ν be two σ-finite measures on the measurable space $(S,\mathcal{F})$. If $\nu \ll \mu$, there exists a unique (up to equality μ-a.e.) non-negative measurable function $f : S \to [0,\infty]$ such that*

$$\nu(A) = \int_A f\,\mathrm{d}\mu\,,$$

for all $A \in \mathcal{F}$. This f is called the Radon-Nikodym derivative of ν with respect to μ and it is often written as

$$f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,.$$

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.1.12. *Take a random variable $X \in L^1$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given a sub-$\sigma$-algebra $\mathcal{G}$ of $\mathcal{F}$, we define the conditional expectation $Y = \mathcal{E}[X|\mathcal{G}]$ to be the unique $(\Omega, \mathcal{G})$ random variable (up to equality almost everywhere) 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.1.13 (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.

$[\mathrm{SC}]_8$:Maybe keep this

Lemma A.1.14 (Bayes' rule)**.** *Let $\mathbb{Q}$ be equivalent to $\mathbb{P}$. Let the Radon–Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$ be $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$. It follows that for all $\mathbb{Q}$-integrable Z,*

$$\mathbb{E}^{\mathbb{Q}}[Z|\mathcal{F}_t] = \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} Z \middle| \mathcal{F}_t\right] \Big/ \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \middle| \mathcal{F}_t\right].$$

Proof. We need to show that for any $A \in \mathcal{F}_t$, the $\mathcal{F}_t$ random variable

$$Y_t := \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} Z \middle| \mathcal{F}_t\right] \Big/ L_t\,, \qquad L_t := \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \middle| \mathcal{F}_t\right],$$

satisfies that

$$\mathbb{E}^{\mathbb{Q}}[1_A\, Y_t] = \mathbb{E}^{\mathbb{Q}}[1_A\, Z]\,.$$

The proof follows from the following equalities

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}[1_A\, Z] &= \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} 1_A\, Z\right] = \mathbb{E}^{\mathbb{P}}\left[\mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} 1_A\, Z \middle| \mathcal{F}_t\right]\right] \\
&= \mathbb{E}^{\mathbb{P}}\left[1_A\, \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} Z \middle| \mathcal{F}_t\right]\right] = \mathbb{E}^{\mathbb{Q}}\left[1_A\, L_t^{-1}\, \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} Z \middle| \mathcal{F}_t\right]\right] \\
&= \mathbb{E}^{\mathbb{Q}}[1_A\, Y_t]\,.
\end{aligned}$$

In the above, we used that for any $\mathcal{F}_t$ random variable X we have that

$$\mathbb{E}^{\mathbb{Q}}[X] = \mathbb{E}^{\mathbb{P}}\left[X\,\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] = \mathbb{E}^{\mathbb{P}}\left[\mathbb{E}^{\mathbb{P}}\left[X\,\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\middle|\mathcal{F}_t\right]\right] = \mathbb{E}^{\mathbb{P}}\left[X\,\mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\middle|\mathcal{F}_t\right]\right] = \mathbb{E}^{\mathbb{P}}\big[X\,L_t\big].$$

See Proposition 1.7.1.1 in [30] for further details about L_t. □

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

Lemma A.1.15 (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*

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

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

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

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.1.16. *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 [16, Theorem 2.5.4] for a proof.

Theorem A.1.17. *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 $\epsilon > 0$ there is a $\delta > 0$ such that, for all $A \in \mathcal{F}$ with $\mathbb{P}(A) \leq \delta$, we have $\mathbb{E}\big[1_A|X(\omega)|\big] < \epsilon$ for all $X \in K$.

Lemma A.1.18 (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 $\sup_{X\in K} \mathbb{E}\big[\phi(|X|)\big] < \infty$. (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 $\epsilon > 0$. Put $a = \epsilon^{-1}\lambda$ 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

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

Taking $c \to \infty$ we see that

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

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 [16, Theorem 2.5.8] for a proof.)

Theorem A.1.19 (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.1.20. *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.1.17. 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 $\epsilon > 0$, we can find a $\delta > 0$ such that $\mathbb{E}[1_{A_\delta(\mathcal{F})}|X|] < \epsilon$, and hence, for any $A \in \mathcal{F}$ with $\mathbb{P}(A) \leq \delta$,

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

for all $X \in K, \mathcal{G} \in \mathfrak{G}$. By Theorem A.1.17, 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.1.21. *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.1.22. *We say an process X is a* submartingale *if, for all $t \le t'$, we have $X_t \le \mathbb{E}[X_{t'}|\mathcal{F}_t]$, 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.

Kolmogorov continuity theorem (for random fields, as in Rogers+Williams)

[SC]$_9$:need to decide which of these things is important to include

[SC]$_{10}$:not sure we use this in the end

Lemma A.1.23. *Let Y be an adapted right-continuous process. Then Y is a submartingale if and only if, for all stopping times $\tau \le \tau'$, we know $\mathbb{E}[Y_\tau] \le \mathbb{E}[Y_{\tau'}]$. 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

$$A = \{\mathbb{E}[Y_{t'}|\mathcal{F}_t] < Y_t\} \in \mathcal{F}_t.$$

Then $\tau = 1_A t + 1_{A^c} t' \le t'$ is a stopping time, and $Y_{t'} - Y_\tau = 1_A(Y_{t'} - Y_t)$. Therefore,

$$0 \le \mathbb{E}[Y_{t'} - Y_\tau] = \mathbb{E}[1_A(Y_{t'} - Y_t)] = \mathbb{E}[\underbrace{1_A(\mathbb{E}[Y_{t'}|\mathcal{F}_t] - Y_t)}_{\le 0}] \le 0.$$

Therefore, A is a null set and

$$Y_t \le \mathbb{E}[Y_{t'}|\mathcal{F}_t]$$

that is, Y is a submartingale. The martingale statement follows by considering Y and $-Y$. □

A.1A Existence of essential suprema

Theorem A.1.24. *[[16], 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*

[SC]$_{11}$:keep this

(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.1B Almost supermartingales and stochastic approximation

Here we prove some useful results related to stochastic approximation theory.

Definition A.1.25. *An adapted process X is called an* almost supermartingale *if there exist nonnegative adapted integrable processes A, B, C such that*

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] \leq X_t(1 + B_t) + A_t - C_t.$$

Theorem A.1.26 (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 $\big\{\omega : \sum_t(A_t + B_t) < \infty\big\}$.*

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 $\beta > 0$,

$$U_t = X'_t - \sum_{n=1}^{t-1}(A'_n - C'_n), \qquad \tau = \min\Big\{t \in \mathbb{N} : \sum_{n\leq t} A'_t > \beta\Big\}.$$

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 $0 \leq Z' \leq Z$ for $Z \in \{X, A, C\}$, we see that U is integrable. We also know that, as X is an almost supermartingale,

$$\begin{aligned}\mathbb{E}[X'_{t+1}|\mathcal{F}_t] &= \mathbb{E}[X_{t+1}|\mathcal{F}_t] \prod_{n=1}^{t}(1 + B_n)^{-1}\\ &\leq \Big(X_n(1 + B_t) + A_t - C_t\Big) \prod_{n=1}^{t}(1 + B_n)^{-1} = X'_t + A'_t - C'_t\end{aligned}$$

and hence

$$\begin{aligned}
&\mathbb{E}[U^\tau_{t+1}|\mathcal{F}_t]\\
&= U_\tau 1_{t\geq\tau} + \mathbb{E}[U_{t+1}|\mathcal{F}_t]1_{t<\tau} = U_\tau 1_{t\geq\tau} + \mathbb{E}\Big[X'_{t+1} - \sum_{n=1}^{t}(A_n - C_n)\Big|\mathcal{F}_t\Big]1_{t<\tau}\\
&= U_\tau 1_{t\geq\tau} + \Big(\mathbb{E}[X'_{t+1} - A_t + C_t|\mathcal{F}_t] - \sum_{n=1}^{t}(A_n - C_n)\Big)1_{t<\tau}\\
&\leq U_\tau 1_{t\geq\tau} + \Big(X'_t - \sum_{n=1}^{t}(A_n - C_n)\Big)1_{t<\tau} = U_\tau 1_{t\geq\tau} + U_t 1_{t<\tau}\\
&= U^\tau_t
\end{aligned}$$

It follows that U^τ is a supermartingale. We also know that X', A', C' are nonnegative, so

$$U^\tau_t \geq U^\tau_t - \sum_{t\leq\tau-1} C'_t = X'_t - \sum_{t\leq\tau-1} A'_t \geq -\beta.$$

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.1.27. *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,

(i) $\mathbb{E}[\varepsilon_{t+1}(i)|\mathcal{F}_t] = 0$

(ii) $\mathbb{V}[\varepsilon_t(i)|\mathcal{F}_t] \leq c(1 + \|Y_t\|_\infty^2)$ *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,

$$\begin{aligned}\mathbb{E}[(Y_{t+1}^\tau)^2|\mathcal{F}_t] &\leq (1-\alpha_t)^2\beta_t^2(Y_t^\tau)^2 + 2(1-\alpha_t)\beta_t|(Y_t^\tau)|\alpha_t\mathbb{E}[\varepsilon_{t+1}|\mathcal{F}_t] \\ &\quad + \alpha_t^2\mathbb{V}[\varepsilon_{t+1}^2|\mathcal{F}_t] + \alpha_t^2\mathbb{E}[\varepsilon_{t+1}|\mathcal{F}_t]^2 \\ &\leq (1-\alpha_t)^2(Y_t^\tau)^2 + \alpha_t^2\mathbb{V}[\varepsilon_{t+1}^2|\mathcal{F}_t] \\ &\leq (1-\alpha_t)^2(Y_t^\tau)^2 + \alpha_t^2 c(1 + \|(Y_t^\tau)\|_\infty^2) \\ &\leq (1+\alpha_t^2)(Y_t^\tau)^2 + \alpha_t^2 c(1+k^2) - 2\alpha_t(Y_t^\tau)^2.\end{aligned}$$

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 [29], whose proof is a variant of the approach below) is useful.

Lemma A.1.28 (cf. Lemma 5.1.3). *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,

(i) $\alpha_t(i) \in [0,1]$, $\sum_{t\in\mathbb{N}} \alpha_t(i) = \infty$, $\sum_{t\in\mathbb{N}} \alpha_t^2(i) < \infty$,

(ii) $|\mathbb{E}[Z_{t+1}(i)|\mathcal{F}_t]| \leq \gamma \|Y_t\|_\infty$, *with* $\gamma < 1$,

(iii) $\mathbb{V}[Z_t(i)|\mathcal{F}_t] \leq c(1 + \|Y_t\|_\infty^2)$ *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.1.27 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 $\mathbb{E}[Z_{t+1}|\mathcal{F}_t]$)

$$\begin{aligned}|\Delta_{t+1}(i)| &\leq (1-\alpha_t(i))\beta_t|\Delta_t(i)| + \alpha_t(i)\gamma\|\beta_t U_t\|_\infty\\ &\leq (1-\alpha_t(i))|\Delta_t(i)| + \alpha_t(i)\beta_t\gamma(\|\Delta_t\|_\infty + \|\Gamma_t\|_\infty).\end{aligned}$$

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 $\beta_t \leq 1$,

$$\begin{aligned}|\Delta_{t+1}(i)| - \gamma\eta k &\leq (1-\alpha_t(i))\Big(|\Delta_t(i)| - \gamma\eta k\Big)\\ &\leq \Big(|\Delta_T(i)| - \gamma\eta k\Big)\prod_{s=T}^{t}(1-\alpha_t(i)) \leq (k-\gamma\eta k)\exp\Big(-\sum_{s=T}^{t}\alpha_t(i)\Big).\end{aligned} \tag{A.1}$$

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 (A.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

[SC]$_{12}$:need to decide which of these to include in detail, and which to state the result.

A.2A Lipschitz SDEs

This appendix is taken, with slight modification, from [16, Chapter 16]. Consider an SDE of the form

[SC]$_{13}$:I think it makes sense to keep a good amount of this, along with a statement of Ito and BDG but no proof

$$X_t = X_0 + \int_{[0,t]} \mu(\omega, s, X_s)\mathrm{d}s + \int_{[0,t]} \sigma(\omega, s, X_s)\mathrm{d}W_s, \tag{A.2}$$

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 $\|\mu(t,x) - \mu(t,x')\| \leq K\|x - x'\|$. Suppose*

$$\int_{[0,T]} \mathbb{E}\big[\|\mu_s(0)\|^2 + \|\sigma_s(0)\|^2\big]\mathrm{d}s < \infty$$

Then (A.2) *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* (A.2) *with μ, σ functions satisfying the linear growth condition*

$$\|\mu_s(x)\| \leq \tilde{\mu}_s + K\|x\|, \qquad \|\sigma_s(x)\| \leq \tilde{\sigma}_s + K\|x\|,$$

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 $p \geq 2$, there exists a real constant C depending on T, K and p such that

$$\mathbb{E}\Big[\sup_{t\leq s\leq T} \|X_s\|^p \Big| \mathcal{F}_t\Big] < C\Big(\|X_t\|^p + \int_{[t,T]} \mathbb{E}\big[\|\tilde{\mu}_s\|^p + \|\tilde{\sigma}_s\|^p \big|\mathcal{F}_t\big]\, ds\Big).$$

Proof. Continuity of X follows immediately from the continuity of the integrals in (A.2). 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 $t \leq t' \leq T$,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{t\leq t'\leq T} \|X_{t'}\|^p \Big| \mathcal{F}_t\Big] \\
&= \mathbb{E}\Big[\sup_{r\in[t,t']} \Big\|X_t + \int_{[t,r]} \mu_s(X_s)\mathrm{d}s + \int_{[t,r]} \sigma_s(X_s)\mathrm{d}W_s\Big\|^p \Big|\mathcal{F}_t\Big] \\
&\leq C\|X_t\|^p + C\int_{[t,t']} \mathbb{E}\big[\|\mu_s(X_s)\|^p \big|\mathcal{F}_t\big]\mathrm{d}s + C\mathbb{E}\Big[\Big(\sup_{r\in[t,t']}\Big|\int_{[t,r]} \sigma_s(X_s)\mathrm{d}W_s\Big|\Big)^p\Big|\mathcal{F}_t\Big] \\
&\leq C\|X_t\|^p + C\int_{[t,t']} \mathbb{E}\big[\|\mu_s(X_s)\|^p \big|\mathcal{F}_t\big]\mathrm{d}s + C\mathbb{E}\Big[\Big(\int_{[t,t']} \|\sigma_s(X_s)\|^2]\mathrm{d}s\Big)^{p/2}\Big|\mathcal{F}_t\Big] \\
&\leq C\|X_t\|^p + C\int_{[t,t']} \mathbb{E}\big[\|\tilde{\mu}_s\|^p + K^p\|X_s\|^p\big|\mathcal{F}_t]ds + C\int_{[t,t']} \mathbb{E}\big[\|\tilde{\sigma}\|^p + K^p\|X_s\|^p\big|\mathcal{F}_t\big]\mathrm{d}s \\
&\leq C\Big(\|X_t\|^p + \int_{[t,t']} \mathbb{E}\big[\|\tilde{\mu}_s\|^p + \|\tilde{\sigma}\|^p\big|\mathcal{F}_t\big]\mathrm{d}s\Big) + C\int_{[t,t']} \mathbb{E}\big[\sup_{s\leq t}\|X_s\|^p\big|\mathcal{F}_t\big]\mathrm{d}s
\end{aligned}$$

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

$$\sup_{t'\in[t,T]} \|X_{t'}\|^p \leq C\Big(\|X_t\|^p + \int_{[t,T]} \mathbb{E}\big[\|\tilde{\mu}_s\|^p + \|\tilde{\sigma}\|^p\big|\mathcal{F}_t\big]\mathrm{d}s\Big)e^{CT} < \infty.$$

[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

Replacing C by Ce^{CT} gives the result. □ □

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 (A.2) with coefficients (μ, σ) and $(\tilde{\mu}, \tilde{\sigma})$ respectively. For any $\beta \geq 0$,*

$$\begin{aligned}
&\mathbb{E}\big[e^{-\beta T}\|X_T - \tilde{X}_T\|^2\big|\mathcal{F}_t\big]\\
&\leq e^{-(\beta-1-4K^2)(T-t)}\Big(\|X_t - \tilde{X}_t\|^2 + \int_{[t,T]} 2e^{-\beta s}\mathbb{E}[\|\mu_s(X_s) - \tilde{\mu}_s(X_s)\|^2\big|\mathcal{F}_t]\\
&\qquad\qquad + 2e^{-2\beta s}\mathbb{E}[\|\sigma_s(X_s) - \tilde{\sigma}_s(X_s)\|^2\big|\mathcal{F}_t]\mathrm{d}s\Big).
\end{aligned}$$

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

$$\begin{aligned}
Y_T = \|X_t - \tilde{X}_t\|^2 &- \beta\int_{[t,T]} e^{-\beta s}\|X_s - \tilde{X}_s\|^2\mathrm{d}s\\
&+ 2\int_{[t,T]} e^{-\beta s}(X_s - \tilde{X}_s)^\top(\mu_s(X_s) - \tilde{\mu}_s(\tilde{X}_s))\mathrm{d}s\\
&+ 2\int_{[t,T]} e^{-\beta s}(X_s - \tilde{X}_s)^\top(\sigma_s(X_s) - \tilde{\sigma}_s(\tilde{X}_s))\mathrm{d}W_s\\
&+ \int_{[t,T]} e^{-2\beta s}\|\sigma_s(X_s) - \tilde{\sigma}_s(\tilde{X}_s)\|^2\mathrm{d}s.
\end{aligned} \tag{A.3}$$

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 'dW' term in (A.3) 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 (A.3), 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 (A.2), 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 (A.2), up to equality in this norm. By continuity of the integrals, $F(X)$ is continuous, which implies the solution satisfies (A.2), 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 $\tau_{n+1} - \tau_n > \epsilon\mathbf{1}_{\{\tau_{n+1}<T\}}$. The result follows. □

A.3 Properties of Viscosity solutions

In this appendix, we collect additional proofs and results which appear in the theory of viscosity solutions.

Proof of Proposition **??**. We restrict ourselves to the superjet $\mathrm{J}^{2,+}w^*(t_0,x_0)$; the other case follows analogously. By definition of the superjet, we already know that

$$\big(\partial_t\varphi(t_0,x_0), D_x\varphi(t_0,x_0), D^2_{xx}\varphi(t_0,x_0)\big) \in \mathrm{J}^{2,+}w^*(t_0,x_0)$$

whenever $\varphi \in C^{1,2}([0,T) \times \mathcal{X})$ satisfies $\varphi \geq w^*$ and $\varphi(t_0, x_0) = w^*(t_0, x_0)$. So it remains to show that if $(q, p, M) \in \mathrm{J}^{2,+}w^*(t_0, x_0)$, there exists a function $\varphi \in C^{1,2}([0,T) \times \mathcal{X})$ such that $\varphi \geq w^*$, $\varphi(t_0, x_0) = w^*(t_0, x_0)$, and

$$(q, p, M) = \big(\partial_t \varphi(t_0, x_0), D_x \varphi(t_0, x_0), D^2_{xx}\varphi(t_0, x_0)\big).$$

Step 1. Construction of φ. For $r > 0$, define

$$h(r) = \sup\Big\{\frac{\max\{\zeta(t,x), 0\}}{\sqrt{(t-t_0)^2 + \|x - x_0\|^4}} : \\ (t,x) \in [0,T) \times \mathcal{X},\ \sqrt{(t-t_0)^2 + \|x-x_0\|^4} \leq r\Big\},$$

where $\zeta : [0,T] \times \overline{\mathcal{X}} \to \mathbb{R}$ is defined by

$$\zeta(t,x) = w^*(t,x) - w^*(t_0, x_0) - q(t - t_0) - p^\top (x - x_0) \\ - \frac{1}{2}(x - x_0)^\top M (x - x_0).$$

Clearly, h is non-decreasing and continuous on $(0, \infty)$. Moreover, since $(q, p, M) \in \mathrm{J}^{2,+}w^*(t_0, x_0)$, we have

$$\zeta(t,x) \leq o(|t - t_0| + \|x - x_0\|^2),$$

and

$$\limsup_{|t-t_0| + \|x - x_0\|^2 \to 0} \frac{|t - t_0| + \|x - x_0\|^2}{\sqrt{(t-t_0)^2 + \|x - x_0\|^4}} < \infty,$$

so that $h(0) = \lim_{r \downarrow 0} h(r) = 0$. Now define

$$\varphi(t,x) = F\big(r(t,x)\big) + w^*(t_0, x_0) + q(t - t_0) + p^\top (x - x_0) \\ + \frac{1}{2}(x - x_0)^\top M (x - x_0), \quad (t,x) \in [0,T) \times \mathcal{X},$$

with

$$\begin{aligned} r(t,x) &= \sqrt{(t-t_0)^2 + \|x - x_0\|^4}, \\ F(r) &= \frac{2}{3r}\int_r^{2r}\int_\xi^{2\xi} h(\rho)\mathrm{d}\rho\mathrm{d}\xi, \quad r > 0, \\ F(0) &= 0. \end{aligned}$$

We claim that φ is the desired function, i.e., $\varphi \in C^{1,2}([0,T) \times \mathcal{X})$, $\varphi \geq w^*$, $\varphi(t_0, x_0) = w^*(t_0, x_0)$, and

$$(q, p, M) = \big(\partial_t \varphi(t_0, x_0), D_x \varphi(t_0, x_0), D^2_{xx}\varphi(t_0, x_0)\big). \tag{A.4}$$

Step 2. Verifying the properties of φ. By monotonicity of h, we have $h(\rho) \leq h(4r)$ for all $\rho \in [0, 4r]$, so

$$0 \leq F(r) \leq r h(4r),$$

and hence F is continuous at 0 with $F(0) = 0$. For $r > 0$, we also compute

$$F'(r) = \frac{4}{3r}\int_{2r}^{4r} h(\xi)\,d\xi - \frac{2}{3r}\int_{r}^{2r} h(\xi)\,d\xi - \frac{1}{r}F(r),$$
$$F''(r) = \frac{2}{3r}\big(8h(4r) - 6h(2r) + h(r)\big) - \frac{2}{r}F'(r).$$

Using monotonicity and non-negativity of h, it follows that there exists a constant $C > 0$ (which we subsequently allow to grow from line to line) such that

$$|F'(r)| \le Ch(4r) \qquad \text{and} \qquad |F''(r)| \le \frac{C}{r}h(4r).$$

In particular, this implies that $F'(r)$ and $F''(r)$ both tend to 0 as $r \downarrow 0$. For $(t,x) \neq (t_0,x_0)$ we have

$$\partial_t F\big(r(t,x)\big) = \frac{t-t_0}{r(t,x)}F'\big(r(t,x)\big),$$
$$D_x F\big(r(t,x)\big) = \frac{2(x-x_0)\|x-x_0\|^2}{r(t,x)}F'\big(r(t,x)\big),$$

and since $|t-t_0|, \|x-x_0\|^2 \le r(t,x)$, it follows that

$$\lim_{(t,x)\to(t_0,x_0)} \partial_t F(r(t,x)) = 0 \quad \text{and} \quad \lim_{(t,x)\to(t_0,x_0)} D_x F(r(t,x)) = 0.$$

Regarding the Hessian, we have

$$\partial^2_{x_ix_j} F\big(r(t,x)\big) = \frac{2\|x-x_0\|^4}{r(t,x)^2}(x_i-(x_0)_i)(x_j-(x_0)_j)F''\big(r(t,x)\big)$$
$$+ \Big(\frac{2\|x-x_0\|^2}{r(t,x)}\delta_{i,j} + \frac{4(x_i-(x_0)_i)(x_j-(x_0)_j)(t-t_0)^2}{r(t,x)^3}\Big)F'\big(r(t,x)\big),$$

where $\delta_{i,j} = 1$ if $i = j$, and 0 otherwise. Since $|t-t_0|, \|x-x_0\|^2 \le r(t,x)$, it follows that

[CK]$_{14}$:Terrible notation...

[SC]$_{15}$:I've seen worse...

$$\big|\partial^2_{x_ix_j} F\big(r(t,x)\big)\big| \le Ch\big(4r(t,x)\big),$$

and so

$$\lim_{(t,x)\to(t_0,x_0)} D^2_{xx} F\big(r(t,x)\big) = 0.$$

Therefore, $F(r(t,x)) \in C^{1,2}([0,T) \times \mathcal{X})$ with

$$\partial_t F\big(r(t_0,x_0)\big) = 0, \quad D_x F\big(r(t_0,x_0)\big) = 0, \quad D^2_{xx} F\big(r(t_0,x_0)\big) = 0,$$

implying that $\varphi \in C^{1,2}([0,T) \times \mathcal{X})$ and (A.4) holds. Finally, we show that $\varphi \ge w^*$. From monotonicity of h, we get:

$$F(r) \ge rh(r). \tag{A.5}$$

Define

$$\eta(t,x) = w^*(t_0,x_0) + q(t-t_0) + p^\top(x-x_0) + \frac{1}{2}(x-x_0)^\top M(x-x_0),$$

so that $\varphi(t,x) = F(r(t,x)) + \eta(t,x)$. By definition of h,

$$w^*(t,x) - \eta(t,x) \le r(t,x)h\big(r(t,x)\big).$$

Using (A.5), we conclude

$$\varphi(t,x) = F\big(r(t,x)\big) + \eta(t,x) \ge r(t,x)h\big(r(t,x)\big) + \eta(t,x) \ge w^*(t,x),$$

which completes the proof. □

Bibliography

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[2] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

[3] Marco Avellaneda and Sasha Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, 2008.

[4] Richard Bellman. On a routing problem. *Quarterly of applied mathematics*, 16(1):87–90, 1958.

[5] V.E. Beneš. Existence of optimal strategies based on specified information, for a class of stochastic decision problems. *SIAM J. Control*, 8:179–188, 1970.

[6] Alain Bensoussan. *Estimation and control of dynamical systems*, volume 48. Springer, 2018.

[7] Dimitri Bertsekas. *A course in reinforcement learning.* Athena Scientific, 2024.

[8] Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

[9] J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5), 2024.

[10] Vladimir I. Bogachev. *Measure Theory*, volume 2 of *Springer Monographs in Mathematics.* Springer, Berlin, 2007.

[11] Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12362–12373. Curran Associates, Inc., 2021.

[12] M. Capiński and P.E. Kopp. *Measure, Integral and Probability.* Springer, 2nd edition, 2004.

[13] Álvaro Cartea and Sebastian Jaimungal. Risk metrics and fine tuning of high-frequency trading strategies. *Mathematical Finance*, 25(3):576–611, 2015.

[14] Álvaro Cartea, Sebastian Jaimungal, and José Penalva. *Algorithmic and high-frequency trading.* Cambridge University Press, 2015.

[15] David Cass. Optimum growth in an aggregative model of capital accumulation. *The Review of economic studies*, 32(3):233–240, 1965.

[16] Samuel N. Cohen and Robert J. Elliott. *Stochastic Calculus and Applications.* Birkhäuser, 2nd edition, 2015.

[17] Samuel N. Cohen, Christoph Knochenhauer, and Alexander Merkel. Optimal adaptive control with separable drift uncertainty. *SIAM Journal on Control and Optimization*, 63(2):1348–1373, 2025.

[18] Timm Faulwasser and Lars Grüne. Chapter 11 - turnpike properties in optimal control: An overview of discrete-time and continuous-time results. In Emmanuel Trélat and Enrique Zuazua, editors, *Numerical Control: Part A*, volume 23 of *Handbook of Numerical Analysis*, pages 367–400. Elsevier, 2022.

[19] A.F. Filippov. On certain questions in the theory of optimal control. *Vestnik Moskov. Univ. Ser. Mat. Meh. Astronom.*, 2:25–42, 1959. English trans. J. Soc. Indust. Appl. Math. Ser. A. Control 1 (1962), 76-84.

[20] W.H. Fleming and H.M. Soner. *Controlled Markov Processes and Viscosity Solutions.* Springer, 2nd edition, 2006.

[21] Hans Föllmer and Alexander Schied. *Stochastic Finance.* De Gruyter, 2025.

[22] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

[23] Siegfried Graf. Selected results on measurable selections. In Zdeněk Frolík, editor, *Proceedings of the 10th Winter School on Abstract Analysis*, pages 87–122, Palermo, 1982. Circolo Matematico di Palermo.

[24] Olivier Guéant. *The Financial Mathematics of Market Liquidity: From optimal execution to market making.* CRC Press, 2016.

[25] Olivier Guéant, Charles-Albert Lehalle, and Joaquin Fernandez-Tapia. Dealing with the inventory risk: a solution to the market making problem. *Mathematics and financial economics*, 7(4):477–507, 2013.

[26] Lars Peter Hansen and Thomas J. Sargent. *Robustness.* Princeton University Press, Princeton, 2007.

[27] Winston Harrington. Enforcement leverage when penalties are restricted. *Journal of Public Economics*, 37(1):29–53, 1988.

[28] Thomas Ho and Hans R Stoll. Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial economics*, 9(1):47–73, 1981.

[29] T. Jaakkola, M.I. Jordan, and S.P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. MIT AI Memo 1441, 1993.

[30] Monique Jeanblanc, Marc Yor, and Marc Chesney. *Mathematical methods for financial markets.* Springer Science & Business Media, 2009.

[31] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

[32] TC Koopmans. On the concept of optimal economic growth. *Study Week on the Econometric Approach to Development Planning*, page 225–287, 1965.

[33] N.V. Krylov. *Controlled Diffusion Processes.* Springer, 1980.

[34] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc., 2020.

[35] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms.* Cambridge University Press, 2020.

[36] E.J. McShane and R.B. Warfield, Jr. On Filippov's implicit functions lemma. *Proceedings of the American Mathematical Society*, 18:41–47, 1967.

[37] Sean Meyn. *Control systems and reinforcement learning.* Cambridge University Press, 2022.

[38] Huyên Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61. Springer Science & Business Media, 2009.

[39] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[40] Frank Plumpton Ramsey. A mathematical theory of saving. *The economic journal*, 38(152):543–559, 1928.

[41] L.C.G. Rogers and D. Williams. *Diffusions, Markov Processes and Martingales*, volume 1. Cambridge University Press, 2nd edition, 2000. Foundations.

[42] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, second edition, 2018.

[43] Nizar Touzi. *Optimal stochastic control, stochastic target problems, and backward SDE*, volume 29. Springer Science & Business Media, 2012.

[44] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

[45] D.J. White. Real applications of markov decision processes. *Interfaces*, 15(6), 1985.

[46] Peter Whittle. *Optimal control: basics and beyond*. John Wiley & Sons, Inc., 1996.

[47] 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), 2011.

[48] Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999.

Notation

$\mathbb{U}$ set of admissible controls. 35

$\mathcal{U}$ set in which controls take values. 35

X state process. 10, 123