Texts in Applied Mathematics 76

Onésimo Hernández-Lerma
Leonardo R. Laura-Guarachi
Saúl Mendoza-Palacios
David González-Sánchez

An Introduction to Optimal Control Theory

The Dynamic Programming Approach

Springer

Texts in Applied Mathematics

Volume 76

The mathematization of all sciences, the fading of traditional scientific boundaries, the impact of computer technology, the growing importance of computer modelling and the necessity of scientific planning all create the need both in education and research for books that are introductory to and abreast of these developments. The aim of this series is to provide such textbooks in applied mathematics for the student scientist. Books should be well illustrated and have clear exposition and sound pedagogy. Large number of examples and exercises at varying levels are recommended. TAM publishes textbooks suitable for advanced undergraduate and beginning graduate courses, and complements the Applied Mathematical Sciences (AMS) series, which focuses on advanced textbooks and research-level monographs.

Onésimo Hernández-Lerma ·
Leonardo R. Laura-Guarachi ·
Saul Mendoza-Palacios ·
David González-Sánchez

An Introduction to Optimal Control Theory

The Dynamic Programming Approach

Springer

Onésimo Hernández-Lerma
Department of Mathematics
CINVESTAV-IPN
Mexico City, Mexico

Leonardo R. Laura-Guarachi
Escuela Superior de Economía
Instituto Politécnico Nacional
Mexico City, Mexico

Saul Mendoza-Palacios
CIDE
Mexico City, Mexico

David González-Sánchez
Conacyt-Universidad de Sonora
Hermosillo, Sonora, Mexico

ISSN 0939-2475 ISSN 2196-9949 (electronic)
Texts in Applied Mathematics
ISBN 978-3-031-21138-6 ISBN 978-3-031-21139-3 (eBook)
https://doi.org/10.1007/978-3-031-21139-3

Mathematics Subject Classification: 49-01, 49L20, 60J25, 90C40, 93E20

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Max and Juli

- OHL

To Justina and Manuel

- LRLG

Dedicated to Edith, Paulina and Saul,

- SMP

To my family,

- DGS

Preface

These lecture notes present material for an introductory course on *optimal control theory* including deterministic and stochastic systems, with discrete- and continuous-time parameter. The course is *introductory* in the sense that no previous knowledge of control theory is required.

There are several well-known techniques to study Optimal Control Problems (OCPs), but here we are mainly interested in the *Dynamic Programming* (DP) approach. This is a very general technique introduced by Richard E. Bellman (1920–1984) in the 1950s, which is applicable to a large class of optimization and control problems. (See the remark at the end of this Preface.)

The DP approach gives *sufficient conditions* for an OCP to have a solution. The main tool is the so-called *Verification Theorem* that can be summarized as follows. Given an Optimal Control Problem (OCP):

1) Write an associated equation, called the *Dynamic Programming Equation* (DPE).

2) If the DPE has a "suitable" solution, then this solution can be used in turn to solve the given OCP.

Actually, part (1) is straightforward. The main difficulty is part (2); that is, finding suitable solutions to the DPE can be a non-trivial matter.

The DPE is known by several names, such as the *optimality equation*, the Bellman equation or, for continuous-time OCPs, the *Hamilton–Jacobi–Bellman equation*.

The notes begin in Chapter 1 with a general introduction to OCPs. This chapter presents basic notions, such as the objective functionals to be optimized, and the notion of control strategy or control policy. Some examples illustrate the classes of OCPs to be studied in the text.

Chapters 2 and 3 concern discrete-time systems. The former chapter considers the *deterministic* case in which the typical dynamic system is of the form

$$x_{t+1} = F_t(x_t, a_t) \text{ for } t = 0, 1, \ldots, T-1, \tag{1}$$

with a given initial condition x_0 is called the planning horizon. In (1), x_t denote the so-called state and control variables, respectively. The control actions a_t are taken according to a given control policy. Chapter 2 introduces finite ($T < \infty$) and infinite ($T = \infty$) horizon OCPs with objective functionals that are typically defined by finite or infinite sums, respectively. In the infinite-horizon case, we also introduce an asymptotic optimality criterion, namely, the long-run average cost.

The remaining chapters have essentially the same structure as Chapter 2, except for the dynamic model. In particular, in Chapter 3 we consider the *stochastic* analogue of (1) that is,

$$x_{t+1} = F_t(x_t, a_t, \xi_t), t = 0, 1, \ldots, T-1, \tag{2}$$

where x_t and a_t are as in (1), and the ξ_t are random variables that represent random perturbations. It is explained that, depending on the situation that they represent, these perturbations form either a *driving process* or a *random noise*.

In addition to the "system model" (2), in Chapter 3 we introduce the "Markov control model" in which we have a *transition probability* instead of a transition function F_t as in (2). The Markov control model was introduced by Bellman (1957b), who also coined the term "Markov decision process", and which today

is also known as a *Markov Control Process* (MCP). MCPs are especially useful in control problems (for instance, control of some queueing systems) in which we do not have an explicit system model as in (2). On the other hand, it is noted in the text that (1) and (2) are particular classes of MCPs.

The second half of these lectures concerns continuous-time OCPs. It begins in Chapter 4 with control problems in which the state process $x(\cdot)$ satisfies an ordinary differential equation (ODE)

$$\dot{x}(t) = F(t, x(t), a(t)) \text{ for } t \in [0, T]. \tag{3}$$

We again consider finite- and (discounted and undiscounted) infinite-horizon problems in which the "summations" that appear in Chapter 2 are replaced by integrals.

In several places of Chapter 4 it is emphasized that (3) defines, in fact, a certain family of continuous-time MCPs. The reason for doing this is that Chapter 5 introduces a *general* continuous-time MCP, in which the evolution of the state process is determined by an "infinitesimal generator" rather than a system function F as in (3). Since this fact is difficult to grasp from a conceptual viewpoint, immediately after introducing some standard dynamic programming facts, we show that the results in Chapter 4 are examples or particular cases of the ideas in Chapter 5.

Finally, as another *example* of a continuous-time MCP, in Chapter 6 we study the control of diffusion processes that, for our present purposes, can be expressed as solutions of certain stochastic differential equations that generalize the ODEs in (3).

Each of the Chapters 2–6 presents examples to illustrate the main results. It also includes a section with exercises.

Remark: Why Dynamic Programming? In other words, if there are several well-known techniques to analyze optimal control problems (OCPs), why do we emphasize DP? The answer is simple: it has more advantages than any of the other approaches to OCPs. Let us explain this.

1. DP is applicable to essentially *all* classes of OCPs, including deterministic and stochastic problems, with discrete- or continuous-time parameter, finite- or infinite-dimensional state space (see Fabbri et al. (2017)), with discrete (that is, finite or countable spaces) or general metric state space, etc. (In many applications, such as control of queues or reinforcement learning (see e.g. Sutton and Barto (2018)), we work in *finite* state spaces.)

2. For discrete–time problems, DP does not require smoothness conditions, which is important in some applications, for instance in model predictive control (see e.g. Raković and Levine (2019).)

3. In DP, there are well-known *approximation algorithms*, such as the (recursive) *value iteration algorithm* or the (monotone) *policy iteration algorithm*, which are the basis for some applications, such as adaptive (or approximate) DP and stabilization problems in control theory.

4. In DP, we automatically obtain feedback (or closed-loop or Markov) optimal controls, in contrast to the open-loop controls obtained when using, for instance, the Maximum Principle.

Some of these advantages, and many others, are practically impossible to obtain when using other solution techniques for OCPs.

On the negative side, the main disadvantage is what Bellman (1957a, 1961) called the *curse of dimensionality*, which basically refers to the difficulty in solving the DPE when the dimension increases. ◇

Mexico City, Mexico
October 2022

Onésimo Hernández-Lerma
Leonardo R. Laura-Guarachi
Saul Mendoza-Palacios
David González-Sánchez

Acknowledgments. The work by Onésimo Hernández-Lerma was partially supported by Consejo Nacional de Ciencia y Tecnologia (CONACYT-México) grant CF-2019/263963. Moreover, the work by Leonardo R. Laura-Guarachi and Saul Mendoza-Palacios was partially supported by CONACYT-México under grant Ciencia Frontera 2019-87787.

Contents

Preface vii

1 Introduction: Optimal Control Problems 1

2 Discrete–Time Deterministic Systems 13
2.1 The Dynamic Programming Equation 14
2.2 The DP Equation and Related Topics 20
2.2.1 Variants of the DP Equation 20
2.2.2 The Minimum Principle 23
2.3 Infinite–Horizon Problems 28
2.3.1 Discounted Case 29
2.3.2 The Minimum Principle 38
2.3.3 The Weighted-Norm Approach 44
2.4 Approximation Algorithms 50
2.4.1 Value Iteration 51
2.4.2 Policy Iteration 54
2.5 Long–Run Average Cost Problems 62
2.5.1 The AC Optimality Equation 63
2.5.2 The Steady–State Approach 67
2.5.3 The Vanishing Discount Approach 72

3 Discrete–Time Stochastic Control Systems 83
3.1 Stochastic Control Models 83
3.2 Markov Control Processes: Finite Horizon 87
3.3 Conditions for the Existence of Measurable Minimizers 92
3.4 Examples 96
3.5 Infinite–Horizon Discounted Cost Problems 100
3.6 Policy Iteration 110
3.7 Long-Run Average Cost Problems 111

3.7.1 The Average Cost Optimality Inequality ... 114
3.7.2 The Average Cost Optimality Equation 116
3.7.3 Examples 117

4 Continuous–Time Deterministic Systems 127
4.1 The HJB Equation and Related Topics 128
4.1.1 Finite–Horizon Problems: The HJB Equation 128
4.1.2 A Minimum Principle from the HJB Equation 135
4.2 The Discounted Case 141
4.3 Infinite–Horizon Discounted Cost 143
4.4 Long-Run Average Cost Problems 147
4.4.1 The Average Cost Optimality Equation (ACOE) 148
4.4.2 The Steady-State Approach 153
4.4.3 The Vanishing Discount Approach 158
4.5 The Policy Improvement Algorithm 163
4.5.1 The PIA: Discounted Cost Problems 164
4.5.2 The PIA: Average Cost Problems 170

5 Continuous–Time Markov Control Processes 183
5.1 Markov Processes 183
5.2 The Infinitesimal Generator 188
5.3 Markov Control Processes 198
5.4 The Dynamic Programming Approach 201
5.5 Long–Run Average Cost Problems 205
5.5.1 The Ergodicity Approach 209
5.5.2 The Vanishing Discount Approach 210

6 Controlled Diffusion Processes 217
6.1 Diffusion Processes 217
6.2 Controlled Diffusion Processes 221
6.3 Examples: Finite Horizon 223
6.4 Examples: Discounted Costs 231
6.5 Examples: Average Costs 237

Appendix A: Terminology and Notation **245**

Appendix B: Existence of Measurable Minimizers . . . **249**

Appendix C: Markov Processes **255**

Bibliography **263**

Index **271**

Chapter 1

Introduction: Optimal Control Problems

In a few words, in an **optimal control problem** (OCP) we are given a dynamical system that is "controllable" in the sense that its behavior depends on some parameters or components that we can choose within certain ranges. These components are called *control actions*. When we look at these control actions throught the whole period of time in which the system is functioning, then they form *control policies* or *strategies*. On the other hand, we are also given an *objective function* or *performance index* that somehow measures the system's response to each control policy. The OCP is then to find a control policy that optimizes the given objective function.

More precisely, in an OCP we are given:

1. A "controllable" dynamical system, which typically, depending on the time parameter, can be a *discrete–time* system or a *continuous-time* system. For instance, in the former case, the dynamical system is of the form

$$x_{t+1} = F_t(x_t, a_t, \xi_t) \quad \text{for} \quad t = 0, 1, \ldots, T-1, \tag{1.1}$$

with some given *initial state* x_0. (In Chaps. 4–6 we consider *continuous–time* systems.) In (1.1), at each time t, x_t denotes the state variable, with values in some *state space* X; a_t is the

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*, Texts in Applied Mathematics 76,
https://doi.org/10.1007/978-3-031-21139-3_1

control or action variable, with values in some *action space* A; and ξ_t is a *disturbance* or *perturbation* in a *disturbance set* S. In most applications of control theory, the spaces X, A, and S are subsets of finite-dimensional spaces. However, for technical reasons (to be briefly discussed in Remark 1.7), it is convenient to assume that they are general Borel spaces. Moreover, depending on the disturbances ξ_t, the system (1.1) is classified in deterministic, stochastic or uncertain. This is explained in Remark 1.2, below.

2. We are also given a set Π of admissible (or feasible) control policies or strategies, which are sequences $\pi = \{a_0, a_1, ...\}$ with values $a_t \in A$.
3. Finally, we are given a real–valued function V on $\Pi \times X$, which is the objective function or performance index. The function V can take many different forms. One of the most common is a *total cost*

$$V(\pi, x_0) := \sum_{t=0}^{T-1} c_t(x_t, a_t) + C_T(x_T), \tag{1.2}$$

where $\pi = \{a_0, \ldots, a_{T-1}\}$ denotes the strategy being used, and x_0 is the initial state for the system (1.1). The term $c_t(x_t, a_t)$, which is called the *stage cost*, denotes the cost incurred at time t given that x_t is the state of the system and a_t is the applied control action. The so–called *terminal cost function* $C_T(\cdot)$ in (1.2) depends on the *terminal state* x_T only. In (1.2), T is called the OCP's *planning horizon* and can be finite or infinite. The infinite–horizon case is obtained from (1.2) taking $C_T(\cdot) \equiv 0$ and letting $T \to \infty$.

With the above components, we can now state the OCP as follows: For each initial state x_0, optimize the objective function $V(\pi, x_0)$ over all $\pi \in \Pi$ subject to (1.1). Here, depending on the context, "optimize" means either "minimize" (for instance, if V is a cost function) or "maximize" (if V is a reward or a utility function). Thus, if we are minimizing V, then the OCP would be: Find $\pi^* \in \Pi$ such that

$$V(\pi^*, x) = \inf_{\pi \in \Pi} V(\pi, x) \quad \forall\, x_0 = x, \tag{1.3}$$

subject to (1.1). If this holds, then π^* is called an *optimal policy* or *optimal solution* to the OCP, and the function

$$V^*(x) := \inf_{\pi \in \Pi} V(\pi, x) = V(\pi^*, x) \quad \forall\, x \in X \tag{1.4}$$

is the OCP's *value function* or *minimum cost*. If, on the other hand, V is a profit or reward function to be maximized, then in (1.3)–(1.4) we write "sup" in lieu of "inf", and the value function V^* in (1.4) is called the OCP's *maximum utility* or *maximum reward*, depending on the context.

To ensure that the value function $V^*(\cdot)$ in (1.4) is finite–valued, we will suppose that the following assumption holds unless stated otherwise.

Basic Assumption.

(a) The cost functions c_t and C_T are nonnegative;

(b) There is a strategy $\pi \in \Pi$ such that $V(\pi, x) < \infty$ for all $x \in X$.

The dynamic programming approach under the condition (a) in the Basic Assumption (or when the cost functions are bounded below) is known as the *positive case*. In the *negative case* the cost functions are nonpositive (or bounded above). In Sect. 2.3.3 we will introduce the so-called *weighted-norm approach*, which allows positive and negative costs but with a restricted growth rate (Assumption 2.33(c)).

There are many conditions ensuring that $V^*(\cdot)$ is finite–valued, but our Basic Assumption greatly simplifies the mathematical presentation. For instance, if the condition (a) holds, and c_t, C_T are bounded above by a constant M, then the total cost in (1.2) satisfies that

$$0 \le V(\pi, x_0) \le (T+1)M \quad \forall\, \pi \text{ and } x_0.$$

The main role of the Basic Assumption is that it helps to fix ideas, and we can move forward to other aspects on an OCP.

Part (a) in the Basic Assumption guarantees that $V(\pi, x) \ge 0$ for all $\pi \in \Pi$ and $x_0 = x \in X$, and so $V^*(\cdot) \ge 0$. On the other hand, (b) yields that $V^*(x) < \infty$ for all $x \in X$. (Part (a) can

be replaced by the apparently weaker condition: c_t and C_T are bounded below.)

Example 1.1 (A production–inventory system). Consider a production system in which the state variable $x_t \in \mathbb{R}$ is the *stock* or *inventory level* of some product. A typical state equation for this system is

$$x_{t+1} = x_t + a_t - \xi_t \quad \forall\, t = 0, 1, \ldots, \tag{1.5}$$

where the control or action variable $a_t \geq 0$ is the amount to be ordered or produced (and immediately supplied) at the beginning of period t, and the disturbance $\xi_t \geq 0$ is the product's *demand*. Observe that (1.5) is simply a balance–like equation.

A negative stock, which occurs if $\xi_t > x_t + a_t$ in (1.5), is interpreted as a backlog that will be fulfilled as soon as the stock is replenished. However, if backlogging is not allowed, then the model (1.5) can be replaced with

$$x_{t+1} = \max\{x_t + a_t - \xi_t, 0\}.$$

Usually, production systems have a finite *capacity* C, so that, at each period t, we must have $x_t + a_t \leq C$ or $a_t \leq C - x_t$. Therefore, if the state is $x_t = x$, then we have the *control constraint*

$$a \in A(x), \quad \text{with} \quad A(x) = [0, C - x].$$

Thus, the action space is $A = [0, \infty)$, which contains the control constraint set $A(x)$ for every state x.

In a production–inventory system the objective function V in (1.2) is usually interpreted as a total profit or revenue function. Hence, given the unit sale price (p), the unit production cost (c), and the unit holding (or maintenance) cost (h), then at each time t, instead of a cost c_t in (1.2), we have a net revenue of the form

$$r_t(x_t, a_t) := py_t - ca_t - h(x_t + a_t),$$

where $y_t = \min\{\xi_t, x_t + a_t\}$ is the sale during period t. ◇

Remark 1.2. (a) The system (1.1) (or (1.5)) is said to be *deterministic* or *stochastic* or *uncertain* depending on whether the disturbances ξ_t are, respectively,

- given constants in (the disturbance set) S,
- S–valued random variables, or
- constants in S but with unknown values.

(b) If (1.1) is a stochastic system, the disturbances ξ_t are said to form a *driving process* if they have a concrete meaning, say, physical or economical. In contrast, if the disturbances are generic random variables (with no specific meaning), the process $\{\xi_t\}$ is called a *random noise*. For instance, in Example 1.1, if the demand variables ξ_t are random, then they form a driving process. On the other hand, in models of economic growth or population growth the disturbances typically form a random noise. (See Example C.4 and Remark C.5 in Appendix C.)

(c) If (1.1) is a stochastic system, the cost in the right–hand side of (1.2) is *random*. In this case, (1.2) is replaced with the *expected total cost*

$$V(\pi, x_0) := E\left[\sum_{t=0}^{T-1} c_t(x_t, a_t) + C_T(x_T)\right]. \tag{1.6}$$

(d) **Strategies.** To introduce a control policy or strategy $\pi = \{a_t\}$ it is important to specify the *information* available to the controller. In the simplest case, the control actions depend on the time parameter only; that is, $a_t = g(t)$ for some function g. In this case, π is called an *open–loop* policy. If, on the other hand, at each time t, the control action $a_t = g(t, x_t)$ depends on t and the current state x_t, then π is said to be a *closed–loop* policy, also known as a *feedback* or *Markov* policy. In general, if the actions are of the form $a_t = g(x_0, a_0, x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t)$ so that, at each time t, the action depends on the whole history of the process up to time t, then π is called a *nonanticipative* or *history–dependent* policy. In these notes, to fix ideas, we will consider only open–loop and closed–loop (or Markov) policies unless stated otherwise. See Fig. 1.1 for a representation of a feedback or closed–loop scheme. ◇

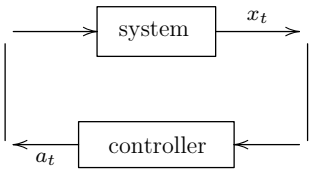

Fig. 1.1 A closed–loop scheme

Example 1.3 (A portfolio selection problem ≡ A consumption–investment problem). Let x_t be an investor's wealth at time $t = 0, 1, \ldots$. At each time t the investor decides how much to consume of his wealth, say c_t, and the rest $x_t - c_t$ is invested in a stock portfolio, which consists of

- a risk–free asset (e.g., bonds) with a fixed interest rate r, and
- a risky asset (e.g., stocks) with a random return rate ξ_t.

Note that $\{\xi_t\}$ is a *driving process* in the sense of Remark 1.2(b). Thus the control variable is $a_t = (c_t, p_t) \in [0, x_t] \times [0, 1]$ with $c_t =$ consumption, and $p_t =$ fraction of $x_t - c_t$ to be invested in the risky asset, and $1 - p_t$ is the fraction of $x_t - c_t$ invested in the riskless asset. The corresponding dynamical model is

$$x_{t+1} = [(1 - p_t)(1 + r) + p_t\xi_t](x_t - c_t), \ t = 0, 1, \ldots,$$

with some given initial wealth $x_0 = x \geq 0$. Clearly, for the investment to be profitable, we impose the condition $E(\xi_t) \geq r$ for all $t = 0, 1, \ldots$. (Otherwise, if the interest rate r is *greater* that the expected return rate $E(\xi_t)$, then obviously the best decision would be to invest in the risk–free asset.)

Hence we have a stochastic control system in which typically we wish to maximize a so-called *expected utility of consumption*

$$V(\pi, x) := E\left[\sum_{t=0}^{T} \beta^t u(c_t)\right], \quad \text{with} \quad T \leq \infty, \tag{1.7}$$

where $u(\cdot)$ is a given utility function. Moreover, given the interest rate $r > 0$, the *discount factor* is $\beta := (1 + r)^{-1}$. ◇

If $T = \infty$ in (1.7), then we obtain an infinite–horizon objective function called an infinite–horizon *discounted utility*. A quite

different infinite–horizon objective function is the *long–run expected average cost* defined as

$$V(\pi, x) := \limsup_{n\to\infty} n^{-1} E\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right], \tag{1.8}$$

where, as usual, $\pi = \{a_t\}$ is the control policy being used, $x_0 = x$ is the given initial state, and the lim sup is in order to minimize the long–run average cost in a worst case scenario. (From a mathematical viewpoint, taking the *lim sup* is convenient because it ensures that (1.8) is well defined, whereas the "limit" might not exist. Moreover, for theoretical reasons, it is more convenient to take *lim sup* rather than *lim inf*. We will come back to this point in the following chapters.) Observe that (1.8) is an *asymptotic value* that does not depend on the expected cost incurred in any finite number of stages. In fact, in many cases it is even independent of the initial state x, that is $V(\pi, x) \equiv V(\pi)$ for all x. Long-run average cost problems appear in both deterministic and stochastic problems. See, for instance, Sects. 2.5 and 5.5 (or 6.5), respectively.

Example 1.4 (A tracking problem). Consider the general, possibly stochastic system (1.1) with state space $X \subset \mathbb{R}^k$ and action set $A \subset \mathbb{R}^l$. In addition, we are given a *fixed* state trajectory, say, $\{x_t^*\}$ in X, and a *fixed* action trajectory $\{a_t^*\}$ in A. Finally, consider the long–run expected average cost (1.8) with running cost

$$c(x_t, a_t) := |x_t - x_t^*|^2 + |a_t - a_t^*|^2 \quad \forall\, t = 0, 1, ...,$$

which is essentially the distance from (x_t, a_t) to (x_t^*, a_t^*). Hence, the corresponding OCP of minimizing $V(\pi, x)$ over all π is called a *tracking problem* because the underlying idea is that the state–action pair (x_t, a_t) should "track" or "follow" or "pursue" or "stay as close as possible to" the given trajectories (x_t^*, a_t^*). In particular, if $a_t^* \equiv 0$ for all t, then we have a tracking problem with *minimum fuel*.

Tracking problems are very common in engineering and economics. As an example, controlling the attitude of a satellite or keeping it as close as possible to a given orbit are tracking problems. (See Fig. 1.2.) For applications in economics, see Kendrick (2002).

Fig. 1.2 Controlling a satellite's attitude

If the state dynamics (1.1) in the tracking problem is stochastic (so that the perturbations ξ_t are random variables), then the process $\{\xi_t\}$ would typically be a *random noise*. This is simply because there is no possibility of assigning a "physical" meaning to ξ_t. ◇

Remark 1.5 (Partially observable systems). In the tracking problem of Example 1.4, suppose that the state x_t of the system is the attitude of a satellite—see Fig. 1.2. Hence we have a particular case of a so-called *partially observable system*, which is so–named precisely because the state of the system is *not* observable. Mathematically speaking, one can model this class of systems as follows.

The evolution of the state process x_t is as in (1.1), where usually the disturbance $\{\xi_t\}$ is a random noise. As already noted, the state x_t is not observable, but there is an *observation process* y_t behaving according to an equation of the form

$$y_t = G_t(x_t, \eta_t) \quad \forall\, t = 0, 1, 2, ..., \tag{1.9}$$

where $\{\eta_t\}$ is a random noise influencing the observation process.

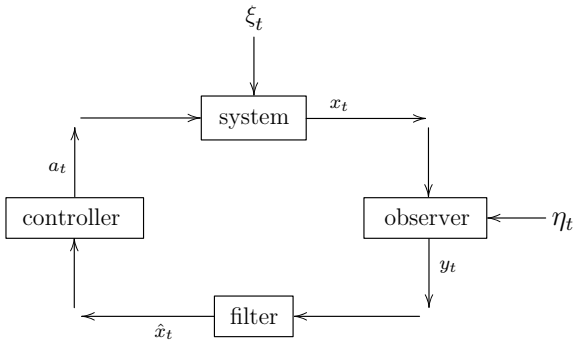

Fig. 1.3 A partially observable control system

Using the observation process $\{y_t\}$ and a suitable *filter* (that is, some statistical device), we obtain an (statistical) *estimator* $\hat{x}_t$ of the state x_t, which is used by the controller to obtain a control of the form $a_t = g(t, \hat{x}_t)$. (See Fig. 1.3 for a graphical representation of a partially observable control system.)

Under suitable assumptions, a partially observable system consisting of Eqs. (1.1) and (1.9) can be transformed into a *completely observable* system in which the original (unobservable) state x_t is replaced by the estimator $\hat{x}_t$. In some particular cases, the estimator $\hat{x}_t$ is a finite–dimensional vector, but in general it is a probability distribution. Hence, the "completely observable" system is an OCP with state variable $\hat{x}_t$ in a space of probability measures! For details see, for instance, Bäuerle and Rieder (2011), Bertsekas and Shreve (1978) or Hernández-Lerma (1989). ◇

We conclude this section with some comments on the continuous–time systems that we will study in Chaps. 4–6.

Remark 1.6 (Continuous–time control problems). For continuous–time control problems we again distinguish (as in the discrete–time case above) between deterministic and stochastic problems. In the former case, the state equation, which is the analogue of (1.1), is an ordinary differential equation

$$\dot{x}(t) = F(t, x(t), a(t)) \quad \text{for} \quad t \in [0, T], \tag{1.10}$$

with some given initial state $x(0) = x_0$. In (1.10), for each t, $x(t)$ and $a(t)$ denote the state variable and the control action that

belong to appropriate spaces, say, $X \subset \mathbb{R}^n$ and $A \subset \mathbb{R}^m$, respectively. Similarly, the cost functional (1.2) is replaced by an integral

$$V(\pi, x_0) := \int_0^T c(t, x(t), a(t))dt + C_T(x(T)), \tag{1.11}$$

where the instantaneous (or running) cost $c(t, x, a)$ and the terminal cost $C_T(x)$ are given functions.

The control variable $a(\cdot)$ in (1.10) and (1.11) depends on the information available to the controller. Here, to simplify the presentation, we only consider

- *open–loop* controls $a(t) := g(t)$, where $g : [0, T] \to A$ is a given (measurable) function; and
- *closed–loop* (or *feedback* or *Markov*) controls $a(t) := g(t, x(t))$, for some (measurable) function $g : [0, T] \times X \to A$.

In the stochastic case, (1.10) is replaced by a *stochastic differential equation*

$$dx(t) = F(t, x(t), a(t))dt + G(t, x(t), a(t))dW(t), \tag{1.12}$$

with $x(\cdot) \in \mathbb{R}^n$, assuming that the state space is $X \subset \mathbb{R}^n$, as in (1.10). Moreover, in (1.12), $W(\cdot) \in \mathbb{R}^d$ is a d–dimensional *Wiener process* (also known as a *Brownian motion*), and $G(t, x, a)$ is a n–by–d matrix function. In this case, the cost functional in (1.11) is a random variable and so we replace it by its expected value:

$$V(\pi, x_0) := E\left[\int_0^T c(t, x(t), a(t))dt + C_T(x(T))\right]. \tag{1.13}$$

The stochastic control problem is to minimize (1.13) subject to (1.12). ◇

The remainder of these notes is organized as follows. Chapters 2 and 3 deal with discrete–time systems. Chapter 2 considers deterministic systems, and Chap. 3 is about the stochastic case.

The remaining chapters deal with continuous-time problems. Chapter 4 concerns control of continuous–time deterministic systems in which the state dynamics is an ordinary differential

equation, as in (1.10). Chapter 5 introduces general continuous–time Markov control processes (MCPs), which include the deterministic systems in Chap. 4 and many other stochastic control systems. In this general framework we show that some aspects of stochastic control problems can be analyzed by exploiting the "Markovian nature" of the involved dynamic systems. As an application of the results for general MCPs we show how to recover some results for the deterministic systems in Chap. 4. (Here, deterministic systems are viewed as a class of, say, "degenerate" MCPs.) Finally, as another application of MCPs, in Chap. 6 we consider a class of controlled stochastic differential equations (1.12), also known as controlled diffusion processes.

Remark 1.7. Why should we use *Borel spaces* (see Appendix A)? To answer this question we should note that the state space X and the action set A of an OCP can be of a different "topological" nature. For instance, in Examples 1.1 and 1.3, X and A are sets in (finite–dimensional) Euclidean spaces $\mathbb{R}$ or $\mathbb{R}^2$. However, in the control of some queueing systems, for instance, X and A are *discrete spaces*, that is, either finite or countably infinite sets. Indeed, in this case, the state variable x_t is typically the "number of customers" (that is, a *nonnegative integer*) in the system at time t. Moreover, the state space X can be finite or infinite, depending on the system's "capacity", and the action set A can be finite or infinite. For example, in a *control of admissions* problem, A is a two-point set $\{a, b\}$, where, for each arriving customer, $a :=$ allow the customer to access the waiting line, and $b :=$ reject the customer. At the other extreme, X and/or A can be *infinite–dimensional* sets, as in the book by Fabbri et al. (2017). As an example, consider the partially observable system in Remark 1.5. If we transform this system into a completely observable one, in which the state variable is the *estimator* $\hat{x}_t$, then the new state space will be a set $\hat{X}$ of probability measures! Clearly, if we consider separately each of these cases, the theory could be a little confusing. To avoid this situation, when dealing with a general OCP we will simply assume that X and A are *Borel spaces*; this includes all the cases mentioned above. (See Appendix A.)

On the other hand, the Borel space context requires a suitable concept of "measurability" of sets and functions. Here, however, to avoid being repetitious, *throughout these lectures we assume that all sets and functions are Borel-measurable*. In fact, this measurability assumption usually holds in our case, because we mostly consider standard, "well-behaved" settings in which functions are, for instance, continuous or differentiable and so forth, and similarly for sets—we mostly deal with nice sets, such as closed or open intervals. To conclude, we believe that to learn from these lecture notes it is not necessary to know Measure Theory. Nevertheless, if the reader wishes to learn about it, he/she can look at several introductory reader-friendly texts, such as Ash (1972), Bartle (1995), Bass (2020), ... ◇

Chapter 2

Discrete–Time Deterministic Systems

In this chapter we consider the discrete–time system (1.1) in the so–called *deterministic* case (see Remark 1.2(a)), so that the disturbances ξ_t are supposed to be given constants in some space S. Since this information is irrelevant for our present purposes, we will omit the notation ξ_t so (1.1) becomes

$$x_{t+1} = F_t(x_t, a_t) \quad \text{for} \quad t = 0, 1, \ldots, T-1, \tag{2.0.1}$$

with a given initial condition x_0. The state and control spaces X and A are given spaces. (Recall our assumption in Remark 1.7: In these lectures, all sets and functions are Borel measurable.)

First, we consider the finite-horizon case, $T < \infty$. Hence, the optimal control problem (OCP) we are concerned with is to minimize the total cost in (1.2), that is,

$$V(\pi, x_0) := \sum_{t=0}^{T-1} c_t(x_t, a_t) + C_T(x_T) \tag{2.0.2}$$

over the set Π of all control policies (or strategies) $\pi = \{a_0, \ldots, a_{T-1}\}$ subject to (2.0.1). At each time t, the set of feasible actions is a set $A(x) \subset A$, which may depend on the current state x. (If necessary, the action set $A(x)$ may also depend on the time t.) For this OCP an optimal policy and the corresponding value function are defined as in (1.3) and (1.4), respectively.

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*, Texts in Applied Mathematics 76,
https://doi.org/10.1007/978-3-031-21139-3_2

In the following Sects. 2.1 and 2.2.1 we introduce the dynamic programming approach to study the OCP (2.0.1)–(2.0.2). This approach gives *sufficient conditions* for the existence of an optimal control policy. In contrast, in Sect. 2.2.2 we introduce the minimum principle (also known as Pontryagin's principle) that gives *necessary conditions*. In Sect. 2.3 we study an *infinite-horizon* problem, so $T = \infty$ in (2.0.1)–(2.0.2), with $C_T(\cdot) \equiv 0$. To this end we use both DP and the minimum principle in Sects. 2.3.1 and 2.3.2, respectively. Moreover, in Sect. 2.3.3 we introduce the *weighted-norm approach* that allows positive and negative cost functions. In Sect. 2.4 we consider again the infinite-horizon problem but now from the viewpoint of two approximation schemes, *value iteration* (VI) and *policy iteration* (PI). Both schemes are very useful in applications, and easily extended to stochastic problems (as in Chap. 3). We conclude this chapter with an analysis, in Sect. 2.5, of *long-run average cost problems*. These problems are very popular in the control of some queueing systems, and also in computer and telecommunications applications, in which we wish to optimize some asymptotic cost.

Remark 2.1. Sometimes, without further comments, we will tacitly assume that the OCPs we are dealing with are *consistent* in the sense that they are well defined and admit optimal solutions. Of course, in all of the particular cases considered below we give conditions ensuring consistency. ◇

Recall also the Basic Assumption at the beginning of Chap. 1 about the cost functions c_t, C_T being nonnegative, and $V^*(\cdot) < \infty$.

2.1 The Dynamic Programming Equation

The dynamic programming (DP) technique is based on the following "principle of optimality" stated by Richard Bellman (1920–1984).

Lemma 2.2 (The principle of optimality (PO)). Let $\pi^* = \{a_0^*, \ldots, a_{T-1}^*\}$ be an optimal strategy for the OCP (2.0.1)–(2.0.2), that is,

$$V(\pi^*, x_0) = \min_{\pi} V(\pi, x_0) \quad \forall\, x_0.$$

Let $\{x_0^*, \ldots, x_T^*\}$ be the corresponding path obtained from (2.0.1), so $x_0^* = x_0$. Then, for any time $s \in \{0, \ldots, T-1\}$, the "truncated" strategy $\pi_s^* = \{a_s^*, \ldots, a_{T-1}^*\}$ from time s onward is an optimal strategy that leads the system (2.0.1) from the point x_s^* to the point x_T^*.

We will next sketch the proof of this lemma. The details are left as an exercise for the reader. (See Exercise 2.4)

Sketch of the proof of Lemma 2.2. Arguing by contradiction, suppose that the lemma's conclusion does not hold; that is, for some $0 \le s < T-1$, the truncated policy π_s^* is *not* optimal. Therefore, by our consistency assumption in Remark 2.1, there exists a policy $\hat{\pi}_s := \{\hat{a}_s, \ldots, \hat{a}_{T-1}\}$ that is optimal in the interval $[s, T-1]$ so that

$$V_s(\hat{\pi}_s, x_s^*) < V_s(\pi_s^*, x_s^*),$$

where

$$V_s(\pi_s, x) := \sum_{t=s}^{T-1} c_t(x_t, a_t) + C_T(x_T) \tag{2.1.1}$$

is the total cost from time s onward when using the truncated policy $\pi_s := \{a_s, \ldots, a_{T-1}\}$, given that $x_s = x$. Now consider the combined policy $\tilde{\pi} = \{\tilde{a}_0, \ldots, \tilde{a}_{T-1}\}$ defined as

$$\tilde{a}_t := \begin{cases} a_t^* \text{ if } 0 \le t < s, \\ \hat{a}_t \text{ if } s \le t \le T-1. \end{cases}$$

Then, by definition of $\tilde{\pi}$, $V_0(\tilde{\pi}, x_0) < V_0(\pi^*, x_0)$, which contradicts the optimality of π^* stated in the lemma. □

We will now show how to use Lemma 2.2 to obtain the dynamic programming equation (2.1.8)–(2.1.9) below.

Consider the OCP (2.0.1)–(2.0.2) but only from time s onward, with initial state $x_s = x$, that is, the cost function to be minimized is as in (2.1.1), i.e.,

$$V_s(\pi_s, x) := \sum_{t=s}^{T-1} c_t(x_t, a_t) + C_T(x_T).$$

Let $v_s(x)$ be the corresponding minimal cost, that is,

$$v_s(x) := \inf_{\pi_s} V_s(\pi_s, x). \tag{2.1.2}$$

Moreover, since no control actions are applied at the terminal time T, we define

$$v_T(x) := C_T(x). \tag{2.1.3}$$

Hence, by Lemma 2.2, (2.1.2) becomes

$$\begin{aligned} v_s(x) &= V_s(\pi_s^*, x) \\ &= \sum_{t=s}^{T-1} c_t(x_t^*, a_t^*) + C_T(x_T^*) \\ &= c_s(x, a_s^*) + V_{s+1}(\pi_{s+1}^*, x_{s+1}^*) \\ &= c_s(x, a_s^*) + v_{s+1}(x_{s+1}^*). \end{aligned}$$

Therefore, by (2.0.1),

$$v_s(x) = c_s(x, a_s^*) + v_{s+1}(F_s(x, a_s^*)). \tag{2.1.4}$$

On the other hand, by definition (2.1.2) of $v_s(x)$,

$$v_s(x) \le c_s(x, a) + v_{s+1}(F_s(x, a)) \quad \forall\, a \in A(x). \tag{2.1.5}$$

Finally, combining (2.1.4) and (2.1.5) we obtain that, for $s \in \{0, \dots, T-1\}$ and $x \in X$,

$$v_s(x) = \min_{a \in A(x)} [c_s(x, a) + v_{s+1}(F_s(x, a))], \tag{2.1.6}$$

with *terminal condition* as in (2.1.3):

$$v_T(x) := C_T(x) \quad \forall\, x \in X. \tag{2.1.7}$$

Equation (2.1.6) with the terminal condition (2.1.7) is called the **dynamic programming (DP) equation** or the **Bellman equation** for the OCP (2.0.1)–(2.0.2). The DP equation is the

basis of the *dynamic programming algorithm* in the following theorem.

Theorem 2.3 (Dynamic programming theorem). *Let $J_0, \dots, J_T$ be the functions defined "backward" (from $s = T$ to $s = 0$) on X by*

$$J_T(x) := C_T(x), \tag{2.1.8}$$

and for $s = T-1, T-2, \dots, 0$,

$$J_s(x) := \min_{a \in A(x)} [c_s(x,a) + J_{s+1}(F_s(x,a))]. \tag{2.1.9}$$

Suppose that for each $s = 0, 1, \dots, T-1$, there exists a function $a_s^* : X \to A$ that attains the minimum in the right hand side of (2.1.9) for all $x \in X$. Then the Markov strategy $\pi^* = \{a_0^*, \dots, a_{T-1}^*\}$ is optimal and the value function coincides with J_0, i.e.,

$$\inf_{\pi} V(\pi, x) = V(\pi^*, x) = J_0(x) \quad \forall\, x \in X. \tag{2.1.10}$$

Actually, for each $s = 0, \dots, T$, J_s coincides with the function in (2.1.2)–(2.1.3), i.e.,

$$v_s(x) = J_s(x) \quad \forall\, s = 0, 1, \dots, T,\ x \in X. \tag{2.1.11}$$

Proof. Let $v_T, v_{T-1}, ..., v_0$ be the functions defined by (2.1.2)–(2.1.3), and let π^* be the Markov strategy defined in the theorem. Then, for each $s \in \{0, 1, \dots, T\}$, the definition of v_s yields

$$v_s(x) \le V_s(\pi^*, x) = J_s(x) \quad \forall\, x \in X,$$

where the second equality follows from (2.1.9), which can be expressed as $J_s(x) = c_s(x, a_s^*) + J_{s+1}(x_{s+1}^*)$. Hence, to complete the proof of (2.1.11), it remains to show that

$$v_s(x) \ge J_s(x) \tag{2.1.12}$$

for all $s \in \{0, 1, \dots, T\}$ and $x \in X$.

We will prove (2.1.12) using backward induction. Indeed, first note that, by (2.1.7) and (2.1.8), the equality holds in (2.1.12) when $s = T$. Now let us suppose that (2.1.12) holds for $s+1$, i.e.,

$$v_{s+1}(x) \geq J_{s+1}(x) \quad \forall\, x \in X. \tag{2.1.13}$$

Take an arbitrary policy $\pi = \{a_0, \ldots, a_{T-1}\}$. Then, for any $x \in X$,

$$\begin{aligned} V_s(\pi, x) &= c_s(x, a_s) + V_{s+1}(\pi, F_s(x, a_s)) \\ &\geq c_s(x, a_s) + v_{s+1}(F_s(x, a_s)) \\ &\geq c_s(x, a_s) + J_{s+1}(F_s(x, a_s)) \quad [\text{by } (2.1.13)] \\ &\geq \min_{a \in A(x)} [c_s(x, a) + J_{s+1}(F_s(x, a))] \\ &= J_s(x). \end{aligned}$$

Since this holds for any policy π, we obtain (2.1.12). □

In the following example we illustrate Theorem 2.3 with an LQ control problem (also known as a "linear regulator problem"), which consists of a *Linear* state equation with a *Quadratic* stage cost.

Example 2.4 (Discrete–time LQ system). Consider the linear system

$$x_{t+1} = \alpha x_t + \beta a_t \quad \forall\, t = 0, 1, \ldots, T-1;\ x_0 \text{ given},$$

with nonzero coefficients α, β. The state and action spaces are $X = A = \mathbb{R}$. The objective function (or performance index) is

$$V(\pi, x_0) := \sum_{t=0}^{T-1} (q x_t^2 + r a_t^2) + q_T x_T^2,$$

where $r > 0$ and $q, q_T \geq 0$.

In this case, the dynamic programming equation (2.1.8)–(2.1.9) becomes

$$J_T(x) := q_T x^2 \tag{2.1.14}$$

and for $s = T-1, T-2, \ldots, 0$:

$$J_s(x) := \min_a [q x^2 + r a^2 + J_{s+1}(\alpha x + \beta a)]. \tag{2.1.15}$$

This equation is solved by backward induction: substituting (2.1.14) into (2.1.15) we obtain

$$J_{T-1}(x) := \min_a [qx^2 + ra^2 + q_T(\alpha x + \beta a)^2].$$

Therefore,

$$J_{T-1}(x) := \min_a [(q + q_T\alpha^2)x^2 + (r + q_T\beta^2)a^2 + 2q_T\alpha\beta xa].$$

The right–hand side of this equation is minimized when

$$a^*_{T-1}(x) = G_{T-1}x, \quad \text{with} \quad G_{T-1} := -(r + q_T\beta^2)^{-1}q_T\alpha\beta,$$

and the minimum is

$$J_{T-1}(x) = K_{T-1}x^2, \quad \text{with} \quad K_{T-1} := (r + q_T\beta^2)^{-1}q_T r\alpha^2 + q.$$

In general, using backward induction we can see that the optimal strategy $\pi^* = \{a^*_0, \ldots, a^*_{T-1}\}$ is given by

$$a^*_s(x) = G_s x, \quad \text{with} \quad G_s := -(r + K_{s+1}\beta^2)^{-1}K_{s+1}\alpha\beta, \quad (2.1.16)$$

with coefficients K_s defined recursively by $K_T := q_T$ and for $s = T-1, \ldots, 0$:

$$K_s = (r + K_{s+1}\beta^2)^{-1}K_{s+1}r\alpha^2 + q.$$

Likewise, the optimal cost (2.1.11) from time s onward becomes

$$J_s(x) = K_s x^2 \quad \text{for} \quad s = 0, 1, \ldots, T-1. \quad (2.1.17)$$

In particular, with $s = 0$ we obtain the minimum cost in (2.1.10). ◇

Remark 2.5. (The nonstationary vector LQ problem.) For notational ease, we have considered above the *scalar* (or one-dimensional) LQ problem in the *stationary* case, in which the state and action spaces are $X = A = \mathbb{R}$ and the coefficients α, β, q, r are all *time-invariant* constants. However, essentially the same arguments (with obvious notational changes) are valid in the general nonstationary vector case with state and action spaces $X = \mathbb{R}^n$ and $A = \mathbb{R}^m$, respectively, and time-varying matrix coefficients $\alpha_t \in \mathbb{R}^{n\times n}, \beta_t \in \mathbb{R}^{n\times m}$, and quadratic stage costs

$$c(x, a) = x'q_t x + a'r_t a, \quad t = 0, 1, \ldots$$

where "prime" ($'$) denotes transpose, and q_t and r_t are symmetric matrices, with q_t nonnegative definite, and r_t positive definite. (The latter means that, for all $t, x \in \mathbb{R}^n$, and $a \in \mathbb{R}^m$, we have $x'q_t x \geq 0$ and $a'r_t a > 0$ for $a \neq 0$.) In this case, (2.1.16)–(2.1.17) become

$$a_s^*(x) = G_s x \quad \forall x \in \mathbb{R}^n$$

with

$$G_s = -(r_s + \beta_s' K_{s+1} \beta_s)^{-1} \beta_s' K_{s+1} \alpha_s,$$

where $K_T = q_T$ and, for $s = T-1, ..., 0$,

$$K_s = \alpha_s' [K_{s+1} - K_{s+1}\beta_s (r_s + \beta_s' K_{s+1} \beta_s)^{-1} \beta_s' K_{s+1}] \alpha_s + q_s.$$

With this value of K_s, the optimal (minimum) cost in (2.1.17) becomes $J_s(x) = x' K_s x$ for $s = 0, 1, ..., T-1$. For further details see, for instance, Sect. 2.1 in Bertsekas (1987). ◇

2.2 The DP Equation and Related Topics

This section is divided in two parts. First, in Sect. 2.2.1 we introduce some variants of the DP equation (2.1.8)–(2.1.9). Then, in Sect. 2.2.2, we consider the OCP (2.0.2)–(2.0.1) from the viewpoint of the *minimum principle*, which gives *necessary conditions* for optimality. The idea is to compare this principle with the DP approach.

2.2.1 Variants of the DP Equation

Discounted costs. Let us suppose that (2.0.2)–(2.0.1) are of the form

$$V(\pi, x_0) = \sum_{t=0}^{T-1} \alpha^t c_t(x_t, a_t) + \alpha^T C_T(x_T)$$

and

$$x_{t+1} = F_t(x_t, a_t) \quad \forall\, t = 0, 1, \ldots, T-1,$$

respectively, where $\alpha \in (0,1)$ is a given *discount factor*. Then the DP equation (2.1.8)–(2.1.9) becomes

$$J_T(x) = \alpha^T C_T(x),$$

$$J_s(x) = \min_a [\alpha^s c_s(x,a) + J_{s+1}(F_s(x,a))].$$

Now consider the change of variable $U_s(x) := \alpha^{-s} J_s(x)$. Then we obtain the so–called DP equation in the *discounted case*:

$$U_s(x) = \min_a [c_s(x,a) + \alpha U_{s+1}(F_s(x,a))] \quad \forall\, s = 0, \ldots, T-1, \tag{2.2.1}$$

with

$$U_T(x) = C_T(x). \tag{2.2.2}$$

Example 2.6 (Optimal advertising, Adukov et al. (2015)). Let us consider a market where a monopolistic firm is entering with a new product. The firm's market share at time t is x_t, and its advertising expenditure rate is a_t. Suppose that the market share evolves according to the nonlinear system

$$x_{t+1} = (1-\delta)x_t + \rho a_t (1-x_t)^{1-\sigma} \quad \text{for} \quad t = 0, 1, \ldots, T-1, \tag{2.2.3}$$

where the state and control spaces are $X = A = [0,1]$, $\rho \in (0,1)$ is the effectiveness of advertising, $\delta \in [0,1]$ is the rate at which consumers lose interest in the product, and $\sigma \in [0,1]$ is the nonlinearity parameter.

In order to interpret the nonlinear part of (2.2.3), first note that from the Taylor expansion for $(1-x)^{1-\sigma}$ at $x = 0$, we have

$$(1-x)^{1-\sigma} = (1-x) + \sigma x(1-x) + \sigma^2 x^2 + \ldots.$$

Therefore, the nonlinear term $\rho a_t (1-x_t)^{1-\sigma}$ can be approximated by the sum of a portion representing the new consumers due to the direct advertising, $\rho a_t (1-x_t)$; and the new consumers due to the word-of-mouth advertising by active consumers x_t, namely, $\sigma \rho a_t x_t (1-x_t)$.

Now, going back to our OCP, given an initial state x_0, we want to maximize the profit function

$$V(\pi, x_0) = \sum_{t=0}^{T-1} \alpha^t \left(p x_t - c a_t^{\frac{1}{\sigma}} \right) + \alpha^T p_T x_T,$$

where $p > 0$ and $p_T \geq 0$ are potential revenues, and $c > 0$ is the advertising cost.

This problem can be solved explicitly by means of Eqs. (2.2.1)–(2.2.2) for maximization problems. That is

$$U_T(x) := p_T x,$$

and for $s = 0, 1, ..., T-1$:

$$U_s(x) := \max_a \left\{ px - ca^{\frac{1}{\sigma}} + \alpha U_{s+1} \left((1-\delta)x + \rho a (1-x)^{1-\sigma} \right) \right\}.$$

By backward induction, for $s = T-1$, we have

$$U_{T-1}(x) = \max_a \left\{ px - ca^{\frac{1}{\sigma}} + \alpha p_T \left((1-\delta)x + \rho a (1-x)^{1-\sigma} \right) \right\}.$$

The maximum is attained at

$$a_{T-1}^*(x) = \left[M\left(p_T\right)(1-x) \right]^\sigma, \text{ where } M(\tau) := \left(\frac{\alpha \sigma \rho \tau}{c} \right)^{\frac{1}{1-\sigma}},$$

and so

$$U_{T-1}(x) = N(p_T)x + c\frac{1-\sigma}{\sigma} M\left(p_T\right),$$

with

$$N(\tau) := p + \alpha(1-\delta)\tau - c\frac{1-\sigma}{\sigma} M(\tau).$$

Continuing with the backward induction, we obtain that the optimal control policy $\pi^* = \{a_0^*, \ldots, a_{T-1}^*\}$ is given by

$$a_s^*(x) = \left[M\left(N^{T-s-1}(p_T) \right)(1-x) \right]^\sigma, \quad s = 0, 1, ..., T-1,$$

where N^{T-s-1} is the $(T-s-1)$th iterate of the function N.

From (2.2.3), the corresponding state path is

$$x_{s+1}^* = (1-\delta)x_s^* + \rho [M(N^{T-s-1}(p_T))]^\sigma (1 - x_s^*),$$

and the optimal benefit from time s onward turns out to be

$$U_s(x) = N^{T-s}(p_T)x + c\frac{1-\sigma}{\sigma}\alpha^{T-s-1}\sum_{i=0}^{T-s-1}\alpha^{-i}M\left(N^i(p_T)\right)$$

for $s = 0, 1, ..., T-1$. ◇

Forward form of the DP equation. Consider the DP equation (2.1.8)–(2.1.9) and let $U_s := J_{T-s}$ for $s = 0, 1, \ldots, T$. Then the DP equation becomes

$$U_s(x) = \min_a \{c_s(x,a) + U_{s-1}(F_s(x,a))\} \tag{2.2.4}$$

for $s = 1, 2, \ldots, T$, with *initial condition*

$$U_0(x) = C_T(x).$$

Moreover, if $f_{T-s}(x)$ minimizes the right–hand side of (2.1.9) at the stage $T-s$, then $g_s := f_{T-s}$ minimizes (2.2.4) at the stage s, and $\pi^* = \{g_T, ..., g_1\}$ is an optimal policy, that is

$$U_T(x) = J_0(x) = V(\pi^*, x).$$

In the *discounted case*, the forward form of (2.2.1)–(2.2.2) becomes

$$u_s(x) = \min_{a\in A(x)}[c_s(x,a) + \alpha u_{s-1}(F_s(x,a))] \tag{2.2.5}$$

for $s = 1, 2, \ldots, T$, with $u_0(x) = C_T(x)$.

2.2.2 The Minimum Principle

Remark 2.7. (a) In this section, we suppose that the state and action spaces X and A are subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. We also assume that the cost functions in (2.0.2) and the system functions in (2.0.1) are differentiable.

(b) Given two column vectors x, y of the same (finite) dimension, we write their inner product as $x \cdot y := \sum_i x_i y_i$. Sometimes we also write $x \cdot y$ as $\langle x, y\rangle$. If x is a row vector and y a column vector, then we write their inner product simply as xy.

(c) Consider the OCP (2.0.1)–(2.0.2) and the *Hamiltonian function*

$$H_t(x, a, \rho) := c_t(x, a) + \rho \cdot F_t(x, a), \quad t = 0, 1, ...$$

for $(x, a, \rho) \in X \times A \times \mathbb{R}^n$. Under the differentiability assumption in (a), the Hamiltonian is also differentiable. ◇

The conditions in Remark 2.7 are one of the key differences between the minimum principle (MP) and the dynamic programming (DP) approach. In the former, typically, the state and action spaces X and A are finite-dimensional and the functions in (2.0.1)–(2.0.2) require some differentiability condition. In contrast, in DP, X and A are general Borel spaces (that is, Borel subsets of complete and separable metric spaces), and the functions in (2.0.1)–(2.0.2) require some mild condition, for instance, piecewise continuity.

The minimum principle gives *necessary conditions* for optimality. Roughly, it states the following: If $\pi^* = \{a_t^*\}$ is an optimal strategy and $\{x_t^*\}$ is the corresponding state path, then the pairs $(x_t^*, a_t^*), t = 0, 1, \ldots$, satisfy, for some vectors $\rho_0, \rho_1, \ldots$, the conditions (2.2.6)–(2.2.8) below. In particular, (2.2.7) yields that the optimal controls a_t^* minimize the Hamiltonian function H_t, which gives the name "the minimum principle".

The minimum principle was originally developed for continuous-time problems (Gamkrelidze 1999). In this section we obtain the discrete–time minimum principle in the form of first–order necessary conditions.

Theorem 2.8. (The minimum principle). *Let* $\pi^*=\{a_0^*, \ldots, a_{T-1}^*\}$ *be an optimal strategy for the OCP (2.0.1)–(2.0.2) and* $\{x_0^*, \ldots, x_T^*\}$ *the corresponding path. Then there exist vectors* $\rho_1, \ldots, \rho_T$ *in* $\mathbb{R}^n$ *such that*

(a) for all $t = 1, \ldots, T-1$,

$$\frac{\partial c_t}{\partial x}(x_t^*, a_t^*) + \rho_{t+1} \frac{\partial F_t}{\partial x}(x_t^*, a_t^*) = \rho_t, \tag{2.2.6}$$

(b) for all $t = 0, \ldots, T-1$,

$$\frac{\partial c_t}{\partial a}(x_t^*, a_t^*) + \rho_{t+1}\frac{\partial F_t}{\partial a}(x_t^*, a_t^*) = 0, \tag{2.2.7}$$

(c) we have the terminal condition (TC)

$$\rho_T = \frac{\partial C_T}{\partial x}(x_T^*). \tag{2.2.8}$$

Proof. Consider the "cost to go function" from time s using π^* given $\tilde{x}_s = x$, i.e.,

$$V_s^*(x) = \sum_{t=s}^{T-1} c_t(\tilde{x}_t, a_t^*) + C_T(\tilde{x}_T), \tag{2.2.9}$$

where each $\tilde{x}_t$ is generated by (2.0.1) using π_s^*. Note that V_s^* is a differentiable function (see (2.2.10) below.)

Define

$$\rho_t := \frac{\partial V_t^*}{\partial x}(x_t^*),$$

for all $t = 1, \ldots, T$.

(a) Taking $a_t = a_t^*$ in (2.1.6), it follows that $V_t^*(x) = c_t(x, a_t^*) + V_{t+1}^*(F_t(x, a_t^*))$ for all $t = 1, \ldots, T-1$. Differentiating and evaluating in x_t^*, we obtain

$$\frac{\partial V_t^*(x_t^*)}{\partial x} = \frac{\partial c_t}{\partial x}(x_t^*, a_t^*) + \frac{\partial V_{t+1}}{\partial x}(F_t(x_t^*, a_t^*))\frac{\partial F_t}{\partial x}(x_t^*, a_t^*), \tag{2.2.10}$$

which is (2.2.6).

(b) By the Principle of Optimality (Lemma 2.2) and (2.1.2), $V_t^*(x_t^*) = v_t(x_t^*)$. This fact, by (2.1.6), implies

$$V_t(x_t^*) = \min_{a \in A(x)} \left\{ c_t(x_t^*, a) + V_{t+1}^*(F_t(x_t^*, a)) \right\},$$

for all $t = 0, \ldots, T-1$. Since a_t^* minimizes the righ–hand side, we obtain

$$\frac{\partial c_t}{\partial a}(x_t^*, a_t^*) + \frac{\partial V_{t+1}^*}{\partial x}(F_t(x_t^*, a_t^*))\frac{\partial F_t}{\partial a}(x_t^*, a_t^*) = 0,$$

which is (2.2.7).

(c) Finally, by (2.2.9), we have $V_T^*(x) = C_T(\tilde{x}_T)$, so

$$\rho_T = \frac{\partial V_T(x_T^*)}{\partial x} = \frac{\partial C_T(x_T^*)}{\partial x}.$$

This completes the proof of the theorem. □

Sufficient conditions can also be provided with additional convexity assumptions, as follows.

Theorem 2.9. *Suppose that there exists a strategy* $\pi^* = \{a_0^*, \ldots, a_{T-1}^*\}$ *and vectors* $\rho_1, \ldots, \rho_T$ *such that (2.2.6), (2.2.7) and (2.2.8) hold. Define* $h_0 : X \times A \to \mathbb{R}$ *as*

$$h_0(x, a) = c_0(x, a) + \rho_1 \cdot F_0(x, a),$$

and for $t = 1, \ldots, T-1$, $h_t : X \times A \to \mathbb{R}$ *as*

$$h_t(x, a) = c_t(x, a) + \rho_{t+1} \cdot F_t(x, a) - \rho_t \cdot x.$$

If h_t *is convex for all* $t = 1, \ldots, T-1$ *and* C_T *is convex, then* π^* *is optimal for the OCP (2.0.1)–(2.0.2).*

Proof. Since (2.2.6) and (2.2.7) are the first order conditions of optimality for each h_t, the convexity of h_t implies that (x_t^*, a_t^*) minimizes h_t. Analogously, x_T^* minimizes $C_T(x) - \rho_T \cdot x$.

Let $\pi = \{a_0, \ldots, a_{T-1}\}$ be an arbitrary policy. We have

$$\begin{aligned}
\sum_{t=0}^{T-1} c_t(a_t^*, x_t^*) + C_T(x_T^*) &= \sum_{t=0}^{T-1} h_t(a_t^*, x_t^*) + \sum_{t=1}^{T-1} \rho_t \cdot x_t^* \\
&\quad - \sum_{t=0}^{T-1} \rho_{t+1} \cdot x_{t+1}^* + C_T(x_T^*) \\
&\leq \sum_{t=0}^{T-1} h_t(x_t, a_t) - \rho_T \cdot x_T^* + C_T(x_T^*) \\
&\leq \sum_{t=0}^{T-1} h_t(x_t, a_t) - \rho_T \cdot x_T + C_T(x_T)
\end{aligned}$$

$$= \sum_{t=0}^{T-1} c_t(a_t, x_t) + C_T(x_T).$$

Thus π^* is optimal. □

Example 2.10 (An economic growth model). This is a model introduced by Brock and Mirman (1972). The state and control variables x_t and a_t denote *capital* and *consumption*, respectively, at time $t = 0, 1, \ldots$. The state and control spaces are $X = A = [0, \infty)$, and the dynamics of the system is given by

$$x_{t+1} = cx_t^\theta - a_t \quad \text{for} \quad t = 0, 1, \ldots, T-1, \tag{2.2.11}$$

with $\theta \in (0, 1)$ and x_0 given. Let $A(x) := (0, cx^\theta]$, and consider the objective function or performance index

$$V(\pi, x_0) = \sum_{t=0}^{T-1} \alpha^t \log a_t + \alpha^T \log x_T^\theta. \tag{2.2.12}$$

In this OCP, the economic interpretation is that the controller wishes to determine a consumption strategy $\{a_t\}$ to *maximize* the total discounted utility (2.2.12), subject to the capital dynamics (2.2.11). The term cx_t^θ in (2.2.11) represents the output as a function of the current capital x_t and the technological parameter c; this output is distributed in consumption a_t and capital x_{t+1} for the next period.

To solve this problem we write the minimum principle equations (2.2.6)–(2.2.8) as follows

$$\rho_{t+1} c\theta x_t^{\theta-1} = \rho_t \quad \text{for} \quad t = 1, ..., T-1, \tag{2.2.13}$$

$$\frac{\alpha^t}{a_t} - \rho_{t+1} = 0 \quad \text{for} \quad t = 0, ..., T-1, \tag{2.2.14}$$

with the terminal condition

$$\rho_T = \frac{\theta \alpha^T}{x_T}. \tag{2.2.15}$$

From (2.2.14) for $t = T-1$ and (2.2.15),

$$a_{T-1} = \frac{\alpha^{T-1}}{\rho_T} = \frac{x_T}{\theta\alpha} = \frac{cx_{T-1}^{\theta} - a_{T-1}}{\theta\alpha}$$

which implies

$$a_{T-1} = \frac{cx_{T-1}^{\theta}}{1+\theta\alpha}.$$

Again, from (2.2.13) and (2.2.14) for $t = T-2$,

$$a_{T-2} = \frac{\alpha^{T-2}}{\rho_{T-1}} = \frac{a_{T-1}}{c\theta\alpha x_{T-1}^{\theta-1}} = \frac{x_{T-1}}{\theta\alpha + (\theta\alpha)^2} = \frac{cx_{T-2}^{\theta} - a_{T-2}}{\theta\alpha + (\theta\alpha)^2},$$

yielding

$$a_{T-2} = \frac{cx_{T-2}^{\theta}}{1+\theta\alpha+(\theta\alpha)^2}.$$

Continuation of this process backward in time yields the optimal control policy

$$a_t^* = \frac{c[x_t^*]^{\theta}}{1+\theta\alpha+\cdots+(\theta\alpha)^{T-t}} = \left[\frac{c(1-\theta\alpha)}{1-(\theta\alpha)^{T-t+1}}\right][x_t^*]^{\theta} \quad (2.2.16)$$

for all $t = 0, ..., T-1$. ◇

Remark 2.11. Theorems 2.8 and 2.9 above, and also Theorem 2.29 below, are simplified versions of results by Domínguez-Corella and Hernández-Lerma (2019). ◇

2.3 Infinite–Horizon Problems

In this section we consider infinite-horizon problems. First, in Sect. 2.3.1 we study the so-called *stationary discounted* case by means of the DP approach. Then in Sect. 2.3.2 we consider general nonstationary problems using the minimum principle (MP). Again, as in Sect. 2.2.2, the idea is to compare the DP and the MP techniques. Finally, in Sect. 2.3.3, we introduce the so-called *weighted-norm approach.*

A relevant question is, why should we consider *infinite horizon* problems? Do they appear in "real" situations? Are they important? Detailed answers will require to refer to the approximation algorithms and the asymptotic or long-run averages in Sects. 2.4 and 2.5, respectively. Therefore, we defer this topic to the end of the chapter.

2.3.1 Discounted Case

Instead of the OCP (2.0.1)–(2.0.2), consider the *stationary discounted* OCP: Minimize

$$V_n(\pi, x) := \sum_{t=0}^{n} \alpha^t c(x_t, a_t) \tag{2.3.1}$$

over all policies $\pi = \{a_0, \ldots, a_n\}$ subject to

$$x_{t+1} = F(x_t, a_t) \quad \forall\, t = 0, \ldots, n-1, \tag{2.3.2}$$

with a given initial state $x_0 = x$. In (2.3.1), $\alpha \in (0, 1)$ is a given *discount factor*.

Remark 2.12. The problem (2.3.1)–(2.3.2) is called *stationary* because the cost $c(x, a)$ and the system function $F(x, a)$ are time–invariant. Similarly, a Markov policy $\pi = \{a_t\}$, with $a_t = g(t, x_t)$ for all $t = 0, 1, \ldots$ (see Remark 1.2(d)) is said to be *stationary* if $g(t, x) \equiv g(x)$ for all t. In other words, the control actions $a_t = g(x_t)$ depend on the time parameter t only through the state x_t. Stationary control policies usually appear in infinite–horizon OCPs only. (See Corollary 2.22, for instance.) ◇

In this section, we are interested in the *infinite horizon* OCP obtained from (2.3.1) by letting $n \to \infty$; that is, the OCP now is to minimize

$$V(\pi, x) := \sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \tag{2.3.3}$$

subject to (2.3.2). Let

$$V^*(x) := \inf_{\pi} V(\pi, x) \quad \forall\, x \in X, \tag{2.3.4}$$

which is called the α–**discount value function**. A policy π^* is said to be α–**discount optimal** if

$$V(\pi^*, x) = V^*(x) \quad \forall\, x \in X.$$

The infinite-horizon discounted OCP (2.3.2)–(2.3.3) was introduced by Blackwell (1965).

Example 2.13 (Tracking problems). As noted in Example 1.4, tracking problems are a standard topic in some areas of economics and engineering, among other fields. The idea is as follows.

We are given a *nominal control path* $\{a_t^*, t = 0, 1, \dots\}$ and a *nominal state trajectory* $\{x_t^*, t = 0, 1, \dots\}$. The OCP is then to minimize the *tracking cost*

$$V(\pi, x) := \sum_{t=0}^{\infty} \alpha^t [(x_t - x_t^*)^2 + (a_t - a_t^*)^2] \tag{2.3.5}$$

subject to a state equation such as (2.3.2). In particular, if this state equation is *linear*, say,

$$x_{t+1} = Fx_t + Ga_t,$$

the tracking problem becomes an LQ (Linear–Quadratic) problem. Observe that, essentially, (2.3.5) is an ℓ^2–distance from the state–control trajectory $\{(x_t, a_t), t = 0, 1, \dots\}$ to the given nominal trajectory $\{(x_t^*, a_t^*), t = 0, 1, 2, ...\}$, so the tracking problem is to keep the state–control trajectory as close as possible to the nominal trajectory. In particular, if $a_t^* \equiv 0$ for all $t = 0, 1, \dots$, in engineering this is called a tracking problem with *minimum fuel*. Similar problems are dealt with in economics. ◇

In this section we are interested, among other things, in providing solutions to the following problems.

Problem 1. Let $V_n^*(x) := \inf_\pi V_n(\pi, x)$ be the **minimal cost** or **value function** for the n–stage OCP. Under what conditions does it hold that, as $n \to \infty$, V_n^* converges to V^*? That is, when do we have

$$\lim_{n\to\infty} V_n^*(x) = V^*(x) \tag{2.3.6}$$

for all $x \in X$?

In (2.3.7), below, $A(x) \subset A$ denotes the set of feasible control actions in the state x, for each $x \in X$.

Problem 2. Under what conditions is V^* a solution of the **dynamic programming equation**, also known as the **Bellman equation** or α–**optimality equation** (α–OE) in the discounted–cost case, that is,

$$V^*(x) = \inf_{a\in A(x)} [c(x,a) + \alpha V^*(F(x,a))] \tag{2.3.7}$$

for all $x \in X$?

To put it in other words, consider the so–called **Bellman operator** K defined as

$$Kv(x) := \inf_{a\in A(x)} [c(x,a) + \alpha v(F(x,a))] \tag{2.3.8}$$

for functions v in a certain family. Then we can restate Problem 2 as follows: when is V^* a fixed point of K? That is, when can we ensure that

$$V^* = KV^* \tag{2.3.9}$$

holds?

Remark 2.14. The fact that V^* satisfies (2.3.7) or (2.3.9) is not quite surprising. Indeed, for the α–discounted OCP (2.3.1)–(2.3.2), the DP equation (2.2.5) becomes

$$V_n^*(x) = \min_{a\in A(x)} [c(x,a) + \alpha V_{n-1}^*(F(x,a))] \quad \forall\, n = 1, 2, \ldots \tag{2.3.10}$$

with $V_0^*(\cdot) \equiv 0$. Therefore, if in (2.3.10) we let $n \to \infty$ and, in addition, (2.3.6) holds, then we would expect to obtain (2.3.7) (if interchanging of the "lim" and "min" operations is allowed—see Lemma 2.15 below). On the other hand, note that, using the operator K in (2.3.8), we may express (2.3.10) as an *iteration* of K, that is,

$$V_n^* = KV_{n-1}^* = K^n V_0^*, \quad \text{with} \quad V_0^* \equiv 0. \tag{2.3.11}$$

Due to this fact, the functions V_n^* are also known as *value iteration* (VI) functions. ◇

Note that (2.3.7) and (2.3.9) always have the *trivial solutions* $v \equiv \infty$ and $v \equiv -\infty$. We are interested, of course, in finite–valued functions.

The following lemma gives conditions that allow to interchange limits and minima.

Lemma 2.15. Let $\{g_k\}$ be a sequence of real–valued functions on a space Y converging to a function g. Each of the following conditions (a), (b), (c) ensures that

$$\lim_k \inf_y g_k(y) = \inf_y \lim_k g_k(y) = \inf_y g(y).$$

(a) $g_k \downarrow g$,

(b) g_k converges uniformly to g; that is, for each $\epsilon > 0$ there exists a number $k(\epsilon)$ such that, for all $y \in Y$,

$$|g_k(y) - g(y)| < \epsilon \quad \text{whenever} \quad k > k(\epsilon).$$

(c) Y is a metric space, the functions g_k are *inf–compact* (that is, for each $k = 1, 2, \ldots$ and $r \in \mathbb{R}$, the set $\{y \in Y : g_k(y) \leq r\}$ is compact), and, moreover, $g_k \uparrow g$.

Proof. See Exercise 2.5. □

We will next use some definitions and facts from Appendix B. Let

$$\mathbb{K} := \{(x, a) \in X \times A | a \in A(x)\} \tag{2.3.12}$$

be the *graph* of the multifunction $x \mapsto A(x)$. (See Definition B.1(b).) We denote by $\mathbb{F}$ the family of (measurable) functions—called *selectors*—$f : X \to A$ such that $f(x) \in A(x)$ for all $x \in X$. (Note that a selector $f \in \mathbb{F}$ is simply a function from X to A whose graph $(x, f(x))$ is in $\mathbb{K}$ for all x.)

Lemma 2.16. Let $u : \mathbb{K} \to \mathbb{R}$ be nonnegative and $\mathbb{K}$*–inf–compact*, that is (as in Definition B.4(a2)), for every compact set $X' \subset X$ and every $r \in \mathbb{R}$, the set

$$\{(x, a) \in G(X') | u(x, a) \leq r\}$$

is compact, where $G(X') := \{(x,a) \in X' \times A | a \in A(x)\}$. Then

(a) There exists $f \in \mathbb{F}$ such that

$$u^*(x) := \min_{a \in A(x)} u(x,a) = u(x, f(x)) \quad \forall\, x \in X,$$

and u^* is l.s.c.

(b) If $u, u_k : \mathbb{K} \to \mathbb{R}$, for $k = 1, 2, \ldots$, are bounded below and $\mathbb{K}$–inf–compact, and $u_k \uparrow u$, then

$$\lim_{k \to \infty} \min_{a \in A(x)} u_k(x,a) = \min_{a \in A(x)} u(x,a)$$

for all $x \in X$.

Proof. (a) This part follows from Theorem B.9. It also follows from Theorem B.8, using the fact that (by Lemma B.6)

$$\mathbb{K} - \text{inf-compactness} \;\Rightarrow\; \text{inf-compactness on } \mathbb{K}, \qquad (2.3.13)$$

and also implies lower semi–continuity of u (by Theorem B.9).

(b) Fix an arbitrary $x \in X$, and let $g(a) := u(x,a)$ and $g_k(a) := u_k(x,a)$. Then $g_k \uparrow g$ and, by (2.3.13) again, the functions g_k are inf–compact on $A(x)$. Since x was arbitrary, (b) follows from Lemma 2.15 (c). □

In the context of the infinite–horizon OCP (2.3.2)–(2.3.3), our **Basic Assumption** (in Chap. 1) states that:

(a) the stage cost $c : \mathbb{K} \to \mathbb{R}$ is nonnegative, and
(b) there exists a control policy $\pi \in \Pi$ such that

$$V(\pi, x) < \infty \quad \text{for all} \quad x \in X.$$

In addition to the Basic Assumption, in the remainder of this section we suppose the following.

Assumption 2.17. (a) The cost function $c \geq 0$ is $\mathbb{K}$–inf–compact;
(b) The system function $F : \mathbb{K} \to X$ in (2.3.2) is continuous.

We will denote by $L(X)$ the family of l.s.c. functions on X, and by $L^+(X)$ the subfamily of nonnegative functions in $L(X)$. Observe that $L^+(X)$ is a *convex cone*. (See Exercise 2.9.)

Lemma 2.18. Suppose that Assumption 2.17 holds, and let $L^+(X)$ be the family of nonnegative and l.s.c. functions on X. Then:

(a) The Bellman operator K in (2.3.8) maps $L^+(X)$ into itself; that is, if v is in $L^+(X)$, then so is Kv.
(b) If v is in $L^+(X)$, then there exists a selector $f \in \mathbb{F}$ that attains the minimum in (2.3.8), i.e.,

$$Kv(x) = c(x, f(x)) + \alpha v(F(x, f(x))) \quad \forall\, x \in X.$$

Proof. Fix an arbitrary function $v \in L^+(X)$, and define

$$u(x, a) := c(x, a) + \alpha v(F(x, a)) \quad \forall\, (x, a) \in \mathbb{K}. \tag{2.3.14}$$

Both results (a) and (b) will follow from Lemma 2.16(a) if we show that the nonnegative function u is $\mathbb{K}$–inf–compact. To this end, take an arbitrary compact set $X' \subset X$ and $r \in \mathbb{R}$, and note that the set

$$G(X')_c := \{(x, a) \in G(X') | c(x, a) \leq r\}$$

is *compact*, because c is $\mathbb{K}$–inf–compact (Assumption 2.17(a)). Moreover, $G(X')_c$ contains the set

$$G(X')_u := \{(x, a) \in G(X') | u(x, a) \leq r\}.$$

Therefore, to see that u is $\mathbb{K}$–inf–compact, it suffices to show that $G(X')_u$ is *closed*, that is, if a sequence of elements $(x_k, a_k) \in G(X')_u$ converges to (x, a), then (x, a) is in $G(X')_u$. This, however, is obvious because the mapping $(x, a) \mapsto u(x, a)$ is l.s.c. (Exercise 2.9). □

Remark 2.19. Consider a selector $f \in \mathbb{F}$. To simplify the notation we will write $c(x, f(x))$ and $F(x, f(x))$ as $c(x, f)$ and $F(x, f)$, respectively. Moreover, we will identify f with the stationary Markov control policy $\pi = \{a_t\}$ such that $a_t(x_t) = f(x_t)$ for all $t = 0, 1, \ldots$. Similarly, if $\pi = f$, we will express the total α–discounted cost in (2.3.3) as $V(f, x)$, that is,

$$V(f, x) = \sum_{t=0}^{\infty} \alpha^t c(x_t, f) \tag{2.3.15}$$

for all initial state $x_0 = x$. Note that expanding the righ–hand side of (2.3.15) we obtain

$$V(f,x) = c(x,f) + \alpha \sum_{t=1}^{\infty} \alpha^{t-1} c(x_t, f).$$

Thus, by (2.3.2), we can express (2.3.15) as

$$V(f,x) = c(x,f) + \alpha V(f, F(x,f)) \quad \forall\, x \in X$$

or, using (2.3.2),

$$V(f,x_t) = c(x_t,f) + \alpha V(f, x_{t+1}) \tag{2.3.16}$$

for all $t = 0, 1, \ldots$. ◇

We need the following "monotonicity property" of the Bellman operator K before going back to Problems 1 and 2.

Lemma 2.20. Consider the Basic Assumption and Assumption 2.17. If $v \in L^+(X)$ is such that $v \geq Kv$, then

(a) There exists $f \in \mathbb{F}$ such that $v(x) \geq V(f,x)$ for all $x \in X$; therefore,
(b) $v \geq V^*$.

Proof. (a) If $v \geq Kv$, then, by Lemma 2.18(b), there exists $f \in \mathbb{F}$ such that

$$v(x) \geq c(x,f) + \alpha v(F(x,f)) \quad \forall\, x \in X.$$

Iteration of this inequality yields, for all $n = 1, 2, \ldots$ and $x \in X$,

$$v(x) \geq \sum_{t=0}^{n-1} \alpha^t c(x_t, f) + \alpha^n v(x_n).$$

Therefore, since $v \geq 0$, letting $n \to \infty$ we obtain

$$v(x) \geq V(f,x) \quad \forall\, x \in X.$$

(b) Thus, since $V(f,\cdot) \geq V^*(\cdot)$, we conclude that $v \geq V^*$. □

We are now ready to give conditions under which the answer to both Problems 1 and 2 is affirmative.

Theorem 2.21. *Consider the Basic Assumption and Assumption 2.17. Then*

(a) *(2.3.6) holds, in fact* $V_n^* \uparrow V^*$, *and*
(b) V^* *is the minimal solution of the* α*-optimality equation (2.3.7); that is,* $V^* = KV^*$ *and if* $v \in L^+(X)$ *also satisfies* $v = Kv$, *then* $v \geq V^*$.
(c) V^* *is the "unique" solution of the DPE in the following sense: If* $v \in L^+(X)$ *is such that* $v = Kv$ *and, in addition, for any policy* $\pi = \{a_t\}$ *and the corresponding state trajectory* $\{x_t\}$ *we have*

$$\lim_{n\to\infty} \alpha^n v(x_n) = 0, \tag{2.3.17}$$

then $v = V^*$. *(The condition (2.3.17) is sometimes called a "transversality condition".)*

Proof. Since $c \geq 0$, the definition of V_n^* gives

$$V_n^*(x) \leq V_n(\pi, x) \leq V(\pi, x)$$

for any policy π and $x \in X$, so

$$V_n^*(x) \leq V^*(x) \quad \forall\, x \in X. \tag{2.3.18}$$

On the other hand, since the sequence $\{V_n^*\}$ is nondecreasing, $V_n^* \uparrow v$ for some function v such that, by (2.3.18),

$$v \leq V^*. \tag{2.3.19}$$

Furthermore, from (2.3.10) and Lemma 2.16(b), v is in $L^+(X)$ and it satisfies (2.3.9), that is,

$$v = \lim_n V_n^* = \lim_n KV_{n-1}^* = Kv.$$

It follows from Lemma 2.20 that $v \geq V^*$. This fact and (2.3.19) give that $v = V^*$. This proves both (a) and (b).

(c) If $v = Kv$, Lemma 2.20 gives that $v \geq V^*$. To obtain the reverse inequality first note that $v = Kv$ implies

$$v(x) \leq c(x, a) + \alpha v(F(x, a)) \quad \forall\, (x, a) \in \mathbb{K}.$$

Therefore, for any policy $\pi = \{a_t\}$ and the associated state trajectory x_t,

$$v(x_t) \le c(x_t, a_t) + \alpha v(x_{t+1}) \quad \forall\, t = 0, 1, \ldots .$$

Iteration of this inequality, given an arbitrary initial state $x_0 = x$, gives

$$v(x) \le \sum_{t=0}^{n-1} \alpha^t c(x_t, a_t) + \alpha^n v(x_n) \quad \forall\, n = 1, 2, \ldots ,$$

i.e., $v(x) \le V_{n-1}(\pi, x) + \alpha^n v(x_n)$. Finally, letting $n \to \infty$, from (2.3.17) we obtain $v(x) \le V(\pi, x)$. Thus, since π and $x_0 = x$ were arbitrary, it follows that $v(\cdot) \le V^*(\cdot)$. This completes the proof of part (c). □

As a consequence of Theorem 2.21(b) and Lemma 2.18(b) we obtain the existence of an optimal stationary policy for the infinite–horizon OCP (2.3.2)–(2.3.3), as follows.

Corollary 2.22. Under the hypotheses of Theorem 2.21, there exists $f^* \in \mathbb{F}$ such that $f^*(x) \in A(x)$ attains the minimum in the right–hand side of (2.3.7), that is,

$$V^*(x) = c(x, f^*) + \alpha V^*(F(x, f^*)) \quad \forall\, x \in X, \tag{2.3.20}$$

and f^* is an optimal stationary policy.

Proof. The existence of f^* as in (2.3.20) follows from Lemma 2.18(b). Moreover, as in (2.3.16), iteration of (2.3.20) gives that $V^*(x) = V(f^*, x)$ for all $x \in X$. Therefore, f^* is α–discount optimal. □

In the following examples we use the notation in (2.3.2)–(2.3.4). These examples show that before using DP results (such as Theorem 2.21) we should carefully verify their hypotheses.

Example 2.23 (Bertsekas 1987, p. 212). Let $X = [0, \infty)$, $c(x, a) \equiv 0$, and $F(x, a) = x/\alpha$. Then, for any constant b, the function $V(x) = bx$, for $x \in X$, satisfies (2.3.7). Hence, the operator K in (2.3.8)–(2.3.9) has an infinite number of fixed points in this

case. However, it has a unique fixed point in the class $B(X)$ of real–valued *bounded functions* on X, namely, the zero function $V^*(\cdot) \equiv 0$, which is the optimal cost function for this example. (See Remark 2.25 below, and also the Exercise 2.7.) ◇

Example 2.24 (Bertsekas 1987, p. 215). Let $X = \mathbb{R}$, $A = A(x) = (0, 1]$ for all $x \in X$, $c(x, a) = |x|$, $F(x, a) = \alpha^{-1}ax$. It can be verified that $V^*(x) = |x|$ for all $x \in X$. Now let π be the stationary policy such that $\pi(x) = 1$ for all $x \in X$. Then the cost function $v(\cdot) = V(\pi, \cdot)$ in (2.3.3) is $v(x) = \infty$ if $x \neq 0$ and $v(0) = 0$, so π is not optimal because $v(\cdot) \neq V^*(\cdot)$. Nevertheless, it can be verified that v is a fixed point of K, so it satisfies (2.3.9). ◇

The Exercise 2.7 and other results below use the following well-known fact.

Remark 2.25 (Banach's fixed point theorem). Let $(\mathcal{X}, \rho)$ be a complete metric space, and $T : \mathcal{X} \to \mathcal{X}$ a *contraction mapping*, that is, there exists a number $\beta \in (0, 1)$ such that

$$\rho(Tu, Tv) \leq \beta\rho(u, v) \quad \forall\, u, v \in \mathcal{X}.$$

Then T has a unique fixed point $u^* \in \mathcal{X}$, i.e.,

$$Tu^* = u^*.$$

Moreover, for any $u \in \mathcal{X}$, $T^n u$ converges to the fixed point u^*; in fact,

$$\rho(T^n u, u^*) \leq \beta^n \rho(u, u^*) \quad \forall\, n = 0, 1, \ldots,$$

where $T^n := T(T^{n-1})$ for all $n = 1, 2, ...$, with $T^0 :=$identity. ◇

2.3.2 The Minimum Principle

Now we introduce *the minimum principle* for an infinite–horizon OCP. As in Sect. 2.2.2, these necessary conditions for optimality are stated under smoothness properties. We assume that the state space X is a subset of $\mathbb{R}^n$, the action space A is a subset of $\mathbb{R}^m$, and the cost functions $c_t : X \times A \to \mathbb{R}$ as well as the system functions $F_t : X \times A \to X$ are differentiable in the interior of $X \times A$.

Given an initial state $x_0 = x$ and the dynamics (2.0.1), let us consider the general performance function

$$V(\pi, x) = \sum_{t=0}^{\infty} c_t(x_t, a_t),$$

for which we assume that $V(\pi, x) > -\infty$ for each admissible policy π. We also suppose that there is a policy π such that $V(\pi, x) < \infty$ for every x. Under these conditions we first study the Gâteaux differential of the real valued function $V(\cdot, x)$.

Remark 2.26 (on notation).

(a) Given $\tau = 0, 1, \dots$, a policy

$$\pi = \{a_0, a_1, \dots, a_{\tau-1}, a_\tau, a_{\tau+1}, \dots\},$$

and an action $a \in A$, we define the policy

$$\pi_{-\tau}(a) := \{a_0, a_1, \dots, a_{\tau-1}, a, a_{\tau+1}, \dots\},$$

which is obtained from π by replacing the action a_τ with a.

(b) We will denote by $x_t^{\pi_{-\tau}(a)}$, $t = 0, 1, \dots$, the state path corresponding to $\pi_{-\tau}(a)$. Note that

$$\begin{aligned} x_{t+1}^{\pi_{-\tau}(a)} &:= F_t(x_t, a_t) \quad \text{if} \quad t < \tau, \\ &:= F_\tau(x_\tau, a) \quad \text{if} \quad t = \tau, \\ &:= F_t(x_t^{\pi_{-\tau}(a)}, a_t) \quad \text{if} \quad t > \tau \end{aligned}$$

for all $t = 0, 1, \dots$.

(c) We use the following notation for the product of square matrices $J_1, J_2, \dots$:

$$\begin{aligned} \prod_{k=\tau}^{t} J_k &:= J_\tau \cdots J_t \quad \text{if} \quad \tau \le t, \\ &:= I \quad \text{if} \quad \tau > t, \end{aligned}$$

where I is the identity matrix. ◇

Lemma 2.27. Fix an arbitrary τ in $\{0, 1, \dots\}$ and $y \in \mathbb{R}^m$ (a row vector). Let $\psi^{\tau,y}$ be defined by

$$\psi_i^{\tau,y} := \begin{cases} y & \text{for } i = \tau, \\ 0 & \text{for } i \neq \tau. \end{cases}$$

Then, if it exists, the Gâteaux differential of $V(\cdot, x)$ at π in the direction $\psi^{\tau,y}$ is

$$dV(\pi; \psi^{\tau,y}) = \frac{\partial c_\tau}{\partial a}(x_\tau, a_\tau)y^* + \lambda_{\tau+1}\frac{\partial F_\tau}{\partial a}(x_\tau, a_\tau)y^*,$$

where y^* denotes the *transpose* of $y \in \mathbb{R}^m$, and

$$\lambda_{\tau+1} := \sum_{s=\tau+1}^{\infty} \frac{\partial c_s}{\partial x}(x_s, a_s) \prod_{k=\tau+1}^{s-1} \frac{\partial F_k}{\partial x}(x_k, a_k).$$

Proof. For $\delta \in [0, 1)$ small enough so that $a_\tau + \delta y$ is in an open neighborhood of a_τ, consider the policy $\pi + \delta\psi^{\tau,y} = \pi_{-\tau}(a_\tau + \delta y)$. Then

$$V(\pi + \delta\psi^{\tau,y}, x) = \sum_{s=0}^{\tau-1} c_s(x_s, a_s) + c_\tau(x_\tau, a_\tau + \delta y) + \sum_{s=\tau+1}^{\infty} c_s\left(x_s^{\pi_{-\tau}(a_\tau+\delta y)}, a_s\right).$$

Thus, the Gateaux differential of V at π in the direction $\psi^{\tau,y}$ is given by

$$\begin{aligned} dV(\pi; \psi^{\tau,y}) &= \frac{d}{d\delta} V(\pi + \delta\psi^{\tau,y})\Big|_{\delta=0} \\ &= \frac{\partial c_\tau}{\partial a}(x_\tau, a_\tau)y^* + \frac{d}{d\delta} \sum_{s=\tau+1}^{\infty} c_s\left(x_s^{\pi_{-\tau}(a_\tau+\delta y)}, a_s\right)\Big|_{\delta=0}. \\ &= \frac{\partial c_\tau}{\partial a}(x_\tau, a_\tau)y^* \\ &\qquad + \left(\sum_{s=\tau+1}^{\infty} \frac{\partial c_s}{\partial x}(x_s, a_s) \prod_{k=\tau+1}^{s-1} \frac{\partial F_k}{\partial x}(x_k, a_k)\right) \frac{\partial F_\tau}{\partial a}(x_\tau, a_\tau)y^*, \end{aligned}$$

where the last equality is obtained applying the chain rule inductively. □

From the above computation, we can see that the existence of the Gâteaux differential of $V(\cdot, x)$ requires the following additional hypothesis.

Assumption 2.28. Let $\pi = \{a_0, a_1, \dots\}$ be a policy. For any $\tau = 0, 1, \dots$, there is an open neighborhood U_τ of a_τ such that the sequence of functions $\rho_{\tau,s} : U_\tau \to \mathbb{R}$ given by

$$\rho_{\tau,s}(a) := \frac{\partial c_s}{\partial x}(x_s^{\pi_{-\tau}(a)}, a_s) \prod_{k=\tau+1}^{s-1} \frac{\partial F_k}{\partial x}(x_k^{\pi_{-\tau}(a)}, a_k) \tag{2.3.21}$$

are such that the series $\sum_{s=\tau+1}^{\infty} \rho_{\tau,s}(a)$ converges uniformly in U_τ.

Theorem 2.29. *Let $\pi = \{a_t, t = 0, 1, \dots\}$ be a policy satisfying the Assumption 2.28 and $\{x_t, t = 0, 1, \dots\}$ is the corresponding state trajectory. If π is an optimal policy for the OCP, then there exists a sequence $\{\lambda_t\}_{t=1}^{\infty}$ in $\mathbb{R}^n$ such that:*

1. *for each $t = 1, 2, \dots$,*

$$\frac{\partial c_t}{\partial x}(x_t, a_t) + \lambda_{t+1} \frac{\partial F_t}{\partial x}(x_t, a_t) = \lambda_t, \tag{2.3.22}$$

2. *for each $t = 0, 1, 2, \dots$,*

$$\frac{\partial c_t}{\partial a}(x_t, a_t) + \lambda_{t+1} \frac{\partial F_t}{\partial a}(x_t, a_t) = 0, \tag{2.3.23}$$

3. *Transversality condition (TC): for each $\tau = 0, 1, 2, \dots$,*

$$\lim_{t\to\infty} \lambda_t \prod_{k=\tau}^{t-1} \frac{\partial F_k}{\partial x}(x_k, a_k) = 0. \tag{2.3.24}$$

Proof. Define, for each $t = 1, 2, \dots$,

$$\lambda_t := \sum_{s=t}^{\infty} \frac{\partial c_s}{\partial x}(x_s, a_s) \prod_{k=t}^{s-1} \frac{\partial F_k}{\partial x}(x_k, a_k). \tag{2.3.25}$$

1. A direct calculation gives

$$\begin{aligned} \lambda_t &= \sum_{s=t}^{\infty} \frac{\partial c_s}{\partial x}(x_s, a_s) \prod_{k=t}^{s-1} \frac{\partial F_k}{\partial x}(x_k, a_k) \\ &= \frac{\partial c_t}{\partial x}(x_t, a_t) + \left(\sum_{s=t+1}^{\infty} \frac{\partial c_s}{\partial x}(x_s, a_s) \prod_{k=t+1}^{s-1} \frac{\partial F_k}{\partial x}(x_k, a_k) \right) \frac{\partial F_t}{\partial x}(x_t, a_t) \end{aligned}$$

$$= \frac{\partial c_t}{\partial x}(x_t, a_t) + \lambda_{t+1}\frac{\partial F_t}{\partial x}(x_t, a_t).$$

2. From Lemma 2.27, we have

$$dV(\pi; \psi^{\tau,y}) = \frac{\partial c_\tau}{\partial a}(x_\tau, a_\tau)y^* + \lambda_{\tau+1}\frac{\partial F_\tau}{\partial a}(x_\tau, a_\tau)y^*.$$

Recalling Theorem 1 from Luenberger (1969), p. 178, a necessary condition for π to be a minimizer of V is that $dV(\pi; \psi^{\tau,y}) = 0$ for all $\tau = 0, 1, 2, \ldots$ and every $y \in \mathbb{R}^m$. Hence

$$\frac{\partial c_\tau}{\partial a}(x_\tau, a_\tau) + \lambda_{\tau+1}\frac{\partial F_\tau}{\partial a}(x_\tau, a_\tau) = 0 \text{ for all } \tau = 0, 1, 2, \ldots.$$

3. Notice that

$$\begin{aligned}\lambda_t \prod_{k=\tau}^{t-1} \frac{\partial F_k}{\partial x}(x_k, a_k) &= \left(\sum_{s=t}^{\infty} \frac{\partial c_s}{\partial x}(x_s, a_s) \prod_{k=t}^{s-1} \frac{\partial F_k}{\partial x}(x_k, a_k)\right) \prod_{k=\tau}^{t-1} \frac{\partial F_k}{\partial x}(x_k, a_k) \\ &= \sum_{s=t}^{\infty} \frac{\partial c_s}{\partial x}(x_s, a_s) \prod_{k=\tau}^{s-1} \frac{\partial F_k}{\partial x}(x_k, a_k).\end{aligned}$$

From Assumption 2.28, this series is convergent, and so its tail tends to zero as in (2.3.24). □

Remark 2.30. Let us assume that in Theorem 2.29 the search of an optimal control is restricted to Markov policies $\pi = \{f_t\}$, so that $a_t = f_t(x_t)$ for all $t = 0, 1, \ldots$. If π is optimal for the OPC and $\{x_t\}_{t=0}^{\infty}$ is the corresponding optimal trajectory $x_{t+1} = F_t(x_t, f_t(x_t))$ for all $t = 0, 1, \ldots$, then (2.3.22)–(2.3.24) are rewritten as follows:

1. for each $t = 1, 2, \ldots$,

$$\frac{\partial c_t}{\partial x}(x_t, f_t) + \lambda_{t+1}\frac{\partial F_t}{\partial x}(x_t, f_t) = \lambda_t, \tag{2.3.26}$$

2. for each $t = 0, 1, 2, \ldots$,

$$\frac{\partial c_t}{\partial a}(x_t, f_t) + \lambda_{t+1}\frac{\partial F_t}{\partial a}(x_t, f_t) = 0, \tag{2.3.27}$$

3. Transversality condition (TC): for each $\tau = 0, 1, 2, \ldots$,

$$\lim_{t\to\infty} \lambda_t \prod_{k=\tau}^{t-1} \left[\frac{\partial F_k}{\partial x}(x_k, f_k) + \frac{\partial F_k}{\partial a}(x_k, f_k) \frac{\partial f_k}{\partial x}(x_k) \right] = 0. \quad (2.3.28)$$

Moreover, for each $t = 1, 2, \dots$ we redefine λ_t in (2.3.25) as

$$\lambda_t := \sum_{s=t}^{\infty} \left[\frac{\partial c_s}{\partial x}(x_s, f_s) \prod_{k=t}^{s-1} A_k + \frac{\partial c_s}{\partial a}(x_s, f_s) \frac{\partial f_s}{\partial x}(x_s) \prod_{k=t}^{s-1} B_k \right], \quad (2.3.29)$$

where

$$A_k := \frac{\partial F_k}{\partial x}(x_k, f_k) + \frac{\partial F_k}{\partial a}(x_k, f_k) \frac{\partial f_k}{\partial x}(x_k),$$
$$B_k := \frac{\partial F_k}{\partial x}(x_k, f_k) + \frac{\partial F_k}{\partial a}(x_k, f_k) \frac{\partial f_k}{\partial x}(x_k).$$

The proof is similar to Theorem 2.29. For details see Domínguez-Corella and Hernández-Lerma (2019). ◇

Example 2.31 (Brock–Mirman infinite-horizon model). We next consider an infinite-horizon version of the Brock and Mirman (1972) model in Example 2.10.

Given an initial state x_0, the system evolves according to the dynamics

$$x_{t+1} = cx_t^\theta - a_t, \qquad t = 0, 1, 2, \dots, \quad (2.3.30)$$

with $\theta \in (0, 1)$. We consider again the control constraint sets $A(x) := (0, cx^\theta]$. The performance index, for a feasible policy

$$\pi = \{a_t, \quad t = 0, 1, \dots\},$$

is given by

$$V(\pi, x_0) = \sum_{t=0}^{\infty} \alpha^t \log(a_t).$$

Thus the minimum principle conditions (2.3.26) and (2.3.27) become

$$\lambda_{t+1} c\theta x_t^{\theta-1} = \lambda_t, \qquad t = 1, 2, \dots, \quad (2.3.31)$$

$$\frac{\alpha^t}{a_t} - \lambda_{t+1} = 0, \qquad t = 0, 1, \dots. \quad (2.3.32)$$

This difference equations can be solved by the "guess and verify method" as in Chow (1997). To this end, consider a policy $a_t := dcx_t^\theta$ for some real number d. Then combining (2.3.31) and (2.3.32) we obtain

$$\lambda_t = \lambda_{t+1} c\theta x_t^{\theta-1} = \frac{\alpha^t}{a_t} c\theta x_t^{\theta-1} = \frac{\alpha^t c\theta x_t^{\theta-1}}{dcx_t^\theta} = \frac{\theta\alpha^t}{dx_t}.$$

This implies that

$$\lambda_{t+1} = \frac{\theta\alpha^{t+1}}{dx_{t+1}} = \frac{\theta\alpha^{t+1}}{d(cx_t^\theta - a_t)} = \frac{\theta\alpha^{t+1}}{d(cx_t^\theta - dcx_t^\theta)}.$$

On the other hand, from (2.3.32), $\lambda_{t+1} = \frac{\alpha^t}{dcx_t^\theta}$. Equating both values, we conclude that $d = 1 - \theta\alpha$. Then an optimal policy is

$$a_t^* = c(1-\theta\alpha)[x_t^*]^\theta \quad \text{for} \quad t = 0, 1, 2, \ldots. \tag{2.3.33}$$

◇

Remark 2.32. From (2.3.33), and in accordance with Corollary 2.22, the optimal stationary policy for the Brock and Mirman model for the infinite horizon case is $f^*(x) = c(1-\theta\alpha)x^\theta$ and the value function is

$$V^*(x) = \frac{1}{1-\alpha}\log[c(1-\theta\alpha)] + \frac{\theta\alpha}{(1-\alpha)(1-\theta\alpha)}\log(c\theta\alpha) + \frac{\theta}{1-\theta\alpha}\log(x).$$

This function satisfies the DP equation (2.3.8). On the other hand, in the spirit of Theorem 2.21, we can deduce that the optimal policy (2.2.16) for the finite horizon case converges to the optimal policy (2.3.33) for the infinite horizon problem. ◇

2.3.3 The Weighted-Norm Approach

In Sect. 2.3.1 we studied the infinite-horizon discounted OCP (2.3.2)–(2.3.3) assuming that the stage cost $c(x, a)$ is possibly unbounded, but *nonnegative*. This nonnegativity yields that the infinite series (2.3.3) is well defined (although it might be infinite). In general, however, considering costs $c(x, a)$ with both positive

and negative values may create technical complications. To avoid them, in this section we consider *weighted norms*, an approach introduced by Wessels (1977). This approach indeed allows $c(x,a)$ to take positive and negative values, but its "growth" is restricted in a suitable sense (see Assumption 2.33(c) and Lemma 2.34(b)). The basis of this approach is in the following conditions.

Assumption 2.33. For every $x \in X$:

(a) the control constraint set $A(x)$ is compact, and the set-valued mapping $x \mapsto A(x)$ is continuous. (It suffices to assume that $x \mapsto A(x)$ is u.s.c.)

(b) For $(x,a) \in \mathbb{K}$, with $\mathbb{K}$ in (2.3.12), the function $(x,a) \mapsto F(x,a)$ is continuous, and $(x,a) \mapsto c(x,a)$ is l.s.c.

(c) There is a continuous function $w(\cdot) \geq 1$ on X, and positive constants $\bar{c}$ and $\beta \geq 1$ such that, for every $x \in X$,

(c1) $\sup_{a \in A(x)} |c(x,a)| \leq \bar{c} w(x)$, and

(c2) $\sup_{a \in A(x)} w(F(x,a)) \leq \beta w(x)$, and $\alpha\beta < 1$.

Assumption 2.33 is supposed to hold throughout this section. The function w in part (c) will be referred to as a *weight function* , but in the control literature is also known as a *majorant*, a *bounding function* or a *gauge function*. Part (c2) is called *Wessels condition*. As an example, if the stage cost c is *bounded*, that is, $|c(x,a)| \leq \bar{c}$ for some constant $\bar{c}$ and all $(x,a) \in \mathbb{K}$, then we may take $w \geq 1$ as a bounded function. On the other hand, a common situation is when c satisfies a *polynomial growth condition*, in which case we may take w as $w(x) := D(1 + |x|^k)$. Another common situation is the *exponential growth* case, with $w(x) := De^{k|x|}$.

The proof of the following lemma is left to the reader (Exercise 2.12). Observe that parts (a) and (b) are a direct consequence of Assumption 2.33(c) and an induction argument. Part (c) in the lemma follows from (a)-(b) and the definition (2.3.3) of $V(\pi,\cdot)$.

Lemma 2.34. Let $\{(x_t, a_t), t = 0, 1, ...\}$ be an arbitrary sequence in $\mathbb{K}$ for which (2.3.2) holds, that is, $x_{t+1} = F(x_t, a_t)$ for all $t = 0, 1, ...$ Then, for every initial state $x_0 = x$ and $t = 0, 1, ...,$

(a) $w(x_t) \leq \beta^t w(x)$; and

(b) $|c(x_t, a_t)| \leq \bar{c}\beta^t w(x)$.
(c) For any control policy $\pi = \{a_t, t = 0, 1, ...\}$ and $x_0 = x$,

$$V(\pi, x) \leq \bar{c}\frac{w(x)}{(1-\gamma)},$$

where $\gamma := \alpha\beta$. (By Assumption 2.33(c2), $\gamma < 1$.)

In this section we will be working in the spaces $M_w(X)$ and $L_w(X)$ defined in the following lemma.

Lemma 2.35. (a) The space $M_w(X)$ of real-valued functions v on X with a finite w-norm, which is defined as

$$\|v\|_w := \sup_{x\in X} \frac{|v(x)|}{w(x)}, \tag{2.3.34}$$

is a Banach space.

(b) The space $L_w(X) := L(X) \cap M_w(X)$ of l.s.c. functions in $M_w(X)$ is a complete metric space with the metric induced by the w-norm, that is, $\text{dist}(v, v') := \|v - v'\|_w$.

Proof. Part (a) is straightforward and is left to the reader (Exercise 2.13). To prove (b), let v_n be a sequence in $L_w(X)$ that converges in the w-norm to a function v. By part (a), v belongs to $M_w(X)$. Thus, to complete the proof of (b) it only remains to show that v is l.s.c. To this end, first observe that

$$v(x) = [v(x) - v_n(x)] + v_n(x) \geq -\|v_n - v\|_w w(x) + v_n(x)$$

for all $x \in X$ and $n = 0, 1, ...$ Now consider a sequence $x^k \to x$, and fix an arbitrary n. Then, since v_n is l.s.c. and w is continuous (by Assumption 2.33(c)),

$$\liminf_{k\to\infty} v(x^k) \geq -\|v_n - v\|_w w(x) + v_n(x).$$

Finally, letting n tend to ∞ we obtain $\liminf_k v(x^k) \geq v(x)$. That is, v is l.s.c. □

In the following proposition we state the *Blackwell conditions* (Blackwell 1965) for an operator to be a contraction. They will

be used below in connection with *Banach's fixed point theorem* in Remark 2.25 in the case that the metric space $\mathcal{X}$ is $L_w(X)$.

Proposition 2.36. Let $T : L_w(X) \to L_w(X)$ be a mapping such that:

(a) T is monotone, that is, if u and v are functions in $L_w(X)$ and $u \le v$, then $Tu \le Tv$; and
(b) there is a positive number $\delta < 1$ such that, for every $v \in L_w(X)$ and every constant $k \in \mathbb{R}$, $T(v + kw) \le Tv + \delta kw$.

Then T is a contraction with modulus δ.

Proof. For any two functions v, v' in $L_w(X)$,

$$v \le v' + |v - v'| \le v' + w\|v - v'\|_w.$$

Therefore, by (a)-(b), with $k = \|v - v'\|_w$,

$$Tv - Tv' \le \delta w\|v - v'\|_w.$$

Interchanging v and v', and then combining with the latter inequality we obtain

$$|Tv - Tv'| \le \delta w\|v - v'\|_w.$$

This implies the desired result. □

To state our main result, Theorem 2.38, we will use the following fact.

Proposition 2.37. Let $x \mapsto A(x)$ and w be as in Assumption 2.33. (It suffices to take $x \mapsto A(x)$ u.s.c.). Suppose that v is a l.s.c. function on $\mathbb{K}$ and such that, for some constant k

$$\sup_{a \in A(x)} |v(x, a)| \le kw(x) \quad \forall x \in X.$$

Then there exists $f \in \mathbb{F}$ such that, for all $x \in X$,

$$v^*(x) := \inf_{a \in A(x)} v(x, a) = v(x, f(x)) \tag{2.3.35}$$

and, moreover, v^* is in $L_w(X)$ with w-norm $\|v^*\|_w \le k$.

Proof. The proposition follows from Theorem B.3(b) applied to the nonnegative l.s.c. function $v'(x,a) := v(x,a) + kw(x)$. □

The following Theorem 2.38 is the main result of the weighted-norm approach to the infinite-horizon discounted OCP (2.3.2)–(2.3.4). It gives the dynamic programming equation (2.3.7), which we already obtained in Theorem 2.21 when the stage cost c is *nonnegative*. In the present case, however, we also obtain the convergence estimate (2.3.37), which is impossible to obtain in Theorem 2.21.

Theorem 2.38. *Under the Assumption 2.33 the following holds:*

(a) *The α-discount value function V^* is the unique solution in $L_w(X)$ of the dynamic programming equation (2.3.7), i.e., for every $x \in X$,*

$$V^*(x) = \inf_{a \in A(x)} [c(x,a) + \alpha V^*(F(x,a))]. \tag{2.3.36}$$

Moreover, for every $n = 1, 2, ...,$

$$\|V_n^* - V^*\|_w \le \bar{c}\gamma^n/(1-\gamma), \tag{2.3.37}$$

where V_n^* is the VI function in (2.3.10)–(2.3.11), and the constants $\bar{c}$ and $\gamma := \alpha\beta < 1$ come from Assumption 2.33.

(b) *There exists $f^* \in \mathbb{F}$ such that, for every $x \in X$, $f^*(x) \in A(x)$ attains the minimum in the right-hand side of (2.3.36), i.e. (using the notation in Remark 2.19),*

$$V^*(x) = c(x, f^*) + \alpha V^*(F(x, f^*)), \tag{2.3.38}$$

and f^* is α-discount optimal.

Proof. (a) To prove (2.3.36) we will show that, equivalently, V^* is the unique fixed-point in $L_w(X)$ of the Bellman operator K in (2.3.8). To this end, we will prove that K is a *contraction operator* on the complete metric space $L_w(X)$, so (2.3.36)–(2.3.37) will follow from Banach's fixed-point theorem (Remark 2.25).

First, we need to show that indeed K maps $L_w(X)$ into itself. To do this, pick an arbitrary function v in $L_w(X)$. Hence, since v is l.s.c. and $(x,a) \mapsto F(x,a)$ is continuous (Assumption 2.33(b)), the

function $(x,a) \mapsto v(F(x,a))$ is l.s.c on $\mathbb{K}$. In addition, by Assumption 2.33(b), $(x,a) \mapsto c(x,a)$ is l.s.c., so $v'(x,a) := c(x,a) + \alpha v(F(x,a))$ is l.s.c. on $\mathbb{K}$. On the other hand, by definition (2.3.34) of the w-norm and Assumption 2.33(c2), we have

$$|v(F(x,a))| \leq \|v\|_w w(F(x,a)) \leq \beta \|v\|_w w(x).$$

This inequality together with Assumption 2.33(c1) on c and Proposition 2.37 give that $Kv(x) := \inf_{a \in A(x)} v'(x,a)$ is in $L_w(X)$. To conclude, K maps $L_w(X)$ into itself.

Now, to prove that K is a contraction operator on $L_w(X)$ we will verify the Blackwell conditions in Proposition 2.36. First note that, obviously, K is monotone. On the other hand, for any function $v \in L_w(X)$ and any constant k, Assumption 2.33(c2) gives

$$\begin{aligned} K(v + kw)(x) &= \inf_a [c(x,a) + \alpha v(F(x,a)) + \alpha k w(F(x,a))] \\ &\leq Kv(x) + k\gamma w(x) \end{aligned}$$

with $\gamma := \alpha\beta < 1$. Therefore, by Proposition 2.36, K is a contraction with modulus γ. It follows that K has a unique fixed point v^* in $L_w(X)$, i.e., $v^* = Kv^*$ or, more explicitly,

$$v^*(x) = \inf_{a \in A(x)} [c(x,a) + \alpha v^*(F(x,a))] \tag{2.3.39}$$

for all $x \in X$.

To complete the proof of part (a) we will show that $v^* = V^*$, the OCP's value function. First, by Proposition 2.37, there exists $f \in \mathbb{F}$ that minimizes the right-hand side of (2.3.39), i.e.,

$$v^*(x) = c(x,f) + \alpha v^*(F(x,f)) \quad \forall x \in X.$$

Iteration of this equality gives

$$v^*(x) = \sum_{t=0}^{n-1} c(x_t, f) + \alpha^n v^*(x_n) \tag{2.3.40}$$

for all $x \in X$ and $n = 1, 2, \ldots$. In addition, the last term $\alpha^n v^*(x_n)$ tends to zero as $n \to \infty$. In fact, for any function v in $L_w(X)$, Lemma 2.34(a) yields

$$\alpha^n|v(x_n)| \le \alpha^n\|v\|_w w(x_n) \le \gamma^n\|v\|_w w(x) \to 0 \qquad (2.3.41)$$

as $n \to \infty$. Thus, for every $x \in X$, (2.3.40) yields $v^*(x)=V(f,x) \ge V^*(x)$, i.e.,

$$v^*(\cdot) \ge V^*(\cdot). \qquad (2.3.42)$$

To obtain the reverse inequality, note that (2.3.39) gives that

$$v^*(x) \le c(x,a) + \alpha v^*(F(x,a)) \quad \forall (x,a) \in \mathbb{K}.$$

Therefore, for any policy $\pi = a_t$ and any initial state $x_0 = x \in X$,

$$v^*(x_t) \le c(x_t,a_t) + \alpha v^*(x_{t+1}),$$

so, for every $n = 1, 2, ...,$

$$v^*(x) \le \sum_{t=0}^{n-1} c(x_t,a_t) + \alpha^n v^*(x_n).$$

Finally, letting $n \to \infty$ and using (2.3.41) again we obtain

$$v^*(x) \le V(\pi,x) \quad \forall x \in X.$$

Thus, since π was arbitrary, it follows that $v^*(\cdot) \le V^*(\cdot)$. This inequality and (2.3.42) give that $v^* = V^*$. This fact together with Banach's fixed-point theorem complete the proof of part (a).

(b) This part follows from Proposition 2.37, as in (2.3.39)–(2.3.41). □

In the following section we will present some applications of the weighted-norm approach.

2.4 Approximation Algorithms

Solving a dynamic programming equation (DPE), such as (2.3.7), is in general a difficult task. Richard Bellman, in his book Bellman (1957a) coined the term "the curse of dimensionality" to refer to the fact that the difficulty in solving a DPE rapidly increases with the number of dimensions.

On the other hand, there are two natural ways in which we can approximate the solution of a DPE, namely, the *value iteration* (VI) and the *policy iteration* (PI) algorithms. These algorithms are difficult to compare because their performances highly depend on the particular features of the OCPs being dealt with. In either case, however, they are the basis for several useful approaches in so-called *adaptive dynamic programming* and *reinforcement learning* to obtain approximate solutions to a DPE and analyze related issues, such as the "stabilizability property" of an optimal control.

First, we will consider the VI algorithm (Sect. 2.4.1), and then the PI algorithm (Sect. 2.4.2). The VI approach is also known as the method of *successive approximations*. See Wessels (1977).

2.4.1 Value Iteration

Let K be the Bellman operator in (2.3.8), and consider the VI functions

$$V_n^* = KV_{n-1}^* = K^n V_0^* \quad \text{for} \quad n = 1, 2, ...$$

in (2.3.11), with $V_0^* \equiv 0$. The *VI algorithm* refers to Problem 1 in Sect. 2.2, that is, the convergence $V_n^* \to V^*$ in (2.3.6). This convergence was already obtained in Theorems 2.21 and 2.38 under two different sets of assumptions. In particular, Theorem 2.21 requires the stage cost $c(x, a)$ to be nonnegative, whereas Theorem 2.38 uses a weighted-norm approach. In this section we analyze additional properties of the VI algorithm, which concern the VI control policies defined as follows.

Definition 2.39. A sequence $\pi_{VI} = \{f_n, \quad n = 1, 2, ...\} \subset \mathbb{F}$ is called a *VI policy* if, for every $n = 1, 2, ...$, $f_n \in \mathbb{F}$ minimizes the right-hand side of (2.3.10), i.e.,

$$V_n^*(x) = c(x, f_n) + \alpha V_{n-1}^*(F(x, f_n)) \quad \forall x \in X. \tag{2.4.1}$$

We will denote by $\mathbb{F}_n \subset \mathbb{F}$ the subfamily of selectors that satisfy (2.4.1). (Recall from Remark 2.14 that $V_0^*(\cdot) \equiv 0$. Thus, for π_{VI} to be a true control policy we may take f_0 as any selector in $\mathbb{F}$.)

To analyze a VI policy we will use the *discrepancy function* $D : \mathbb{K} \to \mathbb{R}$ defined as

$$D(x,a) := c(x,a) + \alpha V^*(F(x,a)) - V^*(x) \tag{2.4.2}$$

for $(x,a) \in \mathbb{K}$. By the DPE (2.3.7), D is *nonnegative*. Moreover, we can rewrite (2.3.7) as

$$\inf_{a \in A(x)} D(x,a) = 0 \quad \forall x \in X. \tag{2.4.3}$$

Similarly, an equality such as (2.3.38) becomes

$$D(x, f^*) = 0 \quad \forall x \in X, \tag{2.4.4}$$

where, as in the Remark 2.19, $D(x, f^*) \equiv D(x, f^*(x))$.

The name discrepancy function comes from the fact that, for any policy $\pi = \{a_t\}$, we can express the difference or "discrepancy" between $V(\pi, \cdot)$ and $V^*(\cdot)$ in terms of D; in fact, for any initial state $x_0 = x$,

$$V(\pi, x) - V^*(x) = \sum_{t=0}^{\infty} \alpha^t D(x_t, a_t). \tag{2.4.5}$$

(See Exercise 2.14) Results such as (2.4.4) and (2.4.5) motivate the following definition.

Definition 2.40. A Markov policy $\pi = \{f_n\}$ is said to be *asymptotically optimal* (for the discounted cost criterion) if, for every $x \in X$,

$$D(x, f_n) \to 0 \quad \text{as} \quad n \to \infty. \tag{2.4.6}$$

The concept of asymptotic optimality was introduced in *adaptive Markov control processes* (adaptive MCPs), which are MCPs that depend on *unknown* parameters, say, θ. In this case, at each time n, the controller computes an estimate θ_n of θ, and then he/she *adapts* his/her control action f_n to this estimate. Thus (in view of (2.4.4)), (2.4.6) holds if f_n is approximating in some sense an optimal control. Alternatively, asymptotic optimality can be used to analyze "approximations" to the OCP (2.3.2)–(2.3.3).

(See, for instance, Sect. 4.6 in Hernández-Lerma and Lasserre (1996).)

By (2.3.6), a VI policy is a natural candidate to be asymptotically optimal. This is not necessarily true, however, in the generality of Theorem 2.21. On the other hand, in the weighted-norm context of Theorem 2.38 we obtain the following nice result.

Proposition 2.41. Suppose that Assumption 2.33 holds, and let $\pi_{VI} = \{f_n\}$ be a VI policy. Then, for every $x \in X$ and $n = 1, 2, ...,$

$$0 \leq D(x, f_n) \leq 2\bar{c}\gamma^n w(x)/(1-\gamma) \tag{2.4.7}$$

with $\bar{c}, w(\cdot)$, and $\gamma := \alpha\beta$ as in Assumption 2.33. Hence π_{VI} is asymptotically optimal.

Proof. From (2.4.2) and (2.4.1),

$$\begin{aligned} D(x, f_n) &= c(x, f_n) + \alpha V^*(F(x, f_n)) - V^*(x) \\ &= V_n^*(x) - V^*(x) + \alpha[V^*(F(x, f_n)) - V_{n-1}^*(F(x, f_n))] \end{aligned}$$

for all $x \in X$ and $n = 1, 2, ...$ Thus (2.4.7) follows from (2.3.37). □

Note that (2.4.7) gives an a priori bound for $D(x, f_n)$ in the sense that the right-hand does not depend on neither V^* nor V_n^*. Now consider the following situation: fix an arbitrary $n = 1, 2, ...$ and let $f_n \in \mathbb{F}_n$ be the corresponding selector in (2.4.1). Furthermore, consider the *stationary Markov policy* $\pi_n = \{g_t\} \subset \mathbb{F}$ such that $g_t \equiv f_n$ for all $t = 0, 1, ...$ In other words, we apply the same control f_n at every stage $t = 0, 1, ...$ Then (2.4.8) below states that π_n can be made "arbitrarily close to an optimal policy" if n is large enough. (See also Proposition 2.43 below.)

Proposition 2.42. With π_n as above,

$$0 \leq V(\pi_n, x) - V^*(x) \leq 2\bar{c}\gamma^n w(x)/(1-\gamma) \tag{2.4.8}$$

for all $x \in X$.

Proof. Fix $x \in X$, and consider

$$V(\pi_n, x) - V^*(x) \leq |V(\pi_n, x) - V_n^*(x)| + |V_n^*(x) - V^*(x)|.$$

By (2.3.37), the second term on the right satisfies that

$$|V_n^*(x) - V^*(x)| \leq \bar{c}\gamma^n w(x)/(1-\gamma). \tag{2.4.9}$$

Now, by definition of π_n, $V(\pi_n, x)$=$c(x, f_n)$+$\alpha V(\pi_n, F(x, f_n))$. Combining this equation with (2.4.1) we obtain

$$|V(\pi_n, x) - V_n^*(x)| \leq \alpha |V(\pi_n, F(x, f_n)) - V_{n-1}^*(F(x, f_n))|.$$

Iteration of this inequality (and recalling that $V_0^*(\cdot) \equiv 0$) gives

$$\begin{aligned} |V(\pi_n, x) - V_n^*(x)| &\leq \alpha^n |V(\pi_n, x_n)| \\ &\leq \bar{c}\alpha^n w(x_n)/(1-\gamma) \quad \text{[by Lemma 2.34(c)]} \\ &\leq \bar{c}\gamma^n w(x)/(1-\gamma) \quad \text{[by Lemma 2.34(a)].} \end{aligned}$$

This inequality and (2.4.9) give (2.4.8). □

Results such as Propositions 2.41 or 2.42 obviously suggest that the VI selectors $f_n \in \mathbb{F}_n$ might converge to an α-optimal control $f^* \in \mathbb{F}$. This is not necessarily true, but we can ensure the following.

Proposition 2.43. Suppose that the hypotheses of Theorem 2.38 are satisfied, and consider a VI policy $\pi_{VI} = \{f_n, \ n = 0, 1, \dots\}$. Then there exists $f^* \in \mathbb{F}$ such that f_n converges to f^* in the sense of Schäl (1975); that is, for each $x \in X$, there is a sequence $n_i = n_i(x)$ such that $f_{n_i}(x) \to f^*(x)$ as $i \to \infty$.

The proof of Proposition 2.43 follows from Proposition B.12 with $v_n \equiv V_n^*$ and $v^* \equiv V^*$ in (2.4.1) and (2.3.38), respectively.

2.4.2 Policy Iteration

The *policy iteration* (PI) algorithm is also known as Howard's *policy improvement* method. The algorithm was introduced by Howard (1960) for a class of discrete-time Markov decision processes with finite state space and finite action sets. Nevertheless, it soon became evident that the PI algorithm could be extended to many classes of OCPs, including all those considered in these

lectures. On the other hand, a key difference with respect to the VI algorithm is that PI gives a *monotone sequence* of functions converging to the optimal value function $V^*(\cdot)$, even if the stage costs $c(x,a)$ take positive and negative values!

The PI algorithm is based on the "monotonicity property" of the Bellman operator K (2.3.8) established in Lemma 2.20, under the assumptions of Theorem 2.21. In the context of Theorem 2.38 the same proof of Lemma 2.20 yields the following.

Lemma 2.44. Suppose that the Assumption 2.33 holds. If $v \in L_w(X)$ is such that $v \geq Kv$, then

(a) There exists $f \in \mathbb{F}$ such that $v(x) \geq V(f,x)$ for all $x \in X$; and, therefore,
(b) $v \geq V^*$.

Now suppose that the conditions of Lemma 2.20 or Lemma 2.44 are satisfied. Consider an arbitrary selector $g_0 \in \mathbb{F}$, and let $v_0(\cdot) := V(g_0,\cdot)$ be the corresponding discounted cost. Then (as in the Remark 2.19)

$$\begin{aligned} v_0(x) &= c(x,g_0) + \alpha v_0(F(x,g_0)) \\ &\geq \inf_{a\in A(x)} [c(x,a) + \alpha v_0(F(x,a))], \end{aligned} \tag{2.4.10}$$

so $v_0 \geq Kv_0$. Therefore, by Lemma 2.20(a) or Lemma 2.44(a), there exists $g_1 \in \mathbb{F}$ such that $v_0 \geq v_1$, where $v_1(x) := V(g_1,x)$ for all $x \in X$. Next, in (2.4.10) replace v_0, g_0 by v_1, g_1 and repeat the same argument to obtain $g_2 \in \mathbb{F}$ and $v_2(\cdot) := V(g_2,\cdot)$, with $v_1 \geq v_2$. In general, the **PI algorithm** is as follows, with $n = 0, 1, \ldots$

(PI_1) Given $g_n \in \mathbb{F}$, compute the discounted cost $v_n(\cdot) \equiv V(g_n,\cdot)$. Then, for all $x \in X$,

$$v_n(x) = c(x,g_n) + \alpha v_n(F(x,g_n)) \geq Kv_n(x). \tag{2.4.11}$$

(PI_2) *Policy improvement*: Find $g_{n+1} \in \mathbb{F}$ such that

$$Kv_n(x) = c(x,g_{n+1}) + \alpha v_n(F(x,g_{n+1})) \quad \forall x \in X;$$

so $v_n \geq v_{n+1}$, where $v_{n+1}(\cdot) \equiv V(g_{n+1},\cdot)$. Replace n by $n+1$ and go back to step (PI_1).

Theorem 2.45. *Suppose that the hypotheses of Theorem 2.21 or Theorem 2.38 are satisfied, and let v_n be as in the PI algorithm. Then:*

(a) *If there exists n for which $v_n(x) = v_{n+1}(x)$ for all $x \in X$, then the function $v^*(\cdot) \equiv v_n(\cdot)$ satisfies the DPE $v^* = Kv^*$. Moreover, $v^* = V^*$ and g_n is an optimal control.*

(b) *In general, as $n \to \infty$, $v_n \downarrow v^*$, where v^* is a solution of the DPE, and $v^* = V^*$.*

Proof. (a) Let $v^*(\cdot) := v_n(\cdot) \equiv v_{n+1}(\cdot)$. Then, by (PI_1) and (PI_2), $v^* \geq Kv^* \geq v^*$, so v^* satisfies the DPE. This completes the proof of part (a) under the assumptions of Theorem 2.38. Similarly, in Theorem 2.21 V^* is the *minimal* solution of the DPE. Therefore, if $v^* \neq V^*$, then, by (PI_2), there exists a control that "improves" $v* = v_{n+1}$. This is a contradiction.

(b) By construction, the functions v_n form a nondecreasing sequence, which (by Lemma 2.20 or Lemma 2.44) is bounded below by V^*. Therefore, $v_n \downarrow v^*$ for some function $v^* \geq V^*$. Moreover, by Lemma 2.15(a) and (2.4.11),

$$v^* \geq Kv^*. \tag{2.4.12}$$

On the other hand, for all $x \in X$ and $n = 0, 1, ...,$

$$\begin{aligned} v^*(x) &\leq v_n(x) \\ &= c(x, g_n) + \alpha v_n(F(x, g_n)) \\ &\leq c(x, g_n) + \alpha v_{n-1}(F(x, g_n)) \\ &= Kv_{n-1}(x). \end{aligned}$$

Hence, by definition of K in (2.3.8),

$$v^*(x) \leq c(x, a) + \alpha v_{n-1}(F(x, a)) \quad \forall\ (x, a) \in \mathbb{K}.$$

As $n \to \infty$, the latter inequality yields, $v^*(x) \leq c(x, a) + \alpha$ $v^*(F(x, a))$, which in turn gives $v^* \leq Kv^*$. This inequality and (2.4.12) give that v^* satisfies the DPE $v^* = Kv^*$. The last statement in part (b) is obtained as in (a). □

Note that Proposition 2.43 remains true if we replace "VI policy" by "PI policy". For a more general statement of this proposition, see Theorem B.10 or Proposition B.12 in Appendix B.

Example 2.46. Consider the LQ control problem consisting of the linear system

$$x_{t+1} = \delta x_t + \eta a_t, \quad t = 0, 1, ..., \tag{2.4.13}$$

with initial state $x_0 = x$, nonzero coefficients δ, η, and a quadratic stage cost

$$c(x, a) = qx^2 + ra^2 \quad \text{for } x, a \in X = A = \mathbb{R} \tag{2.4.14}$$

with $q \geq 0$ and $r > 0$. Hence the OCP is to minimize the α-discounted cost

$$V(\pi, x) = \sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \tag{2.4.15}$$

subject to (2.4.13). Given this OCP, develop:

(a) the VI algorithm, and
(b) a PI algorithm.
(c) Solve the LQ problem (2.4.13)–(2.4.15) by means of the "guess and verify" approach, which is also known as "the method of undetermined coefficients".

Solution of (a). To simplify the notation, we will write the VI functions V_n^* as v_n. Hence, (2.3.10) becomes

$$v_n(x) = \min_{a \in A(x)} [c(x, a) + \alpha v_{n-1}(F(x, a))] \tag{2.4.16}$$

for all $x \in X$ and $n = 1, 2, ...$, with $v_0(\cdot) \equiv 0$. Thus, for $n = 1$, (2.4.16) gives

$$v_1(x) = \min_a (qx^2 + ra^2) = qx^2 \quad \forall x \in X,$$

and the minimum is attained at $a^* = f_1(x) = 0$ for all x. Similarly,

$$\begin{aligned} v_2(x) &= \min_a [qx^2 + ra^2 + \alpha v_1(\delta x + \eta a)] \\ &= \min_a [qx^2 + ra^2 + \alpha q(\delta x + \eta a)^2] \end{aligned} \tag{2.4.17}$$

Computating the derivative of the right-hand side with respect to a, and then equating to 0, gives that the minimizer is

$$f_2(x) = -(r + \alpha q \eta^2)^{-1} \alpha q \delta \eta \cdot x.$$

Replacing $a = f_2(\cdot)$ in (2.4.17), v_2 becomes the quadratic function $v_2(x) = C_2 x^2$, with coefficient

$$C_2 := \frac{qr + (r\delta^2 + q\eta^2)\alpha q}{r + \alpha q \eta^2}.$$

In general, by induction we can see that, for every $x \in X$,

$$v_n(x) = C_n x^2 \quad \forall n = 0, 1, ..., \tag{2.4.18}$$

with $C_0 = 0$, and VI controls

$$f_n(x) = -(r + SC_{n-1})^{-1} \alpha \delta \eta C_{n-1} \cdot x \tag{2.4.19}$$

for all $n = 1, 2, ...$, with $f_0 \in \mathbb{F}$ arbitrary (since $v_0 \equiv 0$), and

$$C_n = \frac{P + QC_{n-1}}{r + SC_{n-1}} \quad \forall n = 1, 2, \ldots. \tag{2.4.20}$$

Here

$$P := qr, \quad Q := (r\delta^2 + q\eta^2)\alpha, \quad S := \alpha\eta^2.$$

By means of some technical arguments (which can be seen, for instance, in Dynkin and Yushkevich (1979), Sect. 2.11, or Hernández-Lerma and Lasserre (1996), Sect. 4.7) based on the fixed-point approach to (2.4.20), it can be seen that $C_n \to C$ as $n \to \infty$, where $C = z$ is the unique positive solution of the quadratic equation

$$z = \frac{P + Qz}{r + Sz},$$

which we may be rewrite as

$$Sz^2 + (r - Q)z - P = 0. \tag{2.4.21}$$

The unique positive solution is

$$C = [-(r - Q) + ((r - Q)^2 + 4PS)^{1/2}]/2S.$$

Finally, from Theorem 2.21 and (2.4.18) we conclude that the α-optimal discounted cost V^* is given by

$$V^*(x) = \lim_{n\to\infty} v_n(x) = Cx^2. \tag{2.4.22}$$

Moreover, from (2.4.19) and Proposition B.12(a) (in Appendix B),

$$\begin{aligned} f^*(x) :&= \lim_{n\to\infty} f_n(x) \\ &= -(r + SC)^{-1}\alpha\delta\eta Cx \end{aligned} \tag{2.4.23}$$

is an α-optimal selector.

Solution of (b). In the policy iteration (PI) algorithm, first, we take an arbitrary control $g_0 \in \mathbb{F}$ and then we compute the corresponding discounted cost $v_0(\cdot) = V(g_0, \cdot)$. (See (2.4.10).) Since we are not given an indication about how to choose g_0, we may select it so that v_0 is easy to compute. This is the case if we choose $g_0(\cdot) \equiv 0$ (which is the same as f_1 in part (a) above). Thus, with $a_t = 0$ for all $t = 0, 1, ...$, the LQ system (2.4.13)–(2.4.15) becomes

$$x_{t+1} = \delta x_t = \delta^{t+1}x \quad \forall t = 0, 1, ...,$$

given the initial state $x_0 = x$, and the stage cost $c(x, a) = qx^2$. Therefore, the corresponding α-discounted cost is

$$v_0(x) = q\sum_{t=0}^{\infty} \alpha^t x_t^2 = D_0 x^2$$

with coefficient $D_0 := q/(1 - \alpha\delta^2)$, assuming that $\alpha\delta^2 < 1$. Having v_0, we then proceed to the *policy improvement* step. That is, we wish to find $g_1 \in \mathbb{F}$ such that, for all $x \in X$, $g_1(x) \in A$ attains the minimum in the right-hand side of the inequality

$$\begin{aligned} v_0(x) &\geq \min_a [c(x, a) + \alpha v_0(F(x, a))] \\ &= \min_a [qx^2 + ra^2 + \alpha D_0(\delta x + \eta a)^2]. \end{aligned}$$

Now, the usual calculations (computing the derivative with respect to a, equating to 0, and so on) give that $g_1(x) = -G_1 x$,

with coefficient

$$G_1 := \frac{\alpha\delta\eta D_0}{r + \alpha\eta^2 D_0}.$$

From (2.4.15), the corresponding α-discounted cost $v_1(x) = V(g_1, x)$ is $v_1(x) = D_1 x^2$, where

$$D_1 := \frac{q + rG_1^2}{1 - \alpha(\delta - \eta G_1)^2},$$

assuming that $|\alpha(\delta - \eta G_1)^2| < 1$. In general, we obtain by induction that, for all $x \in X$ and $n = 0, 1, ...,$

$$g_n(x) = -G_n x \quad \text{and} \quad v_n(x) = D_n x^2, \tag{2.4.24}$$

with coefficients $G_0 = 0$,

$$D_n = \frac{q + rG_n^2}{1 - \alpha(\delta - \eta G_n)^2} \tag{2.4.25}$$

and

$$G_{n+1} = \frac{\alpha\delta\eta D_n}{r + \alpha\eta^2 D_n} \tag{2.4.26}$$

for $n = 0, 1, ...$

Since the functions $v_n(x) = D_n x^2$ form a nonnegative nonincreasing sequence (Theorem 2.45), there is a nonnegative function $v^*(x) = D^* x^2$ such that $v_n(x) \downarrow v^*(x)$ for all $x \in X$. In particular, as $n \to \infty$, $D_n \to D^*$ and, therefore, from (2.4.26), $G_n \to G^*$, with

$$\begin{aligned} G^* &= \frac{\alpha\delta\eta D^*}{r + \alpha\eta^2 D^*} \\ &= \frac{\alpha\delta\eta D^*}{r + SD^*}, \end{aligned}$$

which is the same as the coefficient of (2.4.23), with $C = D^*$. Finally, from (2.4.25) and (2.4.22) we conclude that $v^*(\cdot) = V^*(\cdot)$ with $C = D^*$.

Solution of (c). We now wish to solve the infinite-horizon LQ problem (2.4.13)–(2.4.15) by the "guess and verify" approach. The idea is to "guess" that the solution of the DPE has a certain form

and then we verify that this is indeed the case. In the present LQ problem, from the finite-horizon case en Example 2.4 we know that the optimal cost is a quadratic function—see (2.1.17). Since this is the case for *any finite horizon* $T = 1, 2, ...,$ we immediately guess that the optimal cost in the infinite-horizon problem (2.4.13)–(2.4.15) is also of the form

$$v(x) := Bx^2 \quad \forall x \in X$$

for some constant B. To verify that this is correct, we consider the DPE (2.3.7) with $V^* = v$. Therefore, (2.3.7) becomes

$$v(x) = \min_{a \in A}[qx^2 + ra^2 + \alpha v(\delta x + \eta a)]$$

or

$$Bx^2 = \min_{a}[qx^2 + ra^2 + \alpha B(\delta x + \eta a)^2]. \tag{2.4.27}$$

The minimum in the right-hand side is attained at

$$\begin{aligned} f(x) &= -\frac{\alpha\delta\eta B}{r + \alpha\eta^2 B} \cdot x \\ &= -\frac{\alpha\delta\eta B}{r + SB} \cdot x, \end{aligned} \tag{2.4.28}$$

with $S{=}\alpha\eta^2$ as in (2.4.19)–(2.4.20). Replacing $a{=}f(x)$ in (2.4.27) and comparing both sides of the resulting equation, we conclude that $v(x) = Bx^2$ indeed satisfies the DPE if B is the unique positive solution of the quadratic equation

$$SB^2 + [r - \alpha(r\delta^2 + q\eta^2)]B - qr = 0.$$

Since this equation is the same as (2.4.21), we obtain that $v(\cdot) = V^*(\cdot)$, the α-optimal discounted cost, and that $f(\cdot)$ in (2.4.28) is the optimal control.

◇

2.5 Long–Run Average Cost Problems

In this section, we study undiscounted infinite horizon optimal control problems with the long–run *average cost* (AC). This optimality criterion was originally introduced by Bellman (1957b) for a class of Markov decision processes (as in Chap. 3, below). In this section we study some aspects of the AC criterion for discrete–time deterministic systems. In the following chapters we study AC optimality for other discrete– and continuous–time, deterministic and stochastic control systems.

Given an initial condition $x_0 = x \in X$ and a policy $\pi = \{a_t\}$, let

$$J_T(\pi, x) := \sum_{t=0}^{T-1} c(x_t, a_t).$$

We wish to minimize the long-run average cost (AC) $J(\pi, x)$ defined as

$$J(\pi, x) := \limsup_{T\to\infty} \frac{1}{T} J_T(\pi, x), \tag{2.5.1}$$

subject to

$$x_{t+1} = F(x_t, a_t), \quad t = 0, 1, \ldots. \tag{2.5.2}$$

The *AC value function* is

$$J^*(x) := \inf\{J(\pi, x) : \pi \in \Pi\} \tag{2.5.3}$$

and a control policy π^* is said to be *average–cost optimal* (AC–optimal) if $J(\pi^*, x) = J^*(x)$ for all $x \in X$.

To avoid trivial situations, we will assume that there is a policy $\pi \in \Pi$ such that the mapping $x \mapsto J(\pi, x)$ is finite-valued.

There are several approaches to analyze the AC optimal control problem (2.5.1)–(2.5.3). The most common are (i) the *AC optimality equation* (ACOE), (ii) the *steady state* (or stationary state) approach, and (iii) the *vanishing discount* approach. We will briefly discuss each of them. (There is also an *infinite-dimensional linear programming* approach to study deterministic AC problems,

but it is too technical to include it here. The interested reader may consult, for instance, Borkar et al. (2019.) or Chap. 11 in Hernández-Lerma and Lasserre (1999).)

Remark 2.47. We will use the following notation:

(a) Given a control policy $\pi \in \Pi$, we denote by $\{x_t^\pi, t = 0, 1, ...\}$ the sequence defined by (2.5.2) when a_t is given by the policy π, that is, $x_{t+1}^\pi = F(x_t^\pi, a_t)$ for all $t = 0, 1, ...$ with some initial condition $x_0^\pi = x_0$. In particular if $f \in \mathbb{F}$ is an stationary policy with $a_t = f(x_t)$, then $x_{t+1}^f = F(x_t^f, f)$ for all $t = 0, 1, \ldots$.

(b) For a given function $\xi : X \to \mathbb{R}$, let Π_ξ be the family of control policies $\pi \in \Pi$ such that, for every initial state x_0,

$$\frac{1}{t}\xi(x_t^\pi) \to 0 \quad \text{as} \quad t \to \infty. \tag{2.5.4}$$

Similarly, we denote by $\mathbb{F}_\xi$ the family of stationary policies $f \in \mathbb{F}$ that satisfies (2.5.4) for every initial state x_0.

Note that if ξ is *bounded*, then (2.5.4) holds for all $\pi \in \Pi$ and all $f \in \mathbb{F}$; hence $\Pi_\xi = \Pi$, and $\mathbb{F}_\xi = \mathbb{F}$. For special functions ξ, the relation (2.5.4) is a transversality-like condition, similar to (2.3.17) in Theorem 2.21(c). ◇

2.5.1 The AC Optimality Equation

A pair (j^*, l) consisting of a real number $j^* \in \mathbb{R}$ and a function $l : X \to \mathbb{R}$ is called a solution to the *average cost optimality equation* (ACOE) if, for every $x \in X$,

$$j^* + l(x) = \inf_{a \in A(x)} [c(x, a) + l(F(x, a))]. \tag{2.5.5}$$

It can be shown that if (j^*, l) is a solution to the ACOE, then j^* is *unique*. Moreover, it is obvious that if $l(\cdot)$ satisfies (2.5.5), then so does $l(\cdot) + k$ for any constant k.

A solution (j^*, l) to the ACOE is also known as a *canonical pair* . If, in addition, f^* is a stationary policy that satisfies (2.5.6) below, then (j^*, l, f^*) is called a *canonical triplet*.

Theorem 2.48. *Suppose that (j^*, l) is a solution to the ACOE (2.5.5), and let Π_l be as in Remark 2.47(b) with $\xi = l$ in (2.5.4). Then, for every initial state $x_0 = x$,*

(a) $j^* \le J(\pi, x)$ *for all* $\pi \in \Pi_l$*; hence*
(b) $j^* \le J^*(x)$ *if* $\Pi = \Pi_l$.

Moreover, suppose that there exists a policy $f^* \in \mathbb{F}_l$ such that $f^*(x) \in A(x)$ attains the minimum in the right–hand side of (2.5.5), i.e.,

$$j^* + l(x) = c(x, f^*) + l(F(x, f^*)) \quad \forall\, x \in X. \tag{2.5.6}$$

Then, for all $x \in X$,

(c) $j^* = J(f^*, x) \le J(\pi, x)$ *for all* $\pi \in \Pi_l$*; hence*
(d) f^* *is AC–optimal and* $J(f^*, \cdot) \equiv J^*(\cdot) \equiv j^*$ *if* $\Pi_l = \Pi$.

Proof. (a) By (2.5.5), for every $(x, a) \in \mathbb{K}$ we have

$$j^* + l(x) \le c(x, a) + l(F(x, a)). \tag{2.5.7}$$

Now consider an arbitrary policy $\pi = \{a_t\} \in \Pi_l$, and let $x_t \equiv x_t^\pi$, $t = 0, 1, \ldots$, be the corresponding state trajectory for any given initial state $x_0^\pi = x_0$. Hence, by (2.5.7),

$$j^* \le c(x_t, a_t) + l(x_{t+1}) - l(x_t) \quad \forall\, t = 0, 1, \ldots.$$

Thus summation over $t = 0, 1, ..., T-1$ gives

$$Tj^* \le J_T(\pi, x) + l(x_T) - l(x_0). \tag{2.5.8}$$

Finally, multiplying by $1/T$ both sides of this inequality and then letting $T \to \infty$, (2.5.1) and (2.5.4) yield part (a).

Part (b) follows from (a) if $\Pi_l = \Pi$.

(c) If f^* satisfies (2.5.6), then we have equality throughout (2.5.7)–(2.5.8), which yields the equality in (c). The inequality follows from (a). Finally, (d) is a consequence of (b) and (c). □

As in Remark 2.47, if l is *bounded*, then $\Pi_l = \Pi$, as required in parts (b) and (d) of Theorem 2.48.

Arguments similar to those in the proof of Theorem 2.48 give other useful results, such as the following.

Proposition 2.49. (a) Suppose that instead of (2.5.7), for some $f \in \mathbb{F}$, we have

$$j^* + l(x) \geq c(x, f) + l(F(x, f)) \quad \forall\, x \in X. \tag{2.5.9}$$

If $\lim_{t\to\infty} l(x_t^f)/t \geq 0$, then $j^* \geq J(f, x)$ for all $x \in X$.
(b) If the inequality in (2.5.9) is reversed, i.e.,

$$j^* + l(x) \leq c(x, f) + l(F(x, f)) \quad \forall\, x \in X, \tag{2.5.10}$$

and $\lim_{t\to\infty} l(x_t^f)/t \leq 0$, then $j^* \leq J(f, x)$.

The proof of Proposition 2.49 is left to the reader (Exercise 2.11).

Corollary 2.50. (a) If (2.5.9) holds for some $f \in \mathbb{F}$, then

$$j^* + l(x) \geq \inf_{a\in A(x)} [c(x, a) + l(F(x, a))] \quad \forall\, x \in X.$$

(b) If (2.5.10) holds for all $f \in \mathbb{F}$, then

$$j^* + l(x) \leq \inf_{a\in A(x)} [c(x, a) + l(F(x, a))] \quad \forall\, x \in X.$$

Example 2.51 (The Brock–Mirman model). In the infinite–horizon Brock and Mirman economic growth model studied in Example 2.31, the system evolves according to

$$x_{t+1} = cx_t^\theta - a_t, \qquad t = 0, 1, 2, \ldots,$$

with a given initial state $x_0 \in X$ and $\theta \in (0, 1)$. As in Examples 2.10 and 2.31, we assume that $X = A = [0, \infty)$, and $A(x) := (0, cx^\theta]$. The system function is $F(x, a) = cx^\theta - a$, and the stage reward (or utility) function is $r(x, a) = \log(a)$. Thus the performance index to be optimized is the long–run *average reward* (AR)

$$J(\pi, x_0) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log(a_t). \tag{2.5.11}$$

Concerning the "lim inf" in (2.5.11), instead of the "lim sup" in (2.5.1), see the paragraph and the Remark after expression (5.2.15) in Sect. 5.2.

To find the canonical triplet (j^*, l, f^*) that satisfies the ACOE

$$j^* + l(x) = \max_{a \in A(x)} [r(x,a) + l(F(x,a))] \quad \forall x \in X, \tag{2.5.12}$$

we consider a function $l(x)$ of the form $l(x) := b \log(x)$, where b is an unknown parameter. Under this assumption, the right side of (2.5.12) reaches the maximum when $a = \frac{cx^\theta}{1+b}$, so we can rewrite (2.5.12) as

$$\begin{aligned} j^* + b\log(x) &= \log\left(\frac{cx^\theta}{1+b}\right) + b\log\left(x^\theta\left(\frac{cb}{1+b}\right)\right) \\ &= (1+b)\theta\log(x) + \log\left(\frac{c}{1+b}\right) + b\log\left(\frac{cb}{1+b}\right). \end{aligned}$$

This last equation is satisfied if $b = (1+b)\theta$, which implies that $b = \dfrac{\theta}{1-\theta}$. Therefore the canonical triplet (j^*, l, f^*) is given by

$$j^* = \log\left(c(1-\theta)\right) + \frac{\theta}{1-\theta}\log\left(c\theta\right), \tag{2.5.13}$$

$$l(x) = \frac{\theta}{1-\theta}\log(x), \tag{2.5.14}$$

$$f^*(x) = c(1-\theta)x^\theta. \tag{2.5.15}$$

It can be shown that, for every initial state x_0,

$$x_t^{f^*} = (c\theta)^{\frac{1-\theta^t}{1-\theta}} x_0^{\theta^t} \quad \text{for} \quad t = 1, 2, \ldots \tag{2.5.16}$$

and $x_t^{f^*} \to (c\theta)^{\frac{1}{1-\theta}}$, which implies that f^* satisfies (2.5.4), that is $f^* \in \mathbb{F}_l$. Thus by Theorem 2.48(d), j^* is the optimal average reward and $f^*$is AR-optimal. ◇

The ACOE (2.5.5) provides a "complete solution" to the AC control problem in the sense that, in addition to providing a canonical triplet (j^*, l, f^*), it also allows us to identify *refinements* of this triplet, such as "overtaking optimal" or "bias optimal" controls. However, if we are only interested in obtaining AC-optimal controls, it suffices to obtain an *optimality inequality*. This is explained in Sect. 2.5.3 below.

2.5.2 The Steady–State Approach

Let $\mathbb{K}$ be as in (2.3.12). A pair $(x^*, a^*) \in \mathbb{K}$ is said to be a *steady* (or *stationary*) state-action pair for the system (2.5.2) if $F(x^*, a^*) = x^*$. If, in addition, (x^*, a^*) solves the steady state problem

$$\text{minimize } c(x, a) \text{ subject to } F(x, a) = x, \tag{2.5.17}$$

then (x^*, a^*) is said to be a *minimum steady state-action pair* for the AC control problem (2.5.1)–(2.5.2).

Assumption 2.52. The OCP (2.5.1)–(2.5.2) satisfies:

(a) There exists a minimum steady state-action pair $(x^*, a^*) \in \mathbb{K}$.

(b) *Dissipativity.* The OCP (2.5.1)–(2.5.2) is *dissipative*, which means that there is a so–called *storage function* $\lambda : X \to \mathbb{R}$ such that, for every $(x, a) \in \mathbb{K}$,

$$\lambda(x) - \lambda(F(x, a)) \leq c(x, a) - c(x^*, a^*). \tag{2.5.18}$$

(c) *Stabilizability.* Let Π_λ be as in Remark 2.47, with λ as in (2.5.18). For each initial state $x_0 \in X$, there exists a control policy $\bar{\pi} \in \Pi_\lambda$ (which may depend on x_0) such that the corresponding state–control path $(\bar{x}_t, \bar{a}_t)$ converges to the minimum steady state pair (x^*, a^*) in (a).

Observe that, introducing the constant $w^* := c(x^*, a^*)$, the *dissipativity inequality* (2.5.18) can be expressed as

$$w^* + \lambda(x) \leq c(x, a) + \lambda(F(x, a)) \quad \forall\, (x, a) \in \mathbb{K}, \tag{2.5.19}$$

which in turn gives

$$w^* + \lambda(x) \leq \inf_{a \in A(x)} [c(x, a) + \lambda(F(x, a))] \quad \forall\, x \in X. \tag{2.5.20}$$

Consequently, in view of the similarity between (2.5.20) and the ACOE (2.5.5), one would expect some connection between the constants w^* and v^*. In fact, the following theorem shows that w^* satisfies conditions similar to (a)–(d) in Theorem 2.48. (See also Proposition 2.49(b).)

Theorem 2.53. *Suppose that Assumption 2.52 holds and the stage cost $c : \mathbb{K} \to \mathbb{R}$ is continuous. For the storage function λ in (2.5.18) let Π_λ be as in Remark 2.47. Then, for all $x \in X$,*

(a) $w^* = J(\bar{\pi}, x) \le J(\pi, x)$ *for all* $\pi \in \Pi_\lambda$, *where* $\bar{\pi}$ *satisfies Assumption 2.52(c); hence*

(b) *the AC value function satisfies that* $J^*(x) = c(x^*, a^*)$, *with* (x^*, a^*) *as in Assumption 2.52(a), if* $\Pi_\lambda = \Pi$.

Proof. (a) Let $w^* := c(x^*, a^*)$ be as in (2.5.19). Let $\pi = \{a_t\}$ be an arbitrary control policy in Π_λ with corresponding state–action sequence (x_t, a_t). Then, from (2.5.19),

$$w^* \le c(x_t, a_t) + \lambda(x_{t+1}) - \lambda(x_t)$$

for all $t = 0, 1, \ldots$. This yields, as in (2.5.7)–(2.5.8),

$$Tw^* \le J_T(\pi, x) + \lambda(x_T) - \lambda(x_0) \quad \forall\, T = 1, 2, \ldots,$$

so, by (2.5.4),

$$w^* \le J(\pi, x) \quad \text{for all} \quad x \in X. \tag{2.5.21}$$

On the other hand, by the stabilizability in Assumption 2.52(c), there is a policy $\bar{\pi} \in \Pi_\lambda$ for which the state–action sequence $(\bar{x}_t, \bar{a}_t)$ converges to (x^*, a^*). Therefore, since the stage cost c is continuous, $c(\bar{x}_t, \bar{a}_t)$ converges to w^*, which implies that $J(\bar{\pi}, x) = w^*$ for all x. This prove (a).

(b) If $\Pi_\lambda = \Pi$, then part (a) yields that $\bar{\pi}$ is AC-optimal and also that the AC value function is $J^*(\cdot) \equiv w^*$. □

Example 2.54 (The Brock–Mirman model, cont'd.). In Example 2.51, for the system function $F(x, a) = cx^\theta - a$ and the stage reward (or utility) function $r(x, a) = \log(a)$, it can be verified that the unique solution to the corresponding steady–state problem (2.5.17), namely,

$$\text{maximize:} \quad \log(a) \quad \text{subject to} \quad cx^\theta - a = x,$$

is given by

$$(x^*, a^*) = \left((c\theta)^{\frac{1}{1-\theta}}, c(1-\theta)(c\theta)^{\frac{\theta}{1-\theta}} \right). \tag{2.5.22}$$

Then using Theorem 2.53(a)

$$J(\bar{\pi}, x) = r(x^*, a^*) = \log(a^*) = \log\left(c(1-\theta)\right) + \frac{\theta}{1-\theta}\log\left(c\theta\right); \tag{2.5.23}$$

compare this with (2.5.13). To get Assumption 2.52(b), note that $F(x, a^*) - F(x, a) = a - a^*$. Thus, the strict concavity of the stage reward $r(x, a) := \log(a)$ gives

$$\begin{aligned} r(x,a) - r(x^*,a^*) &\le \frac{\partial r}{\partial x}(x^*,a^*)(x-x^*) + \frac{\partial r}{\partial a}(x^*,a^*)(a-a^*) \\ &= \frac{1}{a^*}(F(x,a^*) - F(x,a)). \end{aligned}$$

Moreover, the system function $F(x, a^*)$ is also concave in x and $\frac{\partial F}{\partial x}(x^*, a^*) = 1$, so

$$F(x,a^*) - F(x^*,a^*) \le \frac{\partial F}{\partial x}(x^*,a^*)(x-x^*) = x - x^*,$$

that is, $F(x, a^*) \le x$ for all $x \in X$. Therefore, from the last two inequalities we get the corresponding dissipativity condition (2.5.18)

$$r(x,a) - r(x^*,a^*) \le \lambda(x) - \lambda(F(x,a))$$

with the storage function $\lambda(x) := \frac{1}{a^*}x$. Notice that λ is different from l given in (2.5.14), Example 2.51.

Observe that for any initial state $x_0 \in X$, the policy f^* in (2.5.15) with corresponding state-control path (x_t, a_t), where

$$x_t = (c\theta)^{\frac{1-\theta^t}{1-\theta}} x_0^{\theta^t}, \qquad a_t = c(1-\theta)(c\theta)^{\frac{\theta-\theta^{t+1}}{1-\theta}} x_0^{\theta^{t+1}},$$

satisfies the stabilizability Assumption 2.52(c), i.e., (x_t, a_t) converges to the optimal stationary pair (x^*, a^*).

Hence by Theorem 2.53, $c(x^*, a^*) = J(\bar{\pi}, x) \ge J(\pi, x)$ for all $\pi \in \Pi_\lambda$ and all $x \in X$. ◇

Example 2.55 (The Mitra-Wan forestry model). Consider a forestland covered by trees of the same species classified by age-classes from 1 to n. After age n, trees have no economic value. The state space in this example (Mitra and Wan Jr 1985) can be

identified with the n–simplex

$$\Delta := \{x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1, x_i \geq 1, i = 1, ..., n\}$$

where each coordinate x_i denotes the proportion of land occupied by i-aged trees.

Let $x_t = (x_{1,t}, ..., x_{n,t}) \in \Delta$ be the forest state at period t. By the end of the period the forester must decide to harvest a proportion of land in any age class, say $a_t = (a_{1,t}, ..., a_{n,t})$ with $0 \leq a_{i,t} \leq x_{i,t}$, $i = 1, ..., n$. Because a tree has no economic value after age n, $a_{n,t} = x_{n,t}$. Thus, for each $x \in \Delta$, the admissible control set is $A(x) = [0, x_1] \times \cdots \times [0, x_{n-1}] \times \{x_n\}$. Suppose that the forest evolves according to the dynamic model

$$x_{1,t+1} = a_{1,t} + \cdots + a_{n,t}, \tag{2.5.24}$$

$$x_{i+1,t+1} = x_{i,t} - a_{i,t}, \ i = 1, ..., n-1, \tag{2.5.25}$$

where (2.5.24) means that all harvested area at the end of period t must be sown by trees of age 1 at the beginning of period $t+1$. On the other hand, (2.5.25) states that trees of age i that have not been harvested until the end of period t become trees of age $i+1$ in period $t+1$.

For a planning horizon T, (2.5.24)–(2.5.25) can be written as a discrete-time linear control system

$$x_{t+1} = f(x_t, a_t) := Ax_t + Ba_t \quad \text{for} \quad t = 0, 1, ..., T-1, \tag{2.5.26}$$

where

$$A := \begin{pmatrix} 0 & 0 & . & 0 & 0 \\ 1 & 0 & . & 0 & 0 \\ 0 & 1 & . & 0 & 0 \\ . & . & . & . & . \\ 0 & 0 & . & 1 & 0 \end{pmatrix}, \quad \text{and} \quad B := \begin{pmatrix} 1 & 1 & . & 1 & 1 \\ -1 & 0 & . & 0 & 0 \\ 0 & -1 & . & 0 & 0 \\ . & . & . & . & . \\ 0 & 0 & . & -1 & 0 \end{pmatrix}. \tag{2.5.27}$$

Now, assume the timber production per unit area is related to the tree age-classes by the *biomass vector*

$$\xi = (\xi_1, \xi_2, ..., \xi_n) \in \mathbb{R}^n, \quad \xi_i \geq 0, i = 1, 2, \ldots, n,$$

where ξ_i represents the amount of timber produced by i-aged trees occupying a unit of land. Hence, the total amount of timber collected at the end of period t is given by

$$\xi \cdot a_t = \xi_1 a_{1,t} + \cdots + \xi_n a_{n,t}.$$

Consider a timber price function $p : [0, \infty) \to [0, \infty)$, assumed to be increasing and concave. Given a forest state x and an admissible harvest control a, the stage income is $r(x, a) := p(\xi \cdot a)$. Therefore, the performance index to maximize is

$$J_T(\pi, x) := \sum_{t=0}^{T-1} r(x_t, a_t). \tag{2.5.28}$$

It can be shown that the control system (2.5.26) has a set of stationary states given by the pairs (x, a) satisfying $x_1 \geq x_2 \geq \cdots \geq x_n$ and

$$a_1 = x_1 - x_2, a_2 = x_2 - x_3, ..., a_n = x_n.$$

Moreover, for each age class i there is a pair of stationary state and control (x^i, a^i), known as *normal forest*, defined as follows: the state is $x^i := (1/i, ..., 1/i, 0, ..., 0)$, where each of the first i coordinates are $1/i$, and the remaining are 0; and the control is $a^i := (0, ..., 0, 1/i, 0, ..., 0)$, where $1/i$ is in the i–coordinate.

We choose a normal forest (x^*, a^*) such that

$$r(x^*, a^*) = \max \left\{ p\left(\xi \cdot a^i\right) : i = 1, 2, ..., n \right\}.$$

So, given the concavity of p, there is $k \geq 0$ such that

$$r(x, a) - r(x^*, a^*) \leq k\xi \cdot (a - a^*) \quad \text{for all } a \in A(x).$$

Letting $N := (1, 2, ..., n)$ and $\gamma := \max\{\xi \cdot a^i : i = 1, 2, ..., n\}$, we get the vector componentwise inequality $\xi \leq \gamma N$. Moreover, from a straightforward calculation we have $N \cdot [x - F(x, a)] = N \cdot (a - a^*)$ for any $x \in \Delta$ and all $a \in A(x)$. Therefore,

$$r(x, a) - r(x^*, a^*) \leq k\gamma N \cdot [x - F(x, a)] \quad \text{for all } x \in \Delta,\ a \in A(x).$$

Introducing the function $\lambda : \Delta \to \mathbb{R}$, defined by $\lambda(x) = k\gamma N \cdot x$, and the value $j^* := r(x^*, a^*)$, we have the corresponding dissipative inequality (2.5.19),

$$j^* + \lambda(x) \geq r(x, a) + \lambda(F(x, a)) \quad \text{for all } x \in \Delta,\ a \in A(x).$$

Thus in particular we conclude that (x^*, a^*) is an optimal stationary state. Moreover, that optimal stationary state can be reached by a finite sequence of harvest plans from any initial stationary state. Hence, this example satisfies the Assumption 2.52, and the conditions of Theorem 2.53, so the optimal AC value function is

$$J^*(x) \equiv r(x^*, a^*) = \max\left\{ p\left(\frac{\xi_i}{i}\right) : i = 1, 2, ..., n \right\}.$$

◇

2.5.3 The Vanishing Discount Approach

The so-called vanishing discount approach to AC-control problems is based on several connections between discounted cost problems and the average cost. The most straightforward is the following. Given an arbitrary control policy $\pi = \{a_t\}$ and initial state $x_0 = x$, consider the discounted cost $V(\pi, x)$ in (2.3.3), which we now write as $V_\alpha(\pi, x)$ to make explicit the dependence on the discount factor $\alpha \in (0, 1)$, i.e.,

$$V_\alpha(\pi, x) = \sum_{t=0}^{\infty} \alpha^t c(x_t, a_t). \tag{2.5.29}$$

Now, let M be an arbitrary constant and inside the summation replace $c(\cdot, \cdot)$ with $c(\cdot, \cdot) \pm M$. Hence (2.5.29) becomes

$$V_\alpha(\pi, x) = \sum_{t=0}^{\infty} \alpha^t [c(x_t, a_t) - M] + \frac{M}{1 - \alpha},$$

which, multiplying both sides by $1 - \alpha$, we may express as

$$(1-\alpha)V_\alpha(\pi,x) = M + (1-\alpha)\sum_{t=0}^{\infty}\alpha^t[c(x_t,a_t) - M].$$

In particular, taking M as the average cost $J(\pi,x)$ it follows that

$$(1-\alpha)V_\alpha(\pi,x) = J(\pi,x) + (1-\alpha)\sum_{t=0}^{\infty}\alpha^t[c(x_t,a_t) - J(\pi,x)]. \tag{2.5.30}$$

This equation obviously suggests that we can approximate $J(\pi,x)$ by $(1-\alpha)V_\alpha(\pi,x)$ as $\alpha \uparrow 1$.

A second connection between discounted cost problems and the average cost is provided by the Abelian theorem in Part (a) of the following lemma.

Lemma 2.56. Let $\{c_t\}$ be a sequence bounded below, and consider the lower and upper limit averages (also known as Cesàro limits)

$$C^L := \liminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}c_t, \quad C^U := \limsup_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}c_t,$$

and the lower and upper Abelian limits

$$A^L := \liminf_{\alpha\uparrow 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^t c_t, \quad A^U := \limsup_{\alpha\uparrow 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^t c_t.$$

Then

(a) $C^L \le A^L \le A^U \le C^U$.

(b) If $A^L = A^U$, the equality holds in (a), i.e., $C^L = A^L = A^U = C^U$.

For a proof of Lemma 2.56 see the references in Bishop et al. (2014) or Sznajder and Filar (1992). Part (b) in Lemma 2.56 is known as the Hardy-Littlewood Theorem.

Consider now a control policy $\pi = \{a_t\}$, the corresponding state trajectory $\{x_t\}$, and in Lemma 2.56 take $c_t := c(x_t,a_t)$. Then the third inequality in Lemma 2.56(a) gives

$$\limsup_{\alpha\uparrow 1}(1-\alpha)V_\alpha(\pi,x) \le J(\pi,x) \tag{2.5.31}$$

for every initial state $x_0 = x$, with $V_\alpha(\pi, x)$ and $J(\pi, \cdot)$ as in (2.5.30). Moreover, if $V_\alpha^*(\cdot) \equiv V^*(\cdot)$ denotes the α-discount value function in (2.3.4), then (2.5.31) yields

$$\limsup_{\alpha \uparrow 1} (1-\alpha) V_\alpha^*(x) \leq J(\pi, x).$$

In fact, since π in the latter inequality is arbitrary, we obtain from (2.5.3) that, for every $x \in X$,

$$\limsup_{\alpha \uparrow 1} (1-\alpha) V_\alpha^*(x) \leq J^*(x). \tag{2.5.32}$$

In words, (2.5.32) states that, for values of α close to 1, $(1-\alpha)V_\alpha^*(\cdot)$ *is a lower bound for the average cost* $J^*(\cdot)$.

One more connection between discounted cost problems and the AC criterion is provided by the α-discount dynamic programming equation (2.3.7) that we will rewrite as

$$V_\alpha^*(x) = \inf_{a \in A(x)} [c(x,a) + \alpha V_\alpha^*(F(x,a))] \tag{2.5.33}$$

Consider an arbitrary constant m_α, which may depend on $\alpha \in (0,1)$, and define

$$h_\alpha(x) := V_\alpha^*(x) - m_\alpha, \quad \text{and} \quad \rho(\alpha) := (1-\alpha) m_\alpha. \tag{2.5.34}$$

Some typical choices of the constant m_α are $m_\alpha := V_\alpha^*(\bar{x})$, where $\bar{x} \in X$ is an arbitrary (but fixed) state, and $m_\alpha := \inf_{x \in X} V_\alpha^*(x)$, assuming of course that V_α^* is bounded below. The first choice of m_α is useful because $h_\alpha(\bar{x}) = 0$, so that we "fix" h_α at $\bar{x}$. The second choice is also useful because then h_α is a *nonnegative* function. Either way, using (2.5.34) the DPE (2.5.33) becomes

$$\rho(\alpha) + h_\alpha(x) = \inf_{a \in A(x)} [c(x,a) + \alpha h_\alpha(F(x,a))]. \tag{2.5.35}$$

Comparing this equation with (2.5.5), we might try to get conditions for the pair $(\rho(\alpha), h_\alpha)$ in (2.5.35) to converge, as $\alpha \uparrow 1$, to a solution (j^*, l) of (2.5.5).

We will next show by means of examples the *feasibility* of this approach. It should be noted however that, to the best of our knowledge, *there are no general results on the "vanishing discount approach" for deterministic discrete-time systems such as (2.5.1)–(2.5.2)*. All the known results on discrete-time AC control problems refer to *stochastic systems*. See Remark 2.60.

Example 2.57 (An LQ system, cont'd.). We consider again the α-discounted LQ control problem (2.4.13)–(2.4.15), with the α-optimal discounted cost $V_\alpha^*(x) = C(\alpha)x^2$ in (2.4.22). Here $C(\alpha) \equiv C$ is the unique positive solution of (2.4.21) with coefficients $P, Q = Q(\alpha), S = S(\alpha)$ as in (2.4.20), i.e.,

$$P = qr, \quad Q(\alpha) = (r\delta^2 + q\eta^2)\alpha, \quad S(\alpha) = \alpha\eta^2. \tag{2.5.36}$$

We suppose again the conditions in Example 2.46, but now we assume in addition that the coefficients in (2.4.13) are such that $|\delta| < 1$ and $\delta\eta > 0$.

Now, in (2.5.34) take

$$m_\alpha := \inf_{x\in X} V_\alpha^*(x) = V_\alpha^*(\bar{x}) = 0,$$

with $\bar{x} = 0$. Then $\rho(\alpha) = 0$ for every $\alpha \in (0, 1)$, and, as $\alpha \uparrow 1$, we obviously have that $\rho(\alpha) \to j^* = 0$ and

$$h_\alpha(x) = V_\alpha^*(x) \to l(x) := C(1)x^2 \quad \forall x$$

where $C(1)$ is the unique positive solution of the quadratic equation (2.4.21) when $\alpha = 1$. We can also see that, as $\alpha \uparrow 1$, the α-optimal control $f_\alpha(\cdot) \equiv f^*(\cdot)$ in (2.4.23) converges to $g^*(x) := -\theta x$ for all x, where $\theta := (r + S(1)C(1))^{-1}\delta\eta C(1)$ and $S(1)$ is given in (2.5.36) when $\alpha = 1$.

Summarizing, (j^*, l, g^*) is a canonical triplet, as in (2.5.6), with minimum average cost $j^* = 0$. (Here, we are tacitly using the following fact: If we write $a_t = -\theta x_t$ in (2.4.13), then $x_{t+1} = (\delta - \eta\theta)x_t$ is a *stable* system since the coefficient $|\delta - \eta\theta| < 1$.) ◇

Example 2.58 (The Brock-Mirman model, cont'd.). Consider the Brock-Mirman model in Example 2.31. In this α-discount

problem, the performance index to be maximized for a given initial state $x_0 = x$ is

$$V_\alpha(\pi, x) = \sum_{t=0}^{\infty} \alpha^t \log(a_t).$$

The optimal control policy of this problem is $f_\alpha^*(x) := c(1 - \theta\alpha)x^\theta$ (see (2.3.33)) and the optimal state-control pair path (x_t^*, a_t^*) for $t = 1, 2, ...$ is given by

$$x_t^* = (c\theta\alpha)^{\frac{1-\theta^t}{1-\theta}} x^{\theta^t} \quad \text{and} \quad a_t^* = c(1 - \theta\alpha)(c\theta\alpha)^{\frac{\theta-\theta^{t+1}}{1-\theta}} x^{\theta^{t+1}}.$$

Moreover, from Remark 2.32 the corresponding α-discount value function is

$$V_\alpha^*(x) = \frac{1}{1-\alpha} \log[c(1 - \theta\alpha)] + \frac{\theta\alpha}{(1-\alpha)(1-\theta\alpha)} \log(c\theta\alpha) + \frac{\theta}{1-\theta\alpha} \log(x).$$

Rearranging terms, we have

$$(1-\alpha)V_\alpha^*(x) = \log[c(1 - \theta\alpha)] + \frac{\theta\alpha}{1-\theta\alpha} \log(c\theta\alpha) + \frac{\theta(1-\alpha)}{1-\theta\alpha} \log(x),$$

and therefore

$$\lim_{\alpha\uparrow 1}(1-\alpha)V_\alpha^*(x) = \log(c(1-\theta)) + \frac{\theta}{1-\theta} \log(c\theta) = j^* \quad \forall x.$$

In other words, the constant function $J^*(\cdot) \equiv j^*$ is the average optimal value function which, of course, coincides with (2.5.13) and (2.5.23).

Finally, notice that, as $\alpha \uparrow 1$, the optimal path (x_t^*, a_t^*) converges to the steady-state pair (x^*, a^*) in (2.5.22). ◇

Remark 2.59 (The Arzela-Ascoli Theorem). Let $C_b(X)$ be the space of real-valued continuous bounded functions on a metric space X, with the supremum norm $\|f\| := \sup_{x\in X} |f(x)|$. The Arzela-Ascoli theorem characterizes compact subspaces of $C_b(X)$. This result is used, for instance, in the vanishing discount approach, to decide whether a sequence (say, $h_{\alpha_n}(\cdot)$ in (2.5.34)) in $C_b(X)$ has a *convergent subsequence*. A precise statement is as follows.

The Arzela-Ascoli Theorem. Suppose that X is a compact metric space and let f_n be a sequence in $C_b(X)$ such that

(a) f_n is bounded, that is, $\sup_n \|f_n\| < \infty$, and
(b) f_n is equicontinuous, that is, for each $\epsilon > 0$ there exists $\delta > 0$ such that

$$\sup_n |f_n(x) - f_n(y)| < \epsilon \qquad \text{if} \quad |x - y| < \delta.$$

Then f_n has a subsequence converging to some function in $C_b(X)$.

For a proof of this theorem see, for example, Appendix H in Morimoto (2010). ◇

Remark 2.60. As already noted in the paragraph before Example 2.57, as far as we can tell *there are no general results on the "vanishing discount approach" to discrete-time deterministic AC problems.* (For *differential systems*, see part (b) below.) In fact, the closest result we are aware of is the following theorem by Feinberg et al. (2012) for Markov decision processes (MDPs), which we introduce in Chap. 3. Note, however, that this theorem does not give the ACOE (2.5.5); it gives the *optimality inequality* (2.5.37) below, which is the same as the inequality in Corollary 2.50(a). (Vega-Amaya (2015) presents another proof of this theorem.)

(a) For our *deterministic* AC control problem (2.5.1)–(2.5.3) the Feinberg et al. (2012) theorem can be stated as follows.

Theorem (Feinberg et al. (2012)). Suppose that:

(a) Assumption 2.17 holds, that is, the system function $F(\cdot,\cdot)$ is continuous, and the stage cost $c(\cdot,\cdot)$ is nonnegative and $\mathbb{K}$-inf-compact;
(b) there exists a policy π and an initial state x such that $J(\pi, x)$ is finite, and, moreover, the function

$$h(\cdot) := \liminf_{\alpha \uparrow 1} h_\alpha(\cdot)$$

is finite-valued, where $h_\alpha(\cdot) = V^*_\alpha(\cdot) - m_\alpha$ is as in (2.5.34), with $m_\alpha = \inf_{x \in X} V^*_\alpha(x)$.

Then there exists a l.s.c. function $l(\cdot)$ on X and a stationary policy $f \in \mathbb{F}$ such that

$$j^* + l(x) \geq \inf_{a \in A(x)} [c(x,a) + l(F(x,a))] \\ = c(x,f) + l(F(x,f)) \quad \forall x \in X, \quad (2.5.37)$$

where $j^* := \limsup_{\alpha \uparrow 1} m_\alpha$ is the optimal AC, and f is AC-optimal, that is, $J(f,x) = J^*(x) = j^*$ for all $x \in X$.

(b) AC control problems are mainly studied for stochastic systems, as in Chaps. 3, 5, and 6 below. For *discrete-time* deterministic systems the AC problems are practically unexplored, except perhaps for implicit results such as the theorem in part (a). Similarly, for differential systems (as in Chap. 4, below) AC problems have been studied in just a handful of papers such as Arisawa (1997) and Kawaguchi (2003).

◇

For additional comments on deterministic AC control problems see Hernández-Lerma et al. (2023).

Exercises

2.1. Let X and Y be (nonempty) sets, and $D \subset X \times Y$. Assume that the x–section $D(x) := \{y \in Y : (x,y) \in D\}$ is nonempty for all $x \in X$, and similarly for the y–section $D(y) := \{x \in X : (x,y) \in D\}$ for all $y \in Y$. Let v be a real–valued function on D. Prove that:

(a)

$$\sup_{(x,y) \in D} v(x,y) = \sup_{x \in X} \sup_{y \in D(x)} v(x,y) \\ = \sup_{y \in Y} \sup_{x \in D(y)} v(x,y).$$

(b) Show that (a) holds if "sup" is replaced by "inf".

Remark. Parts (a) and (b) in Exercise 2.1 are called "property of the repeated supremum" and "property of the repeated infimum", respectively.

2.2. Let g and h be real–valued functions on a set X.

1. If g and h are bounded from above, then

   (a1) $\sup_x g(x) - \sup_x h(x) \leq \sup_x [g(x) - h(x)]$;
   (a2) $|\sup_x g(x) - \sup_x h(x)| \leq \sup_x |g(x) - h(x)|$.

2. If g and h are bounded from below, then

$$|\inf_x g(x) - \inf_x h(x)| \leq \sup_x |g(x) - h(x)|.$$

2.3. Let X and Y be convex subsets of $\mathbb{R}$, and v a convex function on a convex set $D \subset X \times Y$. Then

$$v^*(x) := \inf_{y \in D(x)} v(x, y)$$

is convex, provided that v^* is finite–valued, where $D(x)$ is the x–section defined in Exercise 2.1 above. If v is strictly convex, then so is v^*.

2.4. Prove Lemma 2.2—Bellman's principle of optimality.

2.5. Prove Lemma 2.15.
Hint. To prove part (a) note that, if $g_k \downarrow g$, then $\lim_k g_k(y) = \inf_k g_k(y) = g(y)$ for all $y \in Y$. In other words, $\lim_k g_k = \inf_k g_k$. Now use Exercise 2.1(b) above. Part (b) in Lemma 2.15 follows from the definition of uniform convergence. Proving part (c) is a little complicated. The reader might wish to see the proof of Lemma 4.2.4 in Hernández-Lerma and Lasserre (1996).

2.6. Give an example of a function on a metric space that is l.s.c. but not inf–compact (as defined in Lemma 2.15(c)).

2.7. Let K be the operator in (2.3.8) defined on the complete metric space $B(X)$ of measurable bounded functions v with the supremum norm $\|v\| := \sup_{x \in X} |v(x)|$. Suppose that the cost function $c(x, a)$ is bounded. Prove that:

(a) K is a contraction on $B(X)$; in fact, $\|Kv - Kv'\| \leq \alpha \|v - v'\|$ for all $v, v' \in B(X)$, and
(b) the α–discount value function V^* is in the space $B(X)$ and it is the unique fixed–point of K (see (2.3.9)) in $B(X)$.

Hint. To prove (a) use Exercise 2.2, part 2. For (b), recall Remark 2.25.

2.8. Let $v : \mathbb{K} \to \mathbb{R}$ be $\mathbb{K}$–inf–compact (see Definition B.4(a2)). Show that v is l.s.c.

2.9. Let $L^+(X)$ be as in Lemma 2.18, that is, the family of nonnegative l.s.c. functions on X. Show that $L^+(X)$ is a *convex cone*, so if u and v are in $L^+(X)$ and $k \geq 0$, then $u + v$ and ku are also in $L(X)$.

2.10. Let $v \in L^+(X)$ and u be as in (2.3.14). Show that, under the Assumption 2.17, the functions $(x, a) \mapsto c(x, a), v(F(x, a))$, $u(x, a)$ are all l.s.c.
Hint. Use Exercise 2.8.

2.11. Prove the Proposition 2.49.

2.12. Prove Lemma 2.34.

2.13. Prove Lemma 2.35(a).

2.14. Prove (2.4.6) for any policy $\pi = \{a_t\}$.
Hint. In (2.4.2), replace $(x, a) \in \mathbb{K}$ by (x_t, a_t) with $t = 0, 1, \ldots$. Next multiply by α^t both sides of (2.4.2), and then sum over all $t = 0, 1, \ldots$. Finally, rearrange terms to obtain (2.4.6).

2.15. Let v_n $(n = 1, 2, ...)$ and v be functions on X such that $v_n \uparrow v$. Show that if the v_n are convex or l.s.c. or monotone, then so is v, respectively.

2.16. Consider the time-varying control system (2.0.1)–(2.0.2) with state and action spaces $X \subset \mathbb{R}^n$ and $A \subset \mathbb{R}^m$, respectively, with A compact. Assume, moreover, that the system function and the stage costs are linear in the state variable, that is,

$$F_t(x, a) = F_1(t)x + F_2(a) \quad \text{and} \quad c_t(x, a) = c_1(t) \cdot x + c_2(a),$$

and terminal cost $C_T(\cdot) \equiv 0$, where $F_2(\cdot)$ and $c_2(\cdot)$ are continuous functions.

(a) Prove that the OCP has an optimal control which is *independent* of the state variable.

Hint. Use the DP algorithm (2.1.8)–(2.1.9). (The result in (a) is due to Midler (1969).)

(b) Show that, under the appropriate conditions (for instance, as in Theorem 2.21), the result in (a) also holds in the infinite-horizon stationary case (2.3.1)–(2.3.2).

2.17. Maximize

$$V(\pi, x_0) = \sum_{t=0}^{T-1} \sqrt{a_t x_t}$$

over all $\pi = \{a_t\}$, with $a_t \in [0, 1]$, subject to $x_{t+1} = \rho(1 - a_t)x_t$, $t = 0, 1, ..., T - 1$, where $x_0 > 0$ and $\rho > 0$.

Answer.

$$a_t^* = \frac{1 - \rho}{1 - \rho^{T-t}}, \quad V_t(x) = \sqrt{x \frac{1 - \rho^{T-t}}{1 - \rho}}, \quad t = 0, 1, ..., T - 1.$$

2.18. (A cake eating or nonrenewable-resource extraction problem.) Maximize

$$V(\pi, x_0) = \sum_{t=0}^{T-1} \beta^t \frac{a_t^{1-\gamma}}{1 - \gamma}$$

over all $\pi = \{a_t\}$, with $a_t \in (0, x_t)$, subject to $x_{t+1} = x_t - a_t, t = 0, 1, ..., T - 1$, where $0 < \gamma < 1$ and $x_0 > 0$.

Answer.

$$a_t^* = x \frac{1 - \beta^{1/\gamma}}{1 - \beta^{(T-t)/\gamma}}, \quad V_t(x) = x^{1-\gamma} \frac{\beta^t \left(1 - \beta^{(T-t)/\gamma}\right)}{(1 - \gamma)\left(1 - \beta^{1/\gamma}\right)}, \quad t = 0, 1, ..., T - 1.$$

Chapter 3

Discrete–Time Stochastic Control Systems

For the discrete–time deterministic systems studied in Chap. 2 there is essentially a unique dynamic model, namely,

$$x_{t+1} = F(x_t, a_t) \quad \forall\, t = 0, 1, \ldots, \tag{3.0.1}$$

with a given initial condition x_0. In contrast, in the *stochastic case* there are two common dynamic models: the so–called **system model**, and the **Markov control model**.

3.1 Stochastic Control Models

In the system model (SM), also known as the *control model*, the controlled system evolves according to a difference equation (similar to (3.0.1)) of the form

$$x_{t+1} = F(x_t, a_t, \xi_t) \quad \forall\, t = 0, 1, \ldots, T-1, \tag{3.1.1}$$

for $T \leq \infty$, with a given—possibly random—initial condition x_0. Here, the state and control variables x_t, a_t have the same meaning as in (3.0.1); in particular, they take values in a *state space* X and an *action set* A, respectively, both assumed to be Borel spaces. Moreover, the ξ_t are independent random variables with values in a Borel space S, and they denote *random perturbations*. These perturbations (as in Remark 1.2(b)) can form a *driving process*

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*, Texts in Applied Mathematics 76,
https://doi.org/10.1007/978-3-031-21139-3_3

or a *random noise*. In the former case, the ξ_t have a physical or economic interpretation, whereas in the latter case they are arbitrary random variables. See Example C.4 and Remark C.5 in Appendix C.

The Markov control model (MCM) approach can be traced back to the paper by Bellman (1957b) who, in addition, coined the term *Markov decision process*, which today is also known as a *Markov control process*. The difference between the MCM and (3.1.1) is that in a MCM the evolution of the system is not specified by a "system function" $F(x, a, s)$ as in (3.1.1); rather it is specified by a *stochastic kernel* or *transition probability*—as in Definition C.2 in Appendix C. (An approach similar to Bellman's was introduced by Shapley (1953) for *stochastic games*.)

More precisely, a MCM is expressed in the form

$$(X, A, \{A(x) : x \in X\}, Q, c), \tag{3.1.2}$$

where, as usual, X and A are Borel spaces denoting the state space and the control or action set, respectively. Moreover, for each $x \in X$, $A(x) \in \mathcal{B}(A)$ represents the set of feasible (or admissible) actions in the state x. Let

$$\mathbb{K} := \{(x, a) \in X \times A : a \in A(x)\} \tag{3.1.3}$$

be the set of feasible state–action pairs, which is assumed to be a Borel subset of $X \times A$. (In the terminology of Definition B.1, $\mathbb{K}$ is the graph of the multifunction $x \mapsto A(x)$.) Then Q is a stochastic kernel on X given $\mathbb{K}$ that represents the transition probability from a state x_t to x_{t+1} under a control action $a_t \in A(x_t)$; that is, for each $t = 0, 1, \ldots$, $B \in \mathcal{B}(X)$, and $(x, a) \in \mathbb{K}$,

$$Q(B|x, a) := \text{Prob}[x_{t+1} \in B | x_t = x, a_t = a]. \tag{3.1.4}$$

Since this holds for all t, we say that the kernel Q is *stationary* or *time–homogeneous* or *time–invariant*. Finally, c in (3.1.2) is a real–valued function on $\mathbb{K}$ that is used to define the OCP's objective function, below.

Remark 3.1. (a) Consider the SM (3.1.1), and suppose that the ξ_t are independent and identically distributed (i.i.d.) random

variables with a common distribution μ on S. Then, from (3.1.4), we can see that the stochastic kernel Q is given by

$$\begin{aligned} Q(B|x,a) &= \text{Prob}[F(x,a,\xi) \in B] \\ &= \mu(\{s \in S : F(x,a,s) \in B\}) \\ &= E[I_B(F(x,a,\xi))] \end{aligned} \tag{3.1.5}$$

where ξ represents a generic random variable with distribution μ, and I_B denotes the *indicator function* of the set B, that is,

$$I_B(x) := \begin{cases} 1 & \text{if} \quad x \in B, \\ 0 & \text{otherwise} . \end{cases}$$

Hence, one can easily go from the SM (3.1.1) to the MCM (3.1.2). One can also go the other way around, from (3.1.2) to (3.1.1), but this is not very helpful because it is obtained by means of an "existence" (nonconstructive) proof. (See Gihman and Skorohod (1979), Sect. 1.1.)

(b) Similarly, if we are given the *deterministic system* (3.0.1), then (3.1.4) gives $Q(B|x,a) = I_B[F(x,a)]$ or, equivalently,

$$Q(B|x,a) = \delta_{F(x,a)}(B), \tag{3.1.6}$$

where $\delta_{F(x,a)}$ denotes the Dirac (or point) measure concentrated at $F(x,a)$, i.e.,

$$\delta_{F(x,a)}(B) := \begin{cases} 1 & \text{if} \quad F(x,a) \in B, \\ 0 & \text{otherwise} . \end{cases}$$

◇

In view of Remark 3.1, in the following we will mainly (but not exclusively) work with the MCM (3.1.2).

Another key difference between a SM and a MCM is that (3.1.1) gives explicitly the state and action processes $\{x_t\}$ and $\{a_t\}$, whereas in a MCM it is unclear, at the outset, how to obtain these processes from (3.1.2). We show next how this is done, but first we need to formalize the notion of "randomized policy" introduced in the Definition 3.2.

The remainder of this section is too technical. The reader might wish to skip it and go directly to Sect. 3.2.

Consider the Markov control model (3.1.2) and, for each $t = 0, 1, \ldots$, define the space H_t of *admissible histories* up to time t as $H_0 := X$, and

$$H_t := \mathbb{K}^t \times X = \mathbb{K} \times H_{t-1} \quad \text{for} \quad t = 1, 2, \ldots, \tag{3.1.7}$$

where $\mathbb{K}$ is the set in (3.1.3). A generic element h_t of H_t, which is called an admissible t–history, or simply a t–history, is a vector of the form

$$h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t), \tag{3.1.8}$$

with $(x_i, a_i) \in \mathbb{K}$ for $i = 0, \ldots, t-1$, and $x_t \in X$.

Definition 3.2. A *randomized control policy*—more briefly, a control policy or simply a *policy*—is a sequence $\pi = \{\pi_t, t = 0, 1, \ldots\}$ of stochastic kernels π_t on the control set A given H_t satisfying the constraint

$$\pi_t(A(x_t)|h_t) = 1 \quad \forall\, h_t \in H_t,\ t = 0, 1, \ldots. \tag{3.1.9}$$

The set of all policies is denoted by Π.

Roughly, a policy $\pi = \{\pi_t\}$ may be interpreted as defining a sequence $\{a_t\}$ of A–valued random variables, called actions (or controls), such that for every t–history h_t as in (3.1.8) and $t = 0, 1, \ldots$, the distribution of a_t is $\pi_t(\cdot|h_t)$, which, by (3.1.9), is concentrated on $A(x_t)$, the set of feasible actions in state x_t. This interpretation of π is made rigorous in equation (3.1.10b), below.

Remark 3.3. The canonical construction. Consider the MCM (3.1.2) and let $(\Omega, \mathcal{F})$ be the measurable space consisting of the (canonical) sample space $\Omega := (X \times A)^\infty$, that is, the space of sequences $\omega = (x_0, a_0, x_1, a_1, \ldots)$ with x_t in X and a_t in A for all $t = 0, 1, \ldots$, and $\mathcal{F}$ is the corresponding product σ–algebra. The proyections (or coordinate variables) $\omega \mapsto x_t$ and $\omega \mapsto a_t$, from Ω to X and A, respectively, are called *state* and *action* variables. Observe that Ω contains the space H_∞, in (3.1.7), of *admissible histories* $\omega = (x_0, a_0, x_1, a_1, \ldots)$ with $(x_t, a_t) \in \mathbb{K}$ (that is, by (3.1.3), $a_t \in A(x_t)$) for all $t = 0, 1, \ldots$.

Let $\pi = \{\pi_t\}$ be an arbitrary control policy and ν an arbitrary probability measure on X, referred to as the "initial distribution". Then, by a theorem of C. Ionescu–Tulcea (Proposition C.8 and Remark C.9 in Appendix C), there exists a unique probability measure P_ν^π on $(\Omega, \mathcal{F})$ which, by (3.1.9), is supported on H_∞, namely, $P_\nu^\pi(H_\infty) = 1$. Moreover, for all $B \in \mathcal{B}(X), C \in \mathcal{B}(A)$, and $h_t \in H_t$ as in (3.1.8), $t = 0, 1, \ldots$, we have:

$$P_\nu^\pi(x_0 \in B) = \nu(B), \tag{3.1.10a}$$

$$P_\nu^\pi(a_t \in C | h_t) = \pi_t(C|h_t), \tag{3.1.10b}$$

$$P_\nu^\pi(x_{t+1} \in B | h_t, a_t) = Q(B|x_t, a_t). \tag{3.1.10c}$$

◇

Definition 3.4. The stochastic process $(\Omega, \mathcal{F}, P_\nu^\pi, \{x_t\})$ is called a discrete–time *Markov control process* (or Markov *decision* process).

The process $\{x_t\}$ in Definition 3.4 depends, of course, on the particular policy π being used and on the given initial distribution ν. Hence, strictly speaking, we should write, for instance, $x_t^{\pi,\nu}$ instead of just x_t. However, we shall keep the simpler notation x_t for it will always be clear from the context what particular π and ν are being used.

The expectation operator with respect to P_ν^π is denoted by E_ν^π. If ν is concentrated at the "initial state" $x \in X$, then we write P_ν^π and E_ν^π as P_x^π and E_x^π, respectively.

3.2 Markov Control Processes: Finite Horizon

In this section we consider the Markov control model (MCM)

$$(X, A, \{A(x) : x \in X\}, Q, c) \tag{3.2.1}$$

introduced in (3.1.2)–(3.1.4), and the Markov control processes (MCP) in Definition 3.4. The optimal control problem (OCP)

we are concerned with is to minimize the finite–horizon objective function (or performance criterion)

$$J(\pi, x) := E_x^{\pi}\left[\sum_{t=0}^{N-1} c(x_t, a_t) + c_N(x_N)\right], \tag{3.2.2}$$

which is a stochastic analogue of (2.0.1). Thus, letting

$$J^*(x) := \inf_{\pi} J(\pi, x),\ x \in X, \tag{3.2.3}$$

be the corresponding **value function** or **minimum cost function**, the OCP we are dealing with is to find an **optimal policy**, that is, a policy $\pi^* \in \Pi$ such that

$$J(\pi^*, x) = J^*(x) \quad \forall\, x \in X. \tag{3.2.4}$$

To this end, we will prove below the stochastic version of the Dynamic Programming (DP) Theorem 2.3.

Let $\pi = \{\pi_t\}$ be an arbitrary policy and, for every $t = 0, 1, \ldots, N$, let $C_t(\pi, x)$ be the "cost–to–go" or cost from time t onwards when using the policy π and given $x_t = x$; that is, for $t = 0, 1, \ldots, N-1$,

$$C_t(\pi, x) := E^{\pi}\left[\sum_{n=t}^{N-1} c(x_n, a_n) + c_N(x_N) | x_t = x\right] \tag{3.2.5}$$

and

$$C_N(\pi, x) := E^{\pi}[c_N(x_N) | x_N = x] = c_N(x). \tag{3.2.6}$$

In particular, from (3.2.2),

$$J(\pi, x) = C_0(\pi, x). \tag{3.2.7}$$

Moreover, for $t = 0, \ldots, N$, let J_t be the *optimal cost from time t* to N, that is,

$$J_t(x) := \inf_{\pi} C_t(\pi, x) \quad \forall\, x \in X,\ t = 0, \ldots, N. \tag{3.2.8}$$

The following lemma states that the functions J_t satisfy the **DP equation** (3.2.9) with the condition (3.2.10).

Lemma 3.5. Suppose that for each $t \in \{0, \dots, N-1\}$ there is a policy π^t for which the minimum is attained in (3.2.8) for all $x \in X$, that is $J_t(x) = C_t(\pi^t, x)$. Then, for each $t = N-1$, $N-2, \dots, 0$ and $x \in X$,

$$J_t(x) = \min_{a \in A(x)} [c(x,a) + \int_X J_{t+1}(y) Q(dy|x,a)] \tag{3.2.9}$$

and

$$J_N(x) = c_N(x). \tag{3.2.10}$$

Remark 3.6. Consider a system model (SM) as in (3.1.1), that is,

$$x_{t+1} = F(x_t, a_t, \xi_t) \quad \forall\, t = 0, 1, \dots$$

with S–valued i.i.d. random disturbances ξ_t with a common distribution G. Then, in analogy with the right–hand side of (2.1.9), the integral in (3.2.9) becomes

$$\begin{aligned} E[J_{t+1}(F(x_t, a_t, \xi_t))|x_t = x, a_t = a] &= E[J_{t+1}(F(x, a, \xi_t))] \\ &= \int J_{t+1}(F(x,a,s)) G(ds). \end{aligned}$$

In a general MCP, the integral in (3.2.9) can be expressed (by (3.1.10c)) as

$$E[J_{t+1}(x_{t+1})|x_t = x, a_t = a] = \int J_{t+1}(y) Q(dy|x,a). \qquad \diamond$$

Proof of Lemma 3.5. From (3.2.6), the terminal condition (3.2.10) is obvious. Now let $t = 0, 1, \dots, N-1$, and consider a policy π such that $\pi_t = f \in \mathbb{F}$ and $\{\pi_{t+1}, \dots, \pi_{N-1}\}$ is an optimal policy from time $t+1$ onwards. Hence, by (3.2.5),

$$\begin{aligned} C_t(\pi, x) &= c(x, f(x)) + E^\pi \left[\sum_{n=t+1}^{N-1} c(x_n, a_n) + c_N(x_N) | x_t = x, a_t = f(x) \right] \\ &= c(x, f(x)) + \int_X J_{t+1}(y) Q(dy|x, f(x)) \\ &\geq \min_{a \in A(x)} \left[c(x,a) + \int_X J_{t+1}(y) Q(dy|x,a) \right]; \end{aligned}$$

therefore,

$$J_t(x) \geq \min_{a \in A(x)} [c(x,a) + \int_X J_{t+1}(y) Q(dy|x,a)].$$

To obtain the reverse inequality observe that, for any $f \in \mathbb{F}$, (3.2.8) yields

$$J_t(x) \leq c(x, f(x)) + \int_X J_{t+1}(y) Q(dy|x, f(x)).$$

Since $f \in \mathbb{F}$ was arbitrary, the latter inequality gives that

$$J_t(x) \leq \min_{a \in A(x)} [c(x,a) + \int_X J_{t+1}(y) Q(dy|x,a)] \ \ \forall\, x \in X.$$

This completes the proof of (3.2.9). □

Remark 3.7. To simplify the notation, if we have a function $g(x,a)$ on $\mathbb{K}$ and use a selector $f \in \mathbb{F}$, we will simply write $g(x,f)$ or $g_f(x)$ in lieu of $g(x, f(x))$. ◇

From Lemma 3.5 we easily obtain the *DP algorithm* in the following theorem.

Theorem 3.8. *Let $\{J_0, \ldots, J_T\}$ be the functions in (3.2.9)–(3.2.10). Suppose that, for each $t = 0, 1, ..., N-1$, there is a selector $f_t \in \mathbb{F}$ such that $f_t(x) \in A(x)$ attains the minimum in (3.2.9) for every $x \in X$, that is (using the notation in Remark 3.7),*

$$J_t(x) = c(x, f_t) + \int_X J_{t+1}(y) Q(dy|x, f_t). \tag{3.2.11}$$

Then the deterministic Markov policy $\pi^* = \{f_0, \ldots, f_{N-1}\}$ is optimal, that is, it satisfies (3.2.4).

Proof. This theorem follows directly from (3.2.11) and (3.2.8), which yield

$$\inf_\pi C_t(\pi, x) = C_t(\pi^*, x)$$

for every $t = 0, \ldots, T$ and $x \in X$. □

Remark 3.9. (a) Lemma 3.5 and Theorem 3.8 hold, of course, if the cost function c and the stochastic kernel Q are time–varying, that is, c_t and Q_t for $t = 0, 1, ...$ as in Theorem 2.3.

(b) Our approach in this section to obtain the DP Theorem 3.8 is a little different from the approach followed in Sect. 2.1 to obtain Theorem 2.3. Indeed, here we introduced the functions J_t in (3.2.8) and then we showed that they satisfy the DP algorithm (3.2.9)–(3.2.10). In contrast, in Theorem 2.3 we started the other way around: first, we introduced the DP algorithm (2.1.8)–(2.1.9), and then we obtained (2.1.10)–(2.1.11). ◇

Remark 3.10. Variants of the D.P. equation.

(a) Nonstationary MCPs. Lemma 3.5 and Theorem 3.8 hold if the MCM (3.2.1) is *nonstationary*, that is, all of the components in (3.2.1) are time–varying, i.e, $X_t, A_t, A_t(x), Q_t, c_t$ for $t = 0, 1, \ldots$. For instance, (3.2.9) would be expressed as

$$J_t(x) = \min_{a \in A_t(x)} \left[c_t(x, a) + \int_{X_{t+1}} J_{t+1}(y) Q_t(dy|x, a) \right].$$

Moreover, there is a well–known *state–augmentation* procedure to transform a nonstationary MCM into a *stationary* one. [See, for instance, Bertsekas and Shreve (1978), Guo et al. (2010), Hernández-Lerma (1989), Hinderer et al. (2016).]

(b) Discounted costs. Given a discount factor $\alpha \in (0, 1)$, the performance criterion (3.2.2) becomes

$$E_x^\pi \left[\sum_{t=0}^{N-1} \alpha^t c(x_t, a_t) + \alpha^N c_N(x_N) \right]$$

Then, using the same change of variable to obtain (2.2.1)–(2.2.2), the DP equation in Lemma 3.5 becomes

$$V_t(x) = \min_{a \in A(x)} \left[c(x, a) + \alpha \int_X V_{t+1}(y) Q(dy|x, a) \right] \qquad (3.2.12)$$

for $t = 0, 1, \ldots, N-1$, and

$$V_N(x) = c_N(x). \tag{3.2.13}$$

(c) System models. Consider the SM in Remark 3.6. Then (3.2.10) remains the same, but (3.2.9) becomes

$$\begin{aligned} J_t(x) &= \min_{a \in A(x)} \{c(x,a) + E[J_{t+1}(F(x,a,\xi))]\} \\ &= \min_{a \in A(x)} \{c(x,a) + \int_S J_{t+1}(F(x,a,s))G(ds)\}. \end{aligned} \tag{3.2.14}$$

In particular, in the discounted case (3.2.12), the expected value in the right–hand side of (3.2.14) is replaced with

$$\alpha E[J_{t+1}(F(x,a,\xi))].$$

Finally, we can also rewrite (3.2.9)–(3.2.10) in a *forward form*, similar to the deterministic case in (2.2.4)–(2.2.5).

3.3 Conditions for the Existence of Measurable Minimizers

For Theorem 3.8 to be useful we need to ensure the existence of measurable selectors (or "minimizers") $f_t \in \mathbb{F}$ as in (3.2.11). There are many ways of doing this—see, for instance, Bertsekas and Shreve (1978), Bäuerle and Rieder (2011), Hernández-Lerma and Lasserre (1996), Hinderer et al. (2016), to mention just a few references. Here we will use some results in Appendix B, below, to see how to ensure the existence of measurable minimizers. First, we consider how to adapt our MCM to Theorem B.3.

Theorem 3.11. *Consider the MCM (3.2.1), and a l.s.c. function $u : X \to \mathbb{R}$ bounded below. Suppose that, for every $x \in X$,*

(a) The action set $A(x)$ is compact;

(b) The function $a \mapsto c(x,a)$ is l.s.c. on $A(x)$;

(c) The function $a \mapsto v'(x,a) := \int_X v(y)Q(dy|x,a)$ *is l.s.c. on* $A(x)$ *for each continuous and bounded function* v *on* X.

Then the function u^* *defined on* X *by*

$$u^*(x) := \inf_{a\in A(x)} \left[c(x,a) + \int_X u(y)Q(dy|x,a) \right] \tag{3.3.1}$$

is measurable and there exists a selector $f \in \mathbb{F}$ *such that* $f(x) \in A(x)$ *attains the minimum in (3.3.1) for every* $x \in X$, *i.e.,*

$$u^*(x) = c(x,f) + \int_X u(y)Q(dy|x,f). \tag{3.3.2}$$

Proof. By the present hypotheses, the result will follow from Theorem B.3 provided that the integral in (3.3.1) is l.s.c. in $a \in A(x)$ for every $x \in X$. Thus, we need to show that, for each $x \in X$, if $\{a_n\}$ is a sequence in $A(x)$ converging to $a \in A(x)$, then

$$\liminf_{n\to\infty} \int_X u(y)Q(dy|x,a_n) \geq \int_X u(y)Q(dy|x,a). \tag{3.3.3}$$

To this end, we will use that u is l.s.c. and bounded below and, therefore, by Proposition A.2, there is a sequence of continuous and bounded functions v_k on X such that $v_k \uparrow u$. Hence, for every n and k,

$$\int_X u(y)Q(dy|x,a_n) \geq \int_X v_k(y)Q(dy|x,a_n),$$

and, letting $n \to \infty$, the assumption (c) yields

$$\liminf_{n} \int_X u(y)Q(dy|x,a_n) \geq \int_X v_k(y)Q(dy|x,a).$$

Finally, letting $k \to \infty$, (3.3.3) follows. □

Remark 3.12. (a) Given a real–valued function v on X, let v' be as in Theorem 3.11(c), i.e.,

$$v'(x,a) := \int_X v(y)Q(dy|x,a) \quad \forall\ (x,a) \in \mathbb{K}. \tag{3.3.4}$$

The stochastic kernel (or transition probability) Q in (3.1.2) is said to be *weakly continuous* (or that it has the *weak–Feller property*) if the mapping $(x,a) \mapsto v'(x,a)$ is continuous on $\mathbb{K}$ for every *continuous* bounded function v. On the other hand, Q is called *strongly continuous* (or that it satisfies the *strong–Feller property*) if v' is continuous for every *measurable* bounded function v. It should be noted that here we will be dealing with the weakly continuous case.

Clearly, since every continuous function is measurable, strong continuity implies weak continuity, but not conversely. The following examples emphasize this fact. First, consider the system model (3.1.1), where the ξ_t are i.i.d. random variables with common distribution G on S. Then we can express (3.3.4) as

$$v'(x,a) = E[v(F(x,a,\xi))] = \int_S v(F(x,a,s))G(ds), \quad (3.3.5)$$

where ξ is a generic random variable with distribution G. Then, for an *arbitrary* measurable function v, the function v' is not necessarily continuous, even if the map $(x,a) \mapsto F(x,a,s)$ is continuous for every $s \in S$. (Take, for instance, $v = I_B$ the indicator function of a set B.) Now suppose that (3.1.1) is the *additive noise* system

$$x_{t+1} = F(x_t, a_t) + \xi_t, \quad t = 0, 1, ... \quad (3.3.6)$$

Moreover, the ξ_t are i.i.d. disturbances as in (3.3.5), but now they take values in $X = S = \mathbb{R}^d$ and, in addition, the distribution G has a continuous and bounded probability density g. Then, using the change of variable $y = F(x,a) + s$, (3.3.5) becomes

$$v'(x,a) = \int_S v(F(x,a)+s)g(s)ds = \int_S v(y)g(y - F(x,a))dy. \quad (3.3.7)$$

Therefore, assuming that the mapping $(x,a) \mapsto F(x,a)$ is continuous, it follows that the function v' in (3.3.7) is continuous, even if the bounded function v is just measurable! In other words, the additive noise system (3.3.6) is strongly continuous.

(b) For the *deterministic system* (3.0.1), the transition probability Q turns out to be the Dirac measure (3.1.6) and so the function v' in (3.3.4) becomes

$$v'(x,a) = v(F(x,a)).$$

Therefore, Q is weakly continuous (in the sense of part (a)) if the system function $F : \mathbb{K} \to X$ is continuous, as in Assumption 2.17. However, requiring that Q is strongly continuous (so that v' is continuous for every measurable bounded function v) would be extremely restrictive.

(c) As in the proof of (3.3.3) (replacing u with v), it can be seen that: *If Q is weakly continuous and $v : X \to \mathbb{R}$ is l.s.c. and bounded below, then the mapping $(x,a) \mapsto v'(x,a)$ in (a) is l.s.c. and bounded below on $\mathbb{K}$.* ◇

Theorem 3.11 requires the action sets $A(x)$ to be compact. This requirement is replaced, in the following theorem, by inf–compactness on $\mathbb{K}$ (see Definition B.4(a3)).

Theorem 3.13. *Consider the MCM (3.2.1) and let $u : X \to \mathbb{R}$ be as in Theorem 3.11, that is, u is l.s.c. and bounded below. In addition, let us assume that*

(a) the cost function c is l.s.c., bounded below, and inf–compact on $\mathbb{K}$;

(b) the transition probability Q is weakly continuous (see Remark 3.12(a)).

Then the conclusions of Theorem 3.11 hold again, that is, the function u^* in (3.3.1) is measurable, and there exists a selector $f \in \mathbb{F}$ such that (3.3.2) holds for every $x \in X$. Moreover, for each $x \in X$, let $A_u(x)$ be the set of actions $a^* \in A(x)$ where the right-hand side of (3.3.1) attains the minimum, i.e.,

$$u^*(x) = c(x,a^*) + \int_X u(y)Q(dy|x,a^*),$$

and suppose that the multifunction $x \mapsto A_u(x)$ is l.s.c. (see Definition B.4(b)). Then the function u^* is l.s.c. on X.

Proof. See Exercise 3.1(a). □

In Theorem 3.13, to conclude that u^* is l.s.c., *we assumed* that the multifunction $x \mapsto A_u(x)$ is l.s.c. In the following Theorem 3.14, to obtain that u^* is l.s.c., we replace the assumption on $x \mapsto A_u(x)$ by the condition that the stage cost c is $\mathbb{K}$–inf–compact (see Lemma 2.16 or Definition B.4(a2)).

Theorem 3.14. *Consider the MCM (3.2.1), and let $u : X \to \mathbb{R}$ be l.s.c. and bounded below. In addition,*

(a) the stage cost $c : \mathbb{K} \to \mathbb{R}$ is nonnegative and $\mathbb{K}$–inf–compact; and

(b) the transition probability Q is weakly continuous.

Then the conclusions in Theorem 3.11 hold, that is, the function u^* in (3.3.1) is measurable and there exists $f \in \mathbb{F}$ that satisfies (3.3.2). Moreover, u^* is l.s.c.

Proof. See Exercise 3.1(b). □

3.4 Examples

Example 3.15 (Stochastic LQ systems) We now consider a stochastic version of the LQ system in Example 2.4. The state and action spaces are $X = A = \mathbb{R}$ and the dynamics is given by

$$x_{t+1} = \gamma x_t + \beta a_t + \xi_t, \ t = 0, 1, \ldots, N-1, \tag{3.4.1}$$

with nonzero coefficients γ and β. The disturbances ξ_t are i.i.d. random variables, independent of the initial state x_0, with a common distribution G that has zero mean and finite variance, i.e.,

$$E(\xi) = 0 \quad \text{and} \quad \sigma^2 := E(\xi^2) < \infty, \tag{3.4.2}$$

where ξ is a generic real–valued random variable with distribution G. The performance criterion is as in (3.2.2), with one–stage costs

$$c(x, a) = qx^2 + ra^2 \quad \text{and} \quad c_N(x) = q_N x^2, \tag{3.4.3}$$

with nonnegative coefficients q and q_N, and $r > 0$. Hence, from Lemma 3.5 and Remark 3.6, the D.P. equation (3.2.9)–(3.2.10) becomes

$$J_t(x) = \min_a [qx^2 + ra^2 + EJ_{t+1}(\gamma x + \beta a + \xi)] \tag{3.4.4}$$

for $t = N-1,\ N-2, ..., 0$, and

$$J_N(x) = q_N x^2. \tag{3.4.5}$$

Note that, from (3.4.5) and (3.4.2),

$$\begin{aligned} EJ_N(\gamma x + \beta a + \xi) &= q_N E(\gamma x + \beta a + \xi)^2 \\ &= q_N(\gamma^2 x^2 + \sigma^2 + 2\gamma\beta x a + \beta^2 a^2). \end{aligned}$$

Inserting this quantity in the right–hand side of (3.4.4) and minimizing over all $a \in A$ we obtain the minimizer $f_{N-1} \in \mathbb{F}$ given by

$$f_{N-1}(x) = G_{N-1}x, \quad \text{with} \quad G_{N-1} := -(r + q_N\beta^2)^{-1} q_N \gamma\beta.$$

Replacing this value of $a = f_{N-1}(x)$ in (3.4.4) we obtain

$$J_{N-1}(x) = K_{N-1}x^2 + q_N\sigma^2 \quad \forall\ x \in \mathbb{R}$$

with

$$K_{N-1} := [1 - (r + q_N\beta^2)^{-1} q_N \beta^2] q_N \gamma^2 + q.$$

Similarly, we replace J_{N-1} in (3.4.4) to obtain the minimizer $f_{N-2} \in \mathbb{F}$ and the function J_{N-2}. In general, by backward induction, we obtain the optimal policy $\pi^* = \{f_0, \ldots, f_{N-1}\}$ with

$$f_t(x) = G_t x, \quad \text{where} \quad G_t := -(r + K_{t+1}\beta^2)^{-1} K_{t+1}\gamma\beta, \tag{3.4.6}$$

with

$$K_t = [1 - (r + K_{t+1}\beta^2)^{-1} K_{t+1}\beta^2] K_{t+1}\gamma^2 + q$$

for $t = N-1, ..., 1, 0$. The optimal cost function from time t to N (see (3.2.8)), with J_N in (3.4.5) and $t = 0, 1, \ldots, N-1$, is

$$J_t(x) = K_t x^2 + \sigma^2 \sum_{n=t+1}^{N} K_n. \tag{3.4.7}$$

In particular, from (3.2.4) and (3.2.8),

$$J^*(x) = J_0(x) = K_0 x^2 + \sigma^2 \sum_{n=1}^{N} K_n$$

for every initial state $x_0 = x$. ◇

Example 3.16. Stochastic LQ system with discounted cost. Let us consider again the LQ system (3.4.1)–(3.4.3), except that now we use the α-*discounted* DP equation (3.2.12) with terminal cost $V_N(\cdot) \equiv 0$ in (3.2.13). The optimal cost is again of the form

$$J_0(x) = K_0 x^2 + \sigma^2 \sum_{t=1}^{N-1} K_t$$

with K_t given by a backward recursion from $t = N-1, ..., 1, 0$ by

$$K_t = [1 - (r\alpha^t + K_{t=1}\beta^2)^{-1} K_{t+1}\beta^2] K_{t+1}\gamma^2 + q\alpha^t$$

with $K_N = 0$.

To complement these results, note that Sect. 3.5 in Hernández-Lerma and Lasserre (1996) as well as Sect. 2.1 in Bertsekas (1987) present the *nonstationary vector case* of the LQ system (3.4.1)–(3.4.2) in which the state and control variables x_t and a_t are vectors in, say, $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and the coefficients $\gamma_t, \beta_t, q_t, r_t$ are matrices of suitable dimensions. On the other hand, observe that the optimal control f_t in (3.4.6) *is the same as in the deterministic case* (2.1.16). This property is sometimes called the *certainty–equivalence* principle of LQ systems. This principle is shared by other stochastic systems. For an extension of the certainty–equivalence principle to a class of stochastic differential *games* see Josa-Fombellida and Rincón-Zapatero (2019). ◇

Example 3.17 (A consumption–investment problem). At each time $t = 0, 1, \ldots, N-1$, an investor wishes to allocate his/her current wealth x_t between investment (a_t) and consumption $(x_t - a_t)$. Hence, if $x_t = x$, the corresponding investment or control set is $A(x) = [0, x]$. The wealth at time $t+1$ is

$$x_{t+1} = a_t \xi_t, \ t = 0, 1, \ldots, N-1,$$

that is, the wealth is proportional to the amount invested at time t, where $\xi_t \geq 0$ denotes a "random interest rate". These "disturbances" ξ_t are i.i.d. random variables, assumed to be independent of the initial wealth x_0. Moreover, to ensure that reinvestment is indeed profitable, we assume that the ξ_t have a (finite) mean $m = E(\xi_t) > 1$. The investor's problem is to find an investment strategy $\pi = \{a_0, a_1, \ldots, a_{N-1}\}$ that maximizes an expected total discounted utility from consumption defined as

$$J(\pi, x) := E_x^\pi \left[\sum_{t=0}^{N-1} \alpha^t u(x_t - a_t) \right],$$

for every initial wealth $x_0 = x \geq 0$, where α is a discount factor, and $u(x - a)$ is a given utility of consumption. We will take the state and action spaces as $X = A = [0, \infty)$.

Since we are now maximizing, the discounted D.P. equation (3.2.12)–(3.2.13) becomes, for every $x \in X$ and $t = N - 1$, $N - 2, \ldots, 0$,

$$J_N(x) = 0, \tag{3.4.8}$$

$$J_t(x) = \max_{a \in A(x)} [u(x - a) + \alpha E J_{t+1}(a\xi_t)]. \tag{3.4.9}$$

The solution to (3.4.8)–(3.4.9) depends, of course, on the utility function u. Here, we will consider two cases.

Linear case. Suppose that $u(x - a) = b \cdot (x - a)$ for some $b > 0$. We will assume that $\alpha m > 1$. In this case, the optimal investment policy is $\pi^* = \{f_0, \ldots, f_{N-1}\}$ with $f_{N-1}(x) = 0$, and

$$f_t(x) = x \quad \forall\, t = 0, 1, \ldots, N - 2. \tag{3.4.10}$$

Moreover, for all $x \in X$, and $t = N - 1, \ldots, 0$,

$$J_t(x) = (m\alpha)^{N-t-1} bx. \tag{3.4.11}$$

In particular, $J^*(x) = J_0(x) = (m\alpha)^{N-1} bx$, for any initial state $x_0 = x$.

Nonlinear case. Assume that $u(x - a) = (b/k)(x - a)^k$. Here $b > 0$ and $0 < k < 1$. The function $u(x) = (b/k)x^k$ is called the

isoelastic utility function or *power utility function*. In this case, the solution of (3.4.8)–(3.4.9) is

$$J_t(x) = (b/k)D_t x^k \quad \forall\, t = N-1, N-2, \ldots, 0, \tag{3.4.12}$$

whereas the optimal maximizers are $f_{N-1}(x) = 0$ and, for $t = N-2, \ldots, 0$,

$$f_t(x) = x/[1 + \delta D_{t+1}^{1/(k-1)}], \tag{3.4.13}$$

with $\delta = (\alpha E\xi^k)^{1/(k-1)}$, and D_t is given recursively by $D_{N-1} = 1$ and for $t = N-2, \ldots, 0$,

$$D_t = \delta^{k-1} D_{t+1}/[1 + \delta D_{t+1}^{1/(k-1)}]^{k-1}. \tag{3.4.14}$$

(See Exercise 3.4.) ◇

3.5 Infinite–Horizon Discounted Cost Problems

In this section we consider the Markov control model (MCM) (3.1.2):

$$(X, A, \{A(x) : x \in X\}, Q, c).$$

The objective function to be minimized is the infinite–horizon expected total **discounted cost**

$$V(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right] \tag{3.5.1}$$

for every policy $\pi \in \Pi$, and every initial state $x \in X$, where $\alpha \in (0, 1)$ is a given discount factor. A policy π^* such that

$$V(\pi^*, x) = \inf_\pi V(\pi, x) =: V^*(x) \quad \forall\, x \in X \tag{3.5.2}$$

is said to be α–*discount optimal*, and V^* is called the α–*discount value function*, also known as the α–*discount optimal cost function*.

We can analyze the infinite-horizon discounted cost problem in the context of any of the Theorems 3.11, 3.13 or 3.14. Here,

however, to fix ideas we will follow Theorem 3.13. (The analysis following Theorem 3.14 is quite similar.)

The following assumption, which is supposed to hold throughout this section, includes some of the hypotheses of Theorem 3.13. (See also Theorem 3.13 or Remark 3.12(a) for the definition of *weak continuity* of Q.)

Assumption 3.18. (a) The stage cost $c : \mathbb{K} \to \mathbb{R}$ is l.s.c., nonnegative, and inf-compact on $\mathbb{K}$, that is (by Definition B.4 (a3)), for each $r > 0$, the set $\{a \in A(x) | c(x, a) \leq r\} \subset A$ is compact.

(b) The multifunction $x \mapsto A_c(x)$ is l.s.c. (see Definition B.4(b)), where, for every $x \in X$, $A_c(x)$ stands for the set of control actions $a^* \in A(x)$ such that $c(x, a^*) = \min_{a \in A(x)} c(x, a)$.

(c) The transition law Q is weakly continuous.

Observe that the LQ system in Examples 3.15 and 3.16 satisfies Assumption 3.18. In particular, since $c(x, a) = qx^2 + ra^2$ (see (3.4.3)) and there are no control constraints (i.e, $A(x) = A = \mathbb{R}$ for all $x \in X$) the multifunction $x \mapsto A_c(x)$ in Assumption 3.18(b) is "constant", that is, $A_c(x) = \{0\}$ for all $x \in X$; hence, it is trivially l.s.c. Similarly, Q is weakly continuous because, from (3.4.1) and the Remark 3.12(a), the mapping

$$(x, a) \mapsto \int_X v(y) Q(dy|x, a) = \int_{\mathbb{R}} v(\gamma x + \beta a + s) G(ds)$$

is continuous for every continuous bounded function v. Assumption 3.18 can also be verified for the Example 3.17. Nevertheless, since the problem aims to *maximize* a utility, the reader might wish to make the "change of variable" $c(x, a) := -u(x - a)$ to put Example 3.17 in the format of Assumption 3.18.

In our present context, the *Basic Assumption* in Chap. 1 is that

(a) the cost function c is nonnegative, and

(b) the value function V^* is finite–valued.

Part (a) is included in Assumption 3.18(a). Part (b) follows from the following Assumption 3.19.

Assumption 3.19. There exists a policy π such that $V(\pi, x) < \infty$ for each $x \in X$.

We denote by Π^0 the family of policies that satisfy Assumption 3.19. For instance, if the cost c is *bounded*, say $0 \le c \le M$, then

$$0 \le V(\pi, x) \le M(1 - \alpha) \quad \forall\, \pi \in \Pi, x \in X.$$

Hence, in this case, $\Pi^0 = \Pi$.

In the remainder of this section we roughly follow the *dynamic programming* (DP) approach in Sect. 2.3.1 for *deterministic* infinite–horizon problems. In particular, we wish to prove that the α–discount value function V^* in (3.5.2) satisfies the DP equation—compare (2.3.7) and (3.5.5). To this end, we again consider the family $L^+(X)$ of nonnegative l.s.c. functions on X, introduced in Lemma 2.18. (See also the Exercise 2.10.)

For each $u \in L^+(X)$, let $Tu : X \to X$ be defined as

$$Tu(x) := \inf_{a \in A(x)} \left[c(x, a) + \alpha \int_X u(y) Q(dy|x, a) \right]. \tag{3.5.3}$$

One of the main results in this section is that the value function V^* is a solution of the *α–discount DP equation* (also known as the *α–discount optimality equation*)

$$u(x) = \inf_{a \in A(x)} \left[c(x, a) + \alpha \int_X u(y) Q(dy|x, a) \right]. \tag{3.5.4}$$

By (3.5.3), a solution $u \in L^+(X)$ to (3.5.4) is a *fixed point* of T, that is, $Tu = u$.

Remark 3.20. The following Theorem 3.21 is quite similar to Theorem 4.2.3 in Hernández-Lerma and Lasserre (1996) except for a key difference: In the latter reference, the transition law Q is supposed to be *strongly continuous* (see Remark 3.12(a)), whereas here Q is *weakly continuous* (Assumption 3.18(c)). This implies (by the Remark 3.12(b)) that Theorem 3.21 *includes as a special case the "deterministic" results in* Theorem 2.21 and Corollary 2.22. This fact is not true in the cited reference. ◇

Theorem 3.21. *Suppose that Assumptions 3.18 and 3.19 hold. Then:*

(a) The α–discount value function V^* is the pointwise minimal solution of (3.5.4); that is, for every $x \in X$,

$$V^*(x) = \min_{a \in A(X)} \left[c(x,a) + \alpha \int_X V^*(y) Q(dy|x,a) \right], \quad (3.5.5)$$

equivalently $V^* = TV^*$, and, in addition, if u is another solution of (3.5.4), then $u(x) \geq V^*(x)$ for all $x \in X$.

(b) There is a selector $f_* \in \mathbb{F}$ such that $f_*(x) \in A(x)$ attains the minimum in (3.5.5), that is,

$$V^*(x) = c(x, f_*) + \alpha \int_X V^*(y) Q(dy|x, f_*) \quad \forall\, x \in X, \quad (3.5.6)$$

and the deterministic stationary policy $f_*^\infty = \{f_*, f_*, \dots\}$ is α–discount optimal; conversely, if $f_*^\infty = \{f_*, f_*, \dots\}$ is α–discount optimal, then $f_* \in \mathbb{F}$ satisfies (3.5.6).

(c) If π^* is a policy such that $V(\pi^*, \cdot)$ satisfies (3.5.4) and, moreover, the condition

$$\lim_{n \to \infty} \alpha^n E_x^\pi V(\pi^*, x) = 0 \quad \forall \pi \in \Pi^0 \text{ and } x \in X \quad (3.5.7)$$

holds, then π^* is α–discount optimal. In other words, if (3.5.7) holds, then π^* is α–optimal if and only if $V(\pi^*, \cdot)$ satisfies the α–discount D.P. equation.

(d) If an α–discount optimal policy exists, then there exists one that is deterministic stationary.

The conclusion of part (d) in Theorem 3.21 is important because it ensures that, even if we work in the space of all randomized, history–dependent policies (see the Remark 1.2(d) or the Remark 3.3), to solve the infinite–horizon discounted cost problem it suffices to consider deterministic stationary policies $f^\infty = \{f, f, ...\}$, with $f \in \mathbb{F}$.

The proof of Theorem 3.21 requires some preliminary results that are important in themselves. The first one is that the operator T in (3.5.3) maps $L^+(X)$ into itself.

Lemma 3.22. If u is a function in $L^+(X)$, then so is Tu and, moreover, there exists a selector $f \in \mathbb{F}$ such that $f(x) \in A(x)$ attains the minimum in the right-hand side of (3.5.3), that is,

$$Tu(x) = c(x, f) + \alpha \int_X u(y) Q(dy|x, f) \quad \forall\, x \in X. \tag{3.5.8}$$

Proof. First, note that the hypotheses (a)-(b) in Theorem 3.13 are the same as the Assumptions 3.18(a),(c). Moreover, our Assumptions 3.18(a),(b),(c) yield that the multifunction $x \mapsto A_u(x)$ in Theorem 3.13 is l.s.c. Therefore, the conclusion in Lemma 3.22 follows from Theorem 3.13. □

The expression (3.5.9) below is the "stochastic version" of (2.3.16). In fact, the arguments in the Remark 2.19 are a simplified version of the arguments in the proof of Lemma 3.23.

Lemma 3.23. Given a selector $f \in \mathbb{F}$, consider the deterministic stationary policy $f^\infty = \{f, f, \dots\}$. Then the α–discounted cost $v_f(x) := V(f^\infty, x)$ satisfies that, for every $x \in X$,

$$v_f(x) = c(x, f) + \alpha \int_X v_f(y) Q(dy|x, f). \tag{3.5.9}$$

Proof. Given $f \in \mathbb{F}$, from (3.5.1) we obtain

$$\begin{aligned} v_f(x) &= E_x^{f^\infty} \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, f) \right] \\ &= E_x^{f^\infty} \left[c(x_0, f) + \sum_{t=1}^{\infty} \alpha^t c(x_t, f) \right] \\ &= c(x, f) + \alpha E_x^{f^\infty}(\Theta), \end{aligned} \tag{3.5.10}$$

where $\Theta := \sum_{t=1}^{\infty} \alpha^{t-1} c(x_t, f)$. Then, by the properties of the conditional expectation and the Markov–like property (3.1.10c),

$$E_x^{f^\infty}(\Theta) = E_x^{f^\infty} \left[E_x^{f^\infty}(\Theta | x_0, a_0, x_1) \right]$$

$$
\begin{aligned}
&= E_x^{f^\infty}\left[E_x^{f^\infty}(\Theta|x_1)\right] \\
&= \int_X E^{f^\infty}\left[\sum_{t=1}^{\infty}\alpha^{t-1}c(x_t,f)|x_1=y\right]Q(dy|x,f) \\
&= \int_X v_f(y)Q(dy|x,f).
\end{aligned}
$$

The latter fact together with (3.5.10) gives (3.5.9). □

Compare part (a) of the following lemma with Lemma 2.20.

Lemma 3.24. (a) If $u \in L^+(X)$ satisfies that $u \geq Tu$, then there exists a selector $f \in \mathbb{F}$ such that $u(x) \geq v_f(x)$ for every $x \in X$. Hence $u \geq V^*$.

(b) Let $u : X \to \mathbb{R}$ be a measurable function such that Tu is well defined. In addition, suppose that $u \leq Tu$ and

$$\lim_{n\to\infty} \alpha^n E_x^\pi[u(x_n)] = 0 \quad \forall\, \pi \in \Pi^0 \text{ and } x \in X. \tag{3.5.11}$$

Then $u \leq V^*$.

Proof. (a) Let $u \in L^+(X)$ be such that $u \geq Tu$. Then, by Lemma 3.22, there exists $f \in \mathbb{F}$ for which, for every $x \in X$,

$$u(x) \geq c(x,f) + \alpha \int u(y)Q(dy|x,f).$$

Iteration of this inequality yields, for every $n = 1, 2, \ldots$ and $x \in X$,

$$u(x) \geq E_x^f\left[\sum_{t=0}^{n-1}\alpha^t c(x_t,f)\right] + \alpha^n E_x^f u(x_n),$$

where $E_x^f u(x_n) = \int u(y)Q^n(dy|x,f)$ and $Q^n(\cdot|x,f)$ is the n–step transition probability of the Markov process $\{x_t\}$ when using the policy f^∞. Therefore, since u is nonnegative,

$$u(x) \geq E_x^f\left[\sum_{t=0}^{n-1}\alpha^t c(x_t,f)\right],$$

and letting $n \to \infty$ we obtain $u(\cdot) \geq v_f(\cdot)$.

(b) First, note that $u \le Tu$ and (3.5.4) imply that, for every $x \in X$ and $a \in A(x)$,

$$u(x) \le c(x,a) + \alpha \int u(y)Q(dy|x,a). \tag{3.5.12}$$

On the other hand, for arbitrary $\pi \in \Pi$ and $x \in X$, the Markov–like property (3.1.10c) yields

$$\begin{aligned} E_x^\pi \left[\alpha^{t+1} u(x_{t+1})|h_t, a_t\right] &= \alpha^{t+1} \int_X u(y)Q(dy|x_t,a_t) \\ &\ge \alpha^t \left[u(x_t) - c(x_t,a_t)\right] \qquad \text{(by (3.5.12))}, \end{aligned}$$

so

$$\alpha^t c(x_t,a_t) \ge -E_x^\pi[\alpha^{t+1}u(x_{t+1}) - \alpha^t u(x_t)|h_t,a_t].$$

Thus, taking expectations E_x^π and summing over $t = 0, \dots, n-1$, we have

$$E_x^\pi \left[\sum_{t=0}^{n-1} \alpha^n c(x_t,a_t)\right] \ge u(x) - \alpha^n E_x^\pi u(x_n) \quad \forall\, n.$$

Finally, letting $n \to \infty$ and using (3.5.11), we obtain that $V(\pi,x) \ge u(x)$ for all $x \in X$. Therefore, since $\pi \in \Pi$ was arbitrary, it follows that $V^* \ge u$. □

The following Lemma 3.25 extends Theorem 2.21 to the stochastic case. It shows the convergence to V^* of the **value iteration** (VI) functions $\{v_n\}$ defined as $v_0 \equiv 0$, and $v_n := Tv_{n-1}$ for all $n = 1, 2, \dots$, that is,

$$v_n(x) := \min_{a \in A(x)} [c(x,a) + \alpha \int_X v_{n-1}(y)Q(dy|x,a)] \tag{3.5.13}$$

for all $x \in X$. The lemma is a key result in the theory and applications of dynamic programming. Before stating it, however, note that the convergence of v_n to V^* is to be expected, because using the *forward form* of the D.P. equation (3.2.9)–(3.2.10) with $c_N \equiv 0$, we have that

$$v_n(x) = \inf_\pi V_n(\pi,x), \tag{3.5.14}$$

where V_n is the n–stage α–discounted cost with zero terminal cost. Hence, since $V_n(\pi, x) \uparrow V(\pi, x)$ as $n \to \infty$, one would expect that, interchanging "lim" and "inf",

$$v_n(x) \uparrow \inf_{\pi} V(\pi, x) = V^*(x) \quad \forall\, x \in X. \tag{3.5.15}$$

Lemma 3.25 confirms that this is indeed the case.

Lemma 3.25 (Convergence of the α—VI functions). Under the Assumptions 3.18 and 3.19:

(a) v_n is in $L^+(X)$ for every n,

(b) the convergence in (3.5.15) holds, and

(c) V^* satisfies the α–discount DP equation (3.5.5).

Proof. Part (a) follows from the Assumption 3.18 (in particular, part (c)) and Lemma 3.22.

(b)–(c) By (3.5.14) and the monotonicity of $\{v_n\}$, there is a function $v \leq V^*$ such that $v_n(x) \uparrow v(x)$ for all $x \in X$. On the other hand, from Lemma 2.15(c) we obtain that v satisfies the α–discount DP equation $v = Tv$. Hence, by Lemma 3.24(a), $v \geq V^*$. This yields (b) and (c). □

We are finally ready for the proof of Theorem 3.21.

Proof of Theorem 3.21.

(a) The DP equation (3.5.5) follows from Lemma 3.25. The fact that V^* is the minimal solution of (3.5.5) is a consequence of Lemma 3.24(a).

(b) The existence of a selector $f_* \in \mathbb{F}$ satisfying (3.5.6) follows from Lemma 3.22. Further, from (3.5.6) and Lemma 3.22, f_*^∞ is optimal. The converse is also obtained from Lemma 3.23.

(c) If $V(\pi^*, .)$ satisfies (3.5.5), then part (a) or Lemma 3.24(a) give that $V(\pi^*, .) \geq V^*(.)$. The reverse inequality follows from (3.5.7) and Lemma 3.24(b).

(d) This part is a consequence of (a) and (b). □

The convergence of the α–VI functions $v_n = Tv_{n-1}$, or $v_n = T^n v_0$, also known as the method of *successive approximations*, is inspired in *Banach's fixed point theorem*. (See Remark 2.25) Value iteration is one of the two most popular algorithms to solve a dynamic programming equation. The other most popular algorithm is the *policy iteration* (PI) algorithm, which we introduce in the next section.

Example 3.26 (Exhaustible resource extraction). The optimal exhaustible resource extraction—also known as a cake-eating problem—is one of the most studied in environmental and resource economics; see for example Hung and Quyen (1994), or Long and Kemp (1984). In this example, we present a discrete version of the continuous model in Dasgupta and Heal (1974); see also Pindyck (1980) for a stochastic version.

Consider an agent that exploits a certain nonrenewable resource. Let x_t and a_t be the stock of the nonrenewable resource and the agent's consumption at time t, respectively. The initial stock of the resource is $x_0 = x > 0$ and the law of motion of x_t is

$$x_{t+1} = \xi_t(x_t - a_t) \quad \text{for} \quad t = 0, 1, 2, ..., \tag{3.5.16}$$

where $\{\xi_t\}_{t=0}^{\infty}$ is a sequence of i.i.d. binomial random variables such that

$$\xi_t = \begin{cases} 1 \text{ with probability } p, \\ d \text{ with probability } 1-p, \end{cases} \tag{3.5.17}$$

at each time step t, where $0 \leq d < 1$, and $0 < p < 1$. The value $1 - d$ can be interpreted as the loss caused by bad natural conditions in the extraction. The state space and control spaces are, respectively, $X = (0, x_0]$, $A = (0, x_0]$, and $A(x) = (0, x]$ for all $x \in X$.

The agent's OCP is to find a consumption trajectory $\{a_t\}_{t=1}^{\infty}$ that maximizes the following discounted utility function, with $\alpha \in (0, 1)$,

$$E_x^{\pi}\left[\sum_{t=0}^{\infty} \alpha^t \log a_t\right] \tag{3.5.18}$$

subject to (3.5.16). The D.P. equation (3.5.4) becomes

$$V(x) = \max_{a\in[0,x]} \big[\log(a) + \alpha E[V(x_{t+1})|x_t = x, a_t = a]\big]. \quad (3.5.19)$$

We next solve (3.5.19) using the method of undetermined coefficients. To this end, let

$$V(x) := b_1 + b_2 \log x, \quad (3.5.20)$$

where b_1 and b_2 are unknown parameters to be determined. Hence

$$E[V(x_{t+1})|x_t = x, a_t = a] = b_1 + b_2 E[\log \xi_t] + b_2 \log(x - a),$$

and (3.5.19) becomes

$$V(x) = \max_{a\in(0,x]} \big[\log(a) + \alpha \left(b_1 + b_2 E[\log \xi_0] + b_2 \log(x - a)\right)\big], \quad (3.5.21)$$

which has a unique solution given by

$$a = \frac{x}{1 + \alpha b_2}.$$

Next, substituting this value of a in (3.5.21) and solving the system for b_1 and b_2, we obtain

$$b_1 = \frac{\alpha}{(1-\alpha)^2} E[\log \xi_0] + \frac{1}{1-\alpha} \log(1-\alpha),$$
$$b_2 = \frac{1}{1-\alpha}.$$

With these values of b_1 and b_2 in (3.5.21), it follows that the optimal control and the corresponding state trajectory are

$$f^*(x) = x(1-\alpha) \quad \text{and} \quad x_{t+1} = \alpha \xi_t x_t, \quad t = 0, 1, ...,$$

and the optimal discounted utility is

$$V(x) = \frac{\alpha}{(1-\alpha)^2} E[\log \xi_0] + \frac{1}{1-\alpha} \log(1-\alpha) + \frac{1}{1-\alpha} \log x,$$

where $E[\log \xi_0] = (1-p)\log(d)$. ◇

3.6 Policy Iteration

Let $g_0 \in \mathbb{F}$ be a selector such that the deterministic stationary policy g_0^∞ has a finite α–discounted cost

$$w_0(x) := V(g_0^\infty, x) < \infty \quad \forall\, x \in X. \tag{3.6.1}$$

By Lemma 3.23,

$$\begin{aligned} w_0(x) &= c(x, g_0) + \alpha \int_X w_0(y) Q(dy|x, g_0) \\ &\geq \min_{a \in A(x)} [c(x, a) + \alpha \int_X w_0(y) Q(dy|x, a)]. \end{aligned}$$

Hence, by Lemma 3.22, there exists $g_1 \in \mathbb{F}$ such that, for all $x \in X$,

$$w_0(x) \geq c(x, g_1) + \alpha \int_X w_0(y) Q(dy|x, g_1).$$

Iteration of the latter inequality, as in the proof of Lemma 3.24(a) gives, for all $x \in X$,

$$w_0(x) \geq w_1(x), \quad \text{with} \quad w_1(x) := V(g_1^\infty, x). \tag{3.6.2}$$

Iteration of these arguments yields the **PI algorithm**:

Step 1. (Initialization step.) Pick an arbitrary selector $g_0 \in \mathbb{F}$ as in (3.6.1).

Step 2. (Policy improvement.) Given $g_n \in \mathbb{F}$, compute the corresponding α–discounted cost $w_n(\cdot) := V(g_n^\infty, \cdot)$ as in (3.5.9). Then, as in (3.6.2), find a selector g_{n+1} such that

$$w_n(x) \geq c(x, g_{n+1}) + \alpha \int_X w_n(y) Q(dy|x, g_{n+1}) \tag{3.6.3}$$

for all $x \in X$, so that $w_n(\cdot) \geq w_{n+1}(\cdot) := V(g_{n+1}^\infty, \cdot)$.

Step 3. (Checking optimality.) Is $w_n(x) = w_{n+1}(x)$ for all $x \in X$? If "yes", then stop. If "no", then go back to step 2 replacing n with $n + 1$.

Thus, the PI algorithm gives a sequence of selectors $g_n \in \mathbb{F}$ with corresponding α–discounted cost $w_n \in L^+(X)$ forming a nonincreasing sequence that satisfies the following.

Theorem 3.27. *(a) If there is an integer n such that $w_n(x) = w_{n+1}(x)$ for all $x \in X$, then $w := w_n$ is a solution to the α–discount DP equation $w = Tw$. If, in addition, w satisfies the condition (3.5.7), that is,*

$$\lim_{t\to\infty} \alpha^t E_x^\pi w(x_t) = 0 \tag{3.6.4}$$

for all $\pi \in \Pi^0$ and $x \in X$, then $w = V^*$ and g_n^∞ is α–discount optimal.

(b) In general, as $n \to \infty$, $w_n \downarrow w$, where w is a solution to the α–discount DP equation. Moreover, if w satisfies (3.6.4), then $w = V^*$ and an α–optimal policy can be determined as in Theorem 3.21(b).

Proof. (a) If there exists n such that $w_n = w_{n+1} =: w$, the arguments leading to (3.6.2) give that w satisfies the α–discount DP equation $w = Tw$. If, moreover, (3.6.4) holds, the desired conclusion follows from Theorem 3.21(c).

(b) In general, using the arguments in the proof of Lemma 3.24(a), it can be seen that (3.6.3) gives $w_n \geq Tw_n \geq w_{n+1}$. Hence, since the sequence $\{w_n\}$ is bounded below ($w_n \geq 0$ for all n), there exists w such that $w_n \downarrow w$ and so, by Lemma 2.15(a), $w \geq Tw \geq w$, i.e., $w = Tw$. Therefore, if (3.6.4) holds, the desired conclusion follows from Theorem 3.21(c). □

3.7 Long-Run Average Cost Problems

In Sect. 2.5 we considered long-run average cost (AC) problems for discrete-time *deterministic systems*. We now briefly introduce AC problems for the stochastic system (3.1.1) and the Markov control model (MCM) $(X, A, \{A(x) : x \in X\}, Q, c)$ in (3.1.2)–(3.1.4). This material requires a working knowledge of probability and

discrete-time stochastic processes, in particular, Markov chains with a general state space. Hence, to avoid excessive technicalities, most of the results in this section are stated without proof. Some references are provided in appropriate places.

For each $T = 1, 2, ..., x \in X$ and $\pi = \{a_t\} \in \Pi$, let

$$J_T(\pi, x) := E_x^{\pi}[\sum_{t=0}^{T-1} c(x_t, a_t)] \tag{3.7.1}$$

be the T-step expected cost when using the policy π, given the initial step $x_0 = x$. The corresponding long-run expected average cost $J(\pi, x)$ is defined as

$$J(\pi, x) := \limsup_{T\to\infty} \frac{1}{T} J_T(\pi, x). \tag{3.7.2}$$

Then, as usual, the *AC value function* is

$$J^*(x) := \inf\{J(\pi, x) : \pi \in \Pi\},$$

and a policy π^* is said to be AC-optimal if $J(\pi^*, x) = J^*(x)$ for every initial state x.

For stochastic control systems, in addition to the expected average cost in (3.7.2) one can study a *pathwise* average cost. In this case, the T-stage expected cost $J_T(\pi, x), T = 1, 2, ...,$ in (3.7.1)–(3.7.2) is replaced by the *pathwise* or *random cost*

$$J_T^0(\pi, x) := \sum_{t=0}^{T-1} c(x_t, a_t). \tag{3.7.3}$$

The analysis of this stochastic control problem is more technical than the expected case and it will not be considered in these notes.

For the AC control problem in the *deterministic* case (2.5.1)–(2.5.3), we considered three approaches or techniques:

(i) the average cost optimality equation (ACOE),

(ii) the vanishing approach, and

(iii) the steady state approach.

In this section we combine the approaches (i) and (ii). (For the steady state approach (iii), which is based on the constrained optimization problem (2.5.17), there is no direct analogue in the stochastic case. In the latter case, it can be seen that the closest thing to (2.5.17) is an infinite-dimensional linear programming problem, such as, for instance, (MP_1) in page 150 of Hernández-Lerma and Lasserre (1996).)

Definition 3.28. A pair (j^*, l) consisting of a real number j^* and a real-valued function l on X is said to be:

(a) a solution to the *average cost optimality equation* (ACOE) if, for every $x \in X$ and $t = 0, 1, ...,$

$$j^* + l(x) = \inf_{a \in A(x)} [c(x, a) + E[l(x_{t+1})|x_t = x, a_t = a]], \tag{3.7.4}$$

where

$$E[l(x_{t+1})|x_t = x, a_t = a] = \int_X l(y)Q(dy|x, a) \tag{3.7.5}$$

denotes the conditional expectation of $l(x_{t+1})$ given $(x_t, a_t) = (x, a)$ for $(x, a) \in \mathbb{K}$; and

(b) a solution to the *average cost optimality inequality* (ACOI) if, for every $x \in X$ and $t = 0, 1, ...,$ the equality in (3.7.4) is replaced by the inequality $\geq$, that is (using (3.7.5)),

$$j^* + l(x) \geq \inf_{a \in A(x)} [c(x, a) + \int_X l(y)Q(dy|x, a)]. \tag{3.7.6}$$

When dealing with the system model (3.1.1) with i.i.d. random disturbances ξ_t with distribution G, the expected value (or integral) in (3.7.4)–(3.7.6) becomes

$$E[l(F(x, a, \xi)] = \int_S l(F(x, a, s))G(ds) \tag{3.7.7}$$

for all $(x, a) \in \mathbb{K}$, where ξ is a generic random variable with distribution μ.

In the remainder of this section we proceed essentially as in Sect. 2.5.3 for deterministic systems. That is, we use the Abelian theorem in Lemma 2.56 and functions as in (2.5.34) to obtain a solution to the ACOI. (Note that inequalities such as (2.5.31) and (2.5.32) are valid in the present, stochastic case because they only depend on the fact that the sequence $\{c_t\}$ in Lemma 2.56 is bounded below.)

3.7.1 The Average Cost Optimality Inequality

Throughout the remainder of this section we suppose that Assumptions 3.18 and 3.19 are satisfied. Consequently, for each $\alpha \in (0,1)$, Theorem 3.21 ensures that the α-discount value function V_α satisfies the dynamic programming equation (3.5.5), i.e.,

$$V_\alpha(x) = \min_{a \in A(x)} [c(x,a) + \alpha \int_X V_\alpha(y) Q(dy|x,a)]. \tag{3.7.8}$$

Now, pick an arbitrary state $\bar{x}$ and, as in (2.5.34), let

$$m_\alpha := V_\alpha(\bar{x})$$

and

$$h_\alpha(x) := V_\alpha(x) - m_\alpha \quad \text{and} \quad \rho(\alpha) := (1-\alpha) m_\alpha. \tag{3.7.9}$$

In addition to Assumptions 3.18 and 3.19, we suppose the following.

Assumption 3.29. There exists $\alpha_0 \in (0,1)$, positive constants N and M, and an u.s.c. function $b(\cdot) \geq 1$ on X such that, for every $\alpha \in [\alpha_0, 1)$ and $x \in X$,

(a) $\rho(\alpha) \leq M$, and

(b) $-N \leq h_\alpha(x) \leq b(x)$.

Assumption 3.29 and many of its variants are well known in the study of MDPs with the AC criterion. The main ideas probably

go back to Sennott (1986) for MDPs with a countable state space and finite action sets.

By Assumption 3.29(a), $\rho(\alpha)$ has a limit point, say $j^* \in [0, M]$, as $\alpha \uparrow 1$. Hence, there is a sequence $\alpha_n \uparrow 1$ such that

$$\rho(\alpha_n) \to j^* \tag{3.7.10}$$

as $n \to \infty$. From (3.7.10) and (3.7.9) we obtain the following fact.

Lemma 3.30. Let α_n be as in (3.7.10). Then

$$\lim_{n\to\infty} (1 - \alpha_n) V_{\alpha_n}(x) = j^* \tag{3.7.11}$$

for all $x \in X$.

Proof. Write $V_\alpha(x) = h_\alpha(x) + m_\alpha$. Multiply this equality by $(1 - \alpha)$ to obtain, from (3.7.9),

$$\begin{aligned} |(1-\alpha)V_\alpha(x) - j^*| &\le |(1-\alpha)|h_\alpha(x)| + |\rho(\alpha) - j^*| \\ &\le (1-\alpha)\max\{N, b(x)\} + |\rho(\alpha) - j^*|, \end{aligned}$$

with N and $b(\cdot)$ as in Assumption 3.29(b). Finally, replace α by α_n and then use (3.7.10) to obtain (3.7.11). □

Now, in (3.7.8) replace $V_\alpha(\cdot)$ with $h_\alpha(\cdot) + m_\alpha$. Then, by (3.7.9), we can express the α-discount DCOE (3.7.8) as

$$\rho(\alpha) + h_\alpha(x) = \min_{a \in A(x)} [c(x, a) + \alpha \int_X h_\alpha(y) Q(dy|x, a)]. \tag{3.7.12}$$

Moreover, by Theorem 3.21, for each $\alpha \in (0, 1)$ there exists $f_\alpha \in \mathbb{F}$ such that, for every $x \in X$, $f_\alpha(x) \in A(x)$ attains the minimum in (3.7.12), i.e.,

$$\rho(\alpha) + h_\alpha(x) = c(x, f_\alpha) + \alpha \int_X h_\alpha(y) Q(dy|x, f_\alpha). \tag{3.7.13}$$

Finally, replacing α in (3.7.12)–(3.7.13) by α_n in the proof of Lemma 3.30 and letting $n \to \infty$, some technical arguments yield the ACOI in the following theorem.

Theorem 3.31. *Suppose that Assumptions 3.18, 3.19 and 3.29 are satisfied and let j^* be as in (3.7.12). Then there exists a function $h^* \in L^+(X)$ and a stationary policy $f^* \in \mathbb{F}$ such that*

$$\begin{aligned} j^* + h^*(x) &\geq \min_{a \in A(x)} [c(x,a) + \int_X h^*(y) Q(dy|x,a)] \\ &= c(x, f^*) + \int_X h^*(y) Q(dy|x, f^*) \quad \forall x \in X. \end{aligned}$$

Moreover, f^* is AC-optimal and $J^*(x) = J(f^*, x) = j^*$ for all $x \in X$.

For the proof of Theorem 3.31 see Costa and Dufour (2012). Similar results appear in Feinberg et al. (2012) or Vega-Amaya (2015). These papers provide many earlier references.

3.7.2 The Average Cost Optimality Equation

To obtain the ACOE starting from the ACOI in Theorem 3.31 we impose another assumption.

Assumption 3.32. Let $b(\cdot) \geq 1$ be as in Assumption 3.29. We suppose that:

(a) $\int_X b(y) Q(dy|x, f) < \infty$ for all $x \in X$ and $f \in \mathbb{F}$.

(b) For each $f \in \mathbb{F}$ there exists a probability measure μ_f on X such that:

(b_1) $\int_X b(y) \mu_f(dy) < \infty$.

(b_2) As $t \to \infty$, $|E_x^f[v(x_t)] - \int_X v(y) \mu_f(dy)| \to 0$ for each initial state $x \in X$ and each real-valued function v on X such that $\sup_{y \in X} |v(y)|/b(y) < \infty$.

We now obtain the ACOE as follows.

Theorem 3.33. *Suppose that 3.18, 3.19, 3.29 and 3.32 are satisfied. Let j^* and f^* be as in Theorem 3.31, and μ_{f^*} as in Assumption 3.32(b). Then there exists a function h on X such that $\sup_y |h(y)|/b(y) < \infty$ and, furthermore, for every $x \in X$,*

$$j^* + h(x) = \min_{a \in A(x)} [c(x,a) + \int_X h(y) Q(dy|x,a)]$$
$$= c(x, f^*) + \int_X h(y) Q(dy|x, f^*) \qquad (3.7.14)$$

for μ_{f^*}-almost all $x \in X$, that is, there exists a set $C \subset X$ such that $\mu_{f^*}(C) = 0$ and (3.7.14) holds for x in the complement of C.

For the proof of Theorem 3.33 see Costa and Dufour (2012). Note, however, that the theorem's conclusion is on the "weak" side in the sense that the ACOE (3.7.14) is valid "almost everywhere" only. To obtain the ACOE "everywhere", that is, *for all* $x \in X$, requires strong conditions that we do not state here. On the other hand, the reader may consult Sect. 1 in Vega-Amaya (2018) for a review of results on the ACOE, as well as a presentation of the *fixed-point approach*, which we have not considered. For surveys of results up to the late 1990s see Chap. 5 and Chap. 11 in Hernández-Lerma and Lasserre (1996, 1999), respectively.

3.7.3 Examples

Example 3.34 (An LQ average cost problem). Consider the discounted LQ problem in Exercise 3.10 below, assuming that the i.i.d. random variables ξ_t have a bounded continuous density.

Thus, for each $\alpha \in (0,1)$, the α-discount value function V_α and the optimal stationary policy f_α are given, respectively, by

$$V_\alpha(x) = C_\alpha x^2 + (1-\alpha)^{-1} C_\alpha \alpha \sigma^2 \qquad (3.7.15)$$

and

$$f_\alpha(x) = -(r + \alpha\beta^2 C_\alpha)^{-1} \alpha\beta\gamma C_\alpha x, \qquad (3.7.16)$$

where $C \equiv C_\alpha$ is the unique positive solution of the quadratic equation

$$Fz^2 + Gz + qr = 0, \qquad (3.7.17)$$

with coefficients

$$F \equiv F_\alpha := \alpha\beta^2 \quad \text{and} \quad G \equiv G_\alpha := r + (r\gamma^2 + q\beta^2)\alpha. \qquad (3.7.18)$$

Therefore, taking $\bar{x} = 0$, m_α and (3.7.9) become

$$m_\alpha = V_\alpha(0) = (1-\alpha)^{-1} C_\alpha \alpha \sigma^2, \quad h_\alpha(x) = V_\alpha(x) - m_\alpha = C_\alpha x^2, \tag{3.7.19}$$

and

$$\rho(\alpha) = (1-\alpha) m_\alpha = C_\alpha \alpha \sigma^2. \tag{3.7.20}$$

Moreover, a direct calculation shows that for all α sufficiently close to 1, say $\alpha_0 < \alpha < 1$ for some $\alpha_0 \in (0,1)$, the positive number C_α is bounded above by

$$L := (r + r\gamma^2 + q\beta^2)/\beta^2 \alpha_0.$$

This yields the Definition 3.28(a), i.e., $\rho(\alpha) \le M$ with $M := L\sigma^2$. Definition 3.28(b) is similarly verified, with $N = 0$ and $b(x) := Lx^2$.

Finally, to obtain the ACOI in Theorem 3.31 take $\alpha_n \in (0,1)$ such that $\alpha_n \uparrow 1$ as $n \to \infty$. Then (as in (3.7.10) and (3.7.11)) we can see from (3.7.20) that

$$\rho(\alpha_n) \to j^* := k\sigma^2 \tag{3.7.21}$$

where k is the unique positive solution of (3.7.17)–(3.7.18) with $\alpha = 1$. Likewise, from (3.7.19), as $n \to \infty$ we obtain

$$h_{\alpha_n}(x) \to h^*(x) := kx^2 \tag{3.7.22}$$

for all $x \in X$. We can now verify that the pair $(j^*, k^*(\cdot))$ is a solution to the ACOI (3.7.6). In fact, if in (3.7.16) we replace α by $\alpha_n \uparrow 1$, then in the limit we obtain

$$f^*(x) = -(r + k\beta^2)^{-1} \beta\gamma k x \quad \forall x \in X$$

and $(j^*, h^*(\cdot), f^*(\cdot))$ is a *canonical triplet* that satisfies the AC optimality equation (ACOE) (3.7.14) for all $x \in X$. (Note that we didn't have to verify Assumption 3.32. This is because of the particular characteristics of LQ problems, but it is not the general case.) ◇

Example 3.35 (The Brock-Mirman model with technological shocks). In the infinite-horizon Brock and Mirman model studied

in Examples 2.10, 2.31 and 2.51, the system evolves according to

$$x_{t+1} = c_t x_t^{\theta} - a_t \quad \text{for} \quad t = 0, 1, 2... \tag{3.7.23}$$

with initial condition $x_0 = x > 0$ and $\theta \in (0, 1)$. Now, we suppose that the technological parameter $\{c_t\}$ behaves as a Markov process such that

$$\log(c_{t+1}) = \rho \log(c_t) + \xi_t \quad \text{for} \quad t = 0, 1, 2... \tag{3.7.24}$$

with $\rho > 0$, initial condition $c_0 > 0$, and where $\{\xi_t\}$ is a collection of independent and identically distributed normal random variables with mean 0 and variance $\sigma > 0$. Under this hypothesis, (3.7.24) yields $c_{t+1} = c_t^{\rho} e^{\xi_t}$, where e^{ξ_t} has a log-normal distribution. This stochastic model was introduced by Brock and Mirman (1972, 1973).

The state space and control spaces are, respectively, $X \times C = [0, \infty) \times [0, \infty)$, $A = (0, \infty)$, and $A(x, c) = (0, cx^{\theta}]$ for $x > 0$.

Consider the objective function to be optimized as the long–run *average reward* (AR)

$$J(\pi, (x, c)) = \liminf_{T \to \infty} \frac{1}{T} E_{x,c}^{\pi} \left[\sum_{t=0}^{T-1} \log(a_t) \right]. \tag{3.7.25}$$

This AR problem is, of course, analogous to the AC problem in (3.7.2), so the *AR value function* is

$$J^*(x, c) := \sup\{J(\pi, (x, c)) : \pi \in \Pi\}. \tag{3.7.26}$$

To find a *canonical triplet* (j^*, l, f^*) that satisfies the average reward optimality equation (as in Definition 3.28(a) or Theorem 3.33), i.e.,

$$j^* + l(x, c) = \max_{a \in A(x,c)} [\log(a) + E[l(x_{t+1}, c_{t+1}) | (x_t, c_t) = (x, c), a_t = a]] \tag{3.7.27}$$

for all $(x, c) \in X \times C$, we consider a function $l(x, c)$ of the form

$$l(x, c) := b_1 \log(x) + b_2 \log(c),$$

where b_1 and b_2 are unknown parameters to be determined.

In this case

$$E[l(x_{t+1}, c_{t+1})|(x_t, c_t) = (x, c), a_t = a] = b_1 \log(cx^\theta - a) + b_2\rho \log(c).$$

Hence, (3.7.27) reaches the maximum at $a = \frac{cx^\theta}{1+b_1}$, and so we can express (3.7.27) as

$$j^* + b_1 \log(x) + b_2 \log(c) = b_1 \log(b_1) - (1 + b_1) \log(1 + b_1) + \theta(1 + b_1) \log(x) + [1 + b_1 + b_2\rho] \log(c).$$

This last equation is satisfied if

$$\begin{aligned} j^* &= b_1 \log(b_1) - (1 + b_1) \log(1 + b_1), \\ b_1 &= \theta(1 + b_1), \\ b_2 &= 1 + b_1 + b_2\rho, \end{aligned}$$

which implies that

$$b_1 = \frac{\theta}{1-\theta}, \quad b_2 = \frac{1}{(1-\theta)(1-\rho)}.$$

Therefore the canonical triplet (j^*, l, f^*) for the stochastic Brock-Mirman model (3.7.23)–(3.7.25) is given by

$$\begin{aligned} j^* &= \frac{\theta}{1-\theta} \log\left(\frac{\theta}{1-\theta}\right) - \frac{1}{1-\theta} \log\left(\frac{1}{1-\theta}\right), \qquad (3.7.28) \\ l(x, c) &= \frac{\theta \log(x)}{1-\theta} + \frac{\log(c)}{(1-\theta)(1-\rho)}, \\ f^*(x, c) &= (1-\theta)cx^\theta. \end{aligned}$$

◇

Example 3.36 (The stochastic Brock-Mirman model). We wish to solve again the stochastic average reward Brock-Mirman model in Example 3.35 (see (3.7.23)–(3.7.25)), but now using the *vanishing discount approach*. To this end, consider the α-optimal discounted utility $V_\alpha(x, c)$ in the Exercise 3.14. Then

$$(1-\alpha)V_\alpha(x, c) = (1-\alpha)B + \frac{(1-\alpha)\alpha\theta}{1-\theta} \log(x) + \frac{1-\alpha}{1-\alpha\theta} \log(c),$$

with B as in Exercise 3.14. Therefore, letting $\alpha \uparrow 1$ we obtain the optimal average reward

$$J^*(x,c) = \lim_{\alpha\uparrow 1}(1-\alpha)V_\alpha(x,c) = j^*$$

for all (x,c), with j^* as in (3.7.28). ◇

Exercises

3.1. (a) Prove Theorem 3.13.
Hint. Use the proof of (3.3.3) or the Remark 3.12(c). Moreover, since c and u are bounded below, without loss of generality you can assume that they are *nonnegative*. Then use Theorem B.7.
(b) Prove Theorem 3.14.

Hint. Recall Remark 3.12(c) and Lemma 2.16(a).

3.2. Consider the MCM (3.1.2), the set $\mathbb{K}$ in (3.1.3), and a real–valued function v on $\mathbb{K}$. (Recall the notation in Remark 3.7: $v(x,f) := v(x,f(x))$.) Show that, for each $x \in X$,

$$\sup_{a\in A(x)} v(x,a) = \sup_{f\in\mathbb{F}} v(x,f).$$

Hint. Fix an arbitrary $x \in X$. For any $f \in \mathbb{F}$, $v(x,f) \le \sup_{a\in A(x)} v(x,a)$. Hence $\sup_{f\in\mathbb{F}} v(x,f) \le \sup_{a\in A(x)} v(x,a)$. To obtain the reverse inequality, fix an arbitrary $a \in A(x)$, and let $h \in \mathbb{F}$ be any selector. Define the mapping $g : X \to A$ as

$$g(y) := \begin{cases} a & \text{if} \quad y = x, \\ h(y) & \text{if} \quad y \neq x. \end{cases}$$

Then g is measurable and so it belongs to $\mathbb{F}$. Therefore,

$$v(x,a) = v(x,g) \le \sup_{f\in\mathbb{F}} v(x,f).$$

Thus, $\sup_a v(x,a) \le \sup_f v(x,f)$.

3.3. Use backward induction to verify (3.4.6)–(3.4.7).

3.4. In Example 3.17, prove (3.4.10)–(3.4.11) and (3.4.12)–(3.4.13).

3.5. In the MCM (3.2.1), suppose that Q is weakly continuous (see Theorem 3.13(b)), and let u be a real–valued l.s.c. function on X that is bounded below. Show that, then, the mapping $(x, a) \mapsto \int_X u(y)Q(dy|x, a)$ is l.s.c. on $\mathbb{K}$.

3.6. Consider a MCM in which the cost–per–stage c is *bounded*—see the paragraph after Assumption 6.2. Show that, then, V^* is the *unique* bounded solution to the α–discount DP equation. (The latter equation, however, may have several *unbounded* solutions—see, for instance, the following exercise.)

3.7. The following MCM by Sennott (1986) shows that the α–discount DP equation may have several *unbounded* solutions. Let $X = \{1, 2, \dots\}$, $A = \{1\}$, $c(x, a) \equiv 0$ for all x, a, and transition law

$$Q(\{1\}|x, 1) = 2x/3(2x - 1) \quad \text{and} \quad Q(\{2x\}|x, 1) = (4x - 3)/3(2x - 1)$$

for all $x \in X$. Let $\alpha = 3/4$. Show that $V^*(\cdot) \equiv 0$, but the identity function $u(x) = x$ for all $x \in X$ is also a solution to the DP equation (3.5.4).

3.8. (Blackwell 1965) Consider the MCM with components $X = \{0\}$, $A = \{1, 2, ...\}$, $c(0, a) = 1/a$ and, of course, $Q(0|0, a) = 1$ for all $a \in A$. Note that the optimal cost function is $V^*(0) = 0$.

(a) Is V^* a solution of the DP equation (3.5.4)? Explain.

(b) Does it exist an optimal policy? Explain.

3.9. Consider a MCM with $X = \{1\}$, $A(1) = \{1, 2, \dots\}$, and $c(1, a) = (1 + a)/a$ for all a. Show that $V^*(1) = 1/(1 - \alpha)$, but there is no optimal control policy; in other words, there is no π such that $V(\pi, \cdot) = V^*(\cdot)$. However, for any $\epsilon > 0$, there exists an ϵ–optimal policy—that is, for each $\epsilon > 0$, there exists a policy $\pi \equiv \pi_\epsilon$ such that $V(\pi, x) \leq V^*(x) + \epsilon$ for all $x \in X$.

3.10. The infinite–horizon LQ discounted problem. Consider the discounted LQ problem in Example 3.16; that is, the system (3.4.1)–(3.4.3) with state and action spaces $X = A = \mathbb{R}$, quadratic cost $c(x, a) = qx^2 + ra^2$, and linear system equation

$$x_{t+1} = \gamma x_t + \beta a_t + \xi_t \quad \forall\, t = 0, 1, \ldots,$$

where the random perturbations ξ_t are i.i.d. random variables with mean zero and finite variance $\sigma^2 = E(\xi^2)$. The coefficients q and r are both positive, and $\gamma \cdot \beta \neq 0$. Show that the α–discount optimal cost is, for every $x \in X$,

$$V^*(x) = Cx^2 + (1-\alpha)^{-1} C\alpha\sigma^2,$$

and the α–optimal stationary policy is determined by

$$f^*(x) = -(r + \alpha\beta^2 C)^{-1} \alpha\beta\gamma C x,$$

where $C = z_1$ is the unique positive solution of the quadratic equation

$$Fz^2 + Gz + qr = 0,$$

with $F = \alpha\beta^2$ and $G = r + (r\gamma^2 + q\beta^2)\alpha$. (Concerning the LQ discounted problem, note that some of the hypotheses in Theorem 3.21 were already verified in the paragraph after Assumption 3.18.)

3.11. Let π be a policy such that its cost function $V(\pi, \cdot) \equiv u(\cdot)$ is "almost" a solution to the α–discount D.P. equation (3.5.4) in the sense that, for some $\epsilon > 0$,

$$u(x) \leq c(x, a) + \alpha \int_X u(y) Q(dy|x, a) + \epsilon(1-\alpha)$$

for all $x \in X$ and $a \in A(x)$. In addition, suppose that u satisfies the condition (3.5.11). Show that π is an ϵ–*optimal* policy, that is, $u(x) \leq V^*(x) + \epsilon$ for all $x \in X$.

3.12. Consider the MCM $(X, A, \{A(x)|x \in X\}, Q, c)$ in (3.1.2), and let $B(X)$ be as in Exercise 2.7. Suppose that the cost function c is bounded, say $0 \leq c \leq M$, and for each $u \in B(X)$, let Tu be as in (3.5.3). Show that the operator T satisfies (a) and (b) in Exercise 2.7; that is,

(a) T is a contraction on $B(X)$, and

(b) the value function V^* in (3.5.2) is the unique fixed point of T in $B(X)$. Moreover,

(c) the α–VI functions $v_n := T^n 0$ in (3.5.13) converge to V^*.

3.13. (a) Consider the time-varying additive-noise system

$$x_{t+1} = F_1(t)x_t + F_2(t, a_t) + \xi_t, \quad t = 0, 1, ..., T-1,$$

with state and action spaces $X \subset \mathbb{R}^n$ and $A \subset \mathbb{R}^m$, respectively. Note that the system is *linear in the state*. The stage cost is of the same form, so the associated performance criterion is

$$J(\pi, x) = E_x^\pi \left[\sum_{t=0}^{T-1} [c_1(t)x_t + c_2(t, a_t)] \right].$$

Assume that A is compact, and the functions $F_2(t, a)$ and $c_2(t, a)$ are continuous in $a \in A$. Moreover, the random variables ξ_t $(t = 0, ..., T-1)$ have finite means, and they are independent and also independent of the initial state x_0. Prove that the OCP has an optimal control that is independent of the state variable.

(b) Show that the conclusion in part (a) also holds in the time-homogeneous infinite-horizon discounted OCP with the system model (3.1.1) and discounted cost (3.5.1) with

$$F(x, a, s) := F_1 x + F_2(a) + s, \quad c(x, a) := c_1 x + c_2(a),$$

respectively, under the Assumption 3.18. (Part (a) is due to Midler (1969). See Exercise 2.16 for a *deterministic* version of this exercise.)

3.14. An stochastic Brock-Mirman (1972) model. Consider a Brock-Mirman economic growth model as in Example 2.10 and Example 3.35

$$x_{t+1} = c_t x_t^\theta - a_t \quad t = 0, 1, ... \tag{3.7.29}$$

except that the "technological parameter" c is now a Markov process that evolves as

$$\log(c_{t+1}) = \rho \log(c_t) + \xi_t, \quad t = 0, 1, ... \tag{3.7.30}$$

where the ξ_t are i.i.d. Gaussian (or normal) random variables with zero mean and standard deviation $\sigma > 0$. In (3.7.30) we assume that c_0 is independent of the sequence $\{\xi_t\}$. As in Example 2.10,

x_t and a_t denote capital and consumption both with values in $[0, \infty)$, but now the *state variable* is the pair (x_t, c_t) with values in $X \times C$ with $X = C = [0, \infty)$. Consider the discounted utility

$$E^{\pi}_{(x,c)} \left[\sum_{t=0}^{\infty} \alpha^t \log(a_t) \right]$$

with initial condition $(x_0, c_0) = (x, c)$. Show that the optimal discounted utility $V_\alpha(x, c)$ is

$$V_\alpha(x, c) = B + \frac{\alpha\theta}{1-\theta} \log(x) + \frac{1}{1-\alpha\theta} \log(c)$$

with

$$B = \frac{1}{1-\alpha} \left[\frac{\alpha\theta}{1-\theta} \log\left(\frac{\alpha\theta}{1-\alpha\theta} \right) - \frac{1}{1-\alpha\theta} \log\left(\frac{1}{1-\alpha\theta} \right) \right].$$

Hint. Solve the corresponding α-DP equation by means of the "guess and verify" approach (or method of undetermined coefficients) with a logarithmic function of the form $v(x, c) = b_1 + b_2 \log(x) + b_3 \log(c)$.

Chapter 4

Continuous–Time Deterministic Systems

We now consider a deterministic continuous–time optimal control problem (OCP) in which the state process $x(\cdot)$ evolves in the *state space* $X := \mathbb{R}^n$ according to an ordinary differential equation

$$\dot{x}(t) = F(t, x(t), a(t)) \quad \text{for} \quad t \in [0, T], \tag{4.0.1}$$

for a given initial state $x(0) = x_0 \in X$, and a given system function $F : [0, T] \times X \times A \to X$, where A is a separable metric space that stands for the *action* (or *control*) *set*. For the time being, in (4.0.1) we consider so-called *open–loop* controls $a(\cdot)$, that is, controls in the family $\mathcal{A}[0, T]$ of piecewise continuous functions $a(\cdot) : [0, T] \to A$.

Remark 4.1. We assume that, for each control function $a(\cdot)$, (4.0.1) has a unique solution $x(\cdot)$. To this end, it suffices to assume that, for instance, F is a continuous function and has continuous first partial derivatives with respect to the components of $x \in X$. (For a more precise statement, see Assumption 4.2 below.) $\Diamond$

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*, Texts in Applied Mathematics 76,
https://doi.org/10.1007/978-3-031-21139-3_4

4.1 The HJB Equation and Related Topics

4.1.1 Finite–Horizon Problems: The HJB Equation

The performance index or cost functional associated to (4.0.1) is

$$J(a(\cdot)) := \int_0^T c(t, x(t), a(t))dt + C(x(T)), \tag{4.1.1}$$

where c and C are given nonnegative functions, called the *running* or *instantaneous* cost and the *terminal* cost, respectively.

Assuming that (4.0.1) has a unique solution and that (4.1.1) is well defined, the OCP we are concerned with is the following:

OCP Minimize (4.1.1) over $\mathcal{A}[0, T]$.

As in previous chapters, when using dynamic programming associated to an OCP we consider, for each time $s \in [0, T)$, an OCP from s to the terminal time T. That is, for each $(s, y) \in [0, T) \times X$, we consider the dynamic system (as in (4.0.1))

$$\dot{x}(t) = F(t, x(t), a(t)) \quad \text{for} \quad t \in [s, T], \quad x(s) = y. \tag{4.1.2}$$

The controls in (4.1.2) are restricted to the interval $[s, T]$, so $a(\cdot)$ is in $\mathcal{A}[s, T]$, the family of piecewise continuous functions $a(\cdot)$ from $[s, T]$ to A, and the performance index in (4.1.1) is replaced with

$$J(s, y; a(\cdot)) := \int_s^T c(t, x(t), a(t))dt + C(x(T)). \tag{4.1.3}$$

This new OCP is called OCP_{sy}. If $(s, y) = (0, x_0)$, then the problem OCP_{sy} reduces, of course, to the original OCP.

If $a^*(\cdot)$ minimizes (4.1.3) over all $a(\cdot) \in \mathcal{A}[s, T]$ and $x^*(\cdot)$ denotes the corresponding solution to (4.1.2), then we say that a^* is an *optimal control* and $(x^*(\cdot), a^*(\cdot))$ is an *optimal pair* of problem OCP_{sy}.

To ensure that OCP_{sy} is well defined for each initial condition (s, y) in $[0, T] \times X$ we impose the following hypotheses.

Assumption 4.2. Let $\varphi(t,x,a)$ be any of the functions $F(t,x,a)$, $c(t,x,a), C(t)$ in (4.0.1)–(4.1.1) (or (4.1.2)–(4.1.3)). The function φ is uniformly continuous and, moreover, there is a constant L such that

(a) $|\varphi(t,x,a) - \varphi(t,\hat{x},a)| \le L|x-\hat{x}| \quad \forall\, t \in [0,T], x, \hat{x} \in X, a \in A.$

(b) $|\varphi(t,0,a)| \le L \quad \forall\, (t,a) \in [0,T] \times A.$

Under this assumption, (4.1.2) has a unique solution $x(\cdot) \equiv x(\cdot; s, y, a(\cdot))$ for every initial condition $(s,y) \in [0,T] \times X$ and every control $a(\cdot) \in \mathcal{A}[s,T]$. (See Exercise 4.4.) Moreover, (4.1.3) is well defined.

Consider now the *value function* corresponding to OCP_{sy}:

$$V(s,y) := \inf_{a(\cdot)\in\mathcal{A}[s,T]} J(s,y;a(\cdot)) \quad \forall\, (s,y) \in [0,T) \times X, \tag{4.1.4}$$

$$V(T,y) = C(y) \quad \forall\ y \in X.$$

The following theorem states that V satisfies *Bellman's principle of optimality* (4.1.5). (See also Corollary 4.4.)

Theorem 4.3. *Suppose that Assumption 4.2 holds. Then, for any $(s,y) \in [0,T) \times X$ and $\hat{s} \in [s,T]$,*

$$V(s,y) = \inf_{a(\cdot)\in\mathcal{A}[s,T]} \left[\int_s^{\hat{s}} c(t,x(t),a(t))dt + V(\hat{s}, x(\hat{s})) \right]. \tag{4.1.5}$$

Proof. Denote by $\bar{V}(s,y)$ the right–hand side of (4.1.5). It is easy to see that

$$V(s,y) \le \bar{V}(s,y). \tag{4.1.6}$$

Indeed, by definition (4.1.4), for any control $a(\cdot) \in \mathcal{A}[s,T]$ we have

$$\begin{aligned} V(s,y) &\le J(s,y;a(\cdot)) \\ &= \int_s^{\hat{s}} c(t,x(t),a(t))dt + J(\hat{s}, x(\hat{s}); a(\cdot)) \end{aligned}$$

and then, taking the infimum over $a(\cdot) \in \mathcal{A}[s,T]$, (4.1.6) follows. To obtain the reverse inequality, fix an arbitrary $\epsilon > 0$, and choose

a control $a_\epsilon(\cdot) \in \mathcal{A}[s,T]$ such that

$$\begin{aligned} V(s,y) + \epsilon &\geq J(s,y;a_\epsilon(\cdot)) \\ &\geq \int_s^{\hat{s}} c(t,x(t),a(t))dt + V(\hat{s},x_\epsilon(\hat{s})) \\ &\geq \bar{V}(s,y), \end{aligned}$$

where $x_\epsilon(\cdot) = x(\cdot;s,y,a_\epsilon)$. The latter inequality and (4.1.6) give (4.1.5). □

Assuming the existence of optimal controls, Theorem 4.3 yields Bellman's principle of optimality in the usual form of Lemma 2.2; namely, if the control $a^*(\cdot)$ is optimal on $[s,T]$, with initial condition (s,y), then restricted to $[\hat{s},T]$, for any $\hat{s} \in (s,T)$, $a^*(\cdot)$ is also optimal with initial condition $(\hat{s},x^*(\hat{s}))$. A more precise statement in terms of the value function V in (4.1.4) is the following.

Corollary 4.4. Suppose that $(x^*(\cdot),a^*(\cdot))$ is an optimal pair of problem OCP_{sy}. Then, for any $\hat{s} \in (s,T)$,

$$V(\hat{s},x^*(\hat{s})) = J(\hat{s},x^*(\hat{s});a^*(\cdot)). \tag{4.1.7}$$

Conversely, if $(x^*(\cdot),a^*(\cdot))$ satisfies (4.1.7) for all $0 \leq s < \hat{s} < T$, then (4.1.5) holds.

Proof. By the optimality of $(x^*(\cdot),a^*(\cdot))$,

$$\begin{aligned} V(s,y) &= J(s,y;a^*(\cdot)) \\ &= \int_s^{\hat{s}} c(t,x^*(t),a^*(t))dt + J(\hat{s},x^*(\hat{s});a^*(\cdot)) \\ &\geq \int_s^{\hat{s}} c(t,x^*(t),a^*(t))dt + V(\hat{s},x^*(\hat{s})) \\ &\geq V(s,y) \qquad\qquad [\text{by (4.1.5)}]. \end{aligned}$$

That is, the latter inequality is in fact an equality and it yields (4.1.7). The converse is obvious. □

Corollary 4.4 gives a necessary condition for $a^*(\cdot)$ to be optimal. For practical purposes, however, this is not very helpful because the corollary does not say how to find neither V nor $a^*(\cdot)$. Never-

theless, proceeding as in Sect. 2.1, we can try to use the principle of optimality (Lemma 2.2) to find the dynamic programming (or Bellman) equation (in (2.1.6)–(2.1.7)). In our present case, the latter procedure shows that, under suitable hypotheses, (4.1.5) yields the DP equation (4.1.8)–(4.1.9) below, which in the continuous–time case is called the *Hamilton–Jacobi–Bellman* (HJB) equation associated to OCP.

Remark 4.5. Given a real-valued function $(t, x) \mapsto v(t, x)$ on $(0, T) \times \mathbb{R}^n$ we denote by v_t the partial derivative of v with respect to t and by v_x the gradient of v, that is, the (row) vector $(v_{x_1}, ..., v_{x_n})$ of partial derivatives. Further, $C^1(Y)$ denotes the family of real-valued continuously differentiable functions on the space Y. ◊

Theorem 4.6. *In addition to Assumption 4.2, suppose that the value function $V : [0, T] \times X \to \mathbb{R}$ in (4.1.4) is continuously differentiable. Then:*

(a) V satisfies the first-order partial differential equation

$$V_t + \inf_{a \in A}[c(t, x, a) + V_x \cdot F(t, x, a)] = 0 \quad \forall\ (t, x) \in [0, T) \times X, \tag{4.1.8}$$

$$V(T, x) = C(x) \ \ \forall \ \ x \in X. \tag{4.1.9}$$

(b) Furthermore, if there exists a control function $a^* : [0, T) \times X \to A$ such that $a^*(t, x) \in A$ attains the minimum in (4.1.8) for every $(t, x) \in [0, T) \times X$, then $a^*(t) := a^*(t, x^*(t))$ is an optimal control.

Before proving Theorem 4.6 we mention other equivalent forms of expressing the HJB equation (4.1.8).

Remark 4.7. (a) The function within brackets in (4.1.8), that is,

$$H(t, x, a, \lambda) := c(t, x, a) + \lambda \cdot F(t, x, a), \quad \text{with} \quad \lambda = V_x \tag{4.1.10}$$

is called the *Hamiltonian* associated to the OCP (4.0.1)–(4.1.1). In terms of the Hamiltonian, the HJB equation (4.1.8) becomes

$$V_t + \inf_{a\in A} H(t,x,a,V_x) = 0 \tag{4.1.11}$$

for all $(t,x) \in [0,T) \times X$, with the terminal condition (4.1.9). We will use this form of the HJB equation to obtain necessary conditions for optimality in Sect. 4.1.2.

(b) For each $a \in A$ and each real–valued continuously differentiable function v on $[0,T] \times X$, let

$$L^a v(t,x) := v_t(t,x) + v_x(t,x) \cdot F(t,x,a). \tag{4.1.12}$$

Interpreting (4.0.1)–(4.1.1) as a *Markov control problem* , the operator L^a in (4.1.12) denotes the *infinitesimal generator* of the "controlled Markov process" $x(\cdot)$ in (4.0.1). (We will come back to this point in Chap. 5.) Moreover, we can rewrite the HJB equation (4.1.8) as

$$\inf_{a\in A}[c(t,x,a) + L^a V(t,x)] = 0. \tag{4.1.13}$$

(c) For future reference, note that (4.1.12) is the derivative of $v(t,x(t))$ with respect to t, given that $(t,x(t),a(t)) = (t,x,a)$; indeed, by the chain rule and (4.1.2),

$$\begin{aligned} L^a v(t,x) &= \frac{d}{dt} v(t,x(t))\Big|_{(t,x,a)} \\ &= [v_t(t,x(t)) + v_x(t,x(t)) \cdot \dot{x}(t)]\,|_{(t,x,a)} \\ &= v_t(t,x) + v_x(t,x) \cdot F(t,x,a). \end{aligned}$$

(d) In the time-homogeneous (or time-invariant) case in which $c(t,x,a)$ and $F(t,x,a)$ in (4.1.1)–(4.1.2) are replaced by $c(x,a)$ and $F(x,a)$, respectively, the operator L^a in (4.1.2) becomes

$$L^a v(x) := v_x(x) \cdot F(x,a) = \frac{d}{dt} v(x(t))\Big|_{(x,a)}$$

for $v \in C^1(X)$.

Summarizing, in addition to the "explicit form" (4.1.8) of the HJB equation, we also have (4.1.11) and (4.1.13). ♢

In part (a) of the following proof we use the fact that, for $\varphi = F$ or c,

$$\lim_{t\downarrow s} \sup_{y\in X, a\in A} |\varphi(t,y,a) - \varphi(s,y,a)| = 0. \tag{4.1.14}$$

This follows from the uniform continuity of F and c in Assumption 4.2.

Proof of Theorem 4.6.

(a) Fix an arbitrary $a \in A$, and let $x(\cdot)$ be the solution of (4.0.1) when using the control $a(\cdot) \equiv a$. Then, from (4.1.5),

$$V(\hat{s}, x(\hat{s})) - V(s,y) + \int_s^{\hat{s}} c(t, x(t), a)dt \geq 0.$$

Multiplying both sides of this inequality by $(\hat{s}-s)^{-1}$ and then letting $\hat{s} \downarrow s$ we obtain

$$c(s,y,a) + V_t(s,y) + V_x(s,y) \cdot F(s,y,a) \geq 0.$$

Therefore, since $a \in A$ was arbitrary, it follows that

$$V_t(s,y) + \inf_{a\in A}[c(s,y,a) + V_x(s,y) \cdot F(s,y,a)] \geq 0. \tag{4.1.15}$$

To obtain the reverse inequality, for any $\epsilon > 0$ and $\hat{s} \in [s,T]$ such that $\hat{s} - s > 0$ is small enough, there exists a control $a(\cdot) \equiv a_{\epsilon,\hat{s}}(\cdot)$ in $\mathcal{A}[s,T]$ such that

$$\int_s^{\hat{s}} c(t, x(t), a(t))dt \leq V(s,y) - V(\hat{s}, x(\hat{s})) + \epsilon(\hat{s}-s).$$

Therefore, since V is continuously differentiable, and from Remark 4.7(c),

$$\begin{aligned}
\epsilon &\geq (\hat{s}-s)^{-1}\left[V(\hat{s}, x(\hat{s})) - V(s,y) + \int_s^{\hat{s}} c(t, x(t), a(t))dt\right] \\
&= (\hat{s}-s)^{-1} \int_s^{\hat{s}} \left[L^{a(t)}V(t, x(t)) + c(t, x(t), a(t))\right] dt \\
&\geq (\hat{s}-s)^{-1} \int_s^{\hat{s}} \inf_{a\in A} \left[c(t, x(t), a) + L^a V(t, x(t))\right] dt \\
&\to \inf_{a\in A} [c(s,y,a) + L^a V(s,y)] \quad \text{as} \quad \hat{s} \downarrow s,
\end{aligned}$$

by (4.1.14). The last inequality and (4.1.13) conclude the proof of part (a).

(b) We will use the HJB equation in the form (4.1.13). Hence, for any $a(\cdot) \in \mathcal{A}[0,T]$ we have

$$c(t,x(t),a(t)) + L^{a(t)}V(t,x(t)) \geq 0 \tag{4.1.16}$$

and so integration on $[s,T]$ yields (by the Remark 4.7(c))

$$\int_s^T c(t,x(t),a(t))dt + V(T,x) - V(s,x) \geq 0. \tag{4.1.17}$$

Therefore, by (4.1.9) and (4.1.3),

$$V(s,x) \leq J(s,x;a(\cdot)). \tag{4.1.18}$$

Finally, let $a^*(t,x)$ be as in (b) and let $x^*(\cdot)$ be the solution of (4.0.1) when using the control $a^*(t) := a^*(t,x^*(t))$. Then, replacing the pair $(x(\cdot),a(\cdot))$ by $(x^*(\cdot),a^*(\cdot))$, we obtain *equalities* in (4.1.16)–(4.1.18); in particular, (4.1.18) becomes

$$V(s,x) = J(s,x;a^*(\cdot)) \quad \forall\ (s,x) \in [0,T] \times X.$$

Thus, $a^*(\cdot)$ is an optimal control. □

Theorem 4.6 is called a *verification theorem*. It is so–named because it gives the following "verification technique" to solve the OCP (4.0.1)–(4.1.1):

Step 1. Solve the HJB equation (4.1.8)–(4.1.9) to find the value function $V(t,x)$.
Step 2. Find the minimizer $a^*(t,x)$ in (4.1.8).
Step 3. Solve (4.0.1) with $a = a^*$ to obtain the optimal pair

$$(x^*(\cdot),a^*(\cdot)).$$

Clearly, step 1 is the most demanding from a technical viewpoint, because it hinges on the fact that (i) V is continuously differentiable, that is, $V \in C^1([0,T] \times X)$, and (ii) the HJB equation admits a unique *classical* solution. Unfortunately, neither (i) nor (ii) are true in general. For instance, Example 2.4 in Yong and

Zhou (1999), pp. 163–164, is an innocent–looking OCP in which V does not satisfy (i). See also Exercise 4.4.

4.1.2 A Minimum Principle from the HJB Equation

In Sects. 2.2.2 and 2.3.2, we respectively considered finite- and infinite-horizon versions of the Minimum Principle in discrete time. We now discuss an analogous principle in continuous time.

Consider the OCP with dynamics (4.0.1) and cost functional (4.1.1). Suppose that Assumption 4.2 holds and the value function V is of class C^2. Let $(x^*(\cdot), a^*(\cdot))$ be an optimal pair. Put

$$\lambda(s) := V_x(s, x^*(s)) \qquad 0 \le s \le T.$$

Recall the Hamiltonian

$$H(s, x, a, \lambda) = c(s, x, a) + \lambda \cdot F(s, x, a)$$

introduced in Remark 4.7(a).

The minimum condition. We start by differentiating, with respect to s, both sides of the equality

$$V(s, x^*(s)) = \int_s^T c(t, x^*(t), a^*(t))dt + C(x^*(T))$$

to obtain

$$V_t(s, x^*(s)) + V_x(s, x^*(s)) \cdot \dot{x}^*(s) = -c(s, x^*(s), a^*(s))$$

which is equivalent to

$$\begin{aligned} -V_t(s, x^*(s)) &= c(s, x^*(s), a^*(s)) + V_x(s, x^*(s)) \cdot F(s, x^*(s), a^*(s)) \\ &= c(s, x^*(s), a^*(s)) + \lambda(s) \cdot F(s, x^*(s), a^*(s)) \\ &= H(x, x^*(s), a^*(s), \lambda(s)) \end{aligned} \tag{4.1.19}$$

for each s in $[0, T)$. On the other hand, by Theorem 4.6, V satisfies the HJB equation (4.1.8)–(4.1.9). In particular,

$$-V_t(s,x^*(s)) = \inf_{a\in A}[c(s,x^*(s),a) + \lambda(s)\cdot F(s,x^*(s),a)], \quad 0\le s<T.$$

From the latter equality and (4.1.19), we conclude that

$$H(s,x^*(s),a^*(s),\lambda(s)) = \min_{a\in A} H(s,x^*(s),a,\lambda(s)), \qquad 0\le s<T. \tag{4.1.20}$$

The adjoint equation. Let us now assume that, for each fixed s,

$$x \mapsto \inf_{a\in A} H(s,x,a,V_x(s,x)) \tag{4.1.21}$$

is differentiable and its derivative at $x^*(s)$ equals

$$c_x(s,x^*(s),a^*(s)) + \lambda(s)F_x(s,x^*(s),a^*(s)) + V_{xx}(s,x^*(s))\cdot F(s,x^*(s),a^*(s)). \tag{4.1.22}$$

Remark 4.8. Results about sufficient conditions for differentiability of mappings like (4.1.21) are usually called *Envelope Theorems*. An example of such sufficient conditions—as well as a formula that yields (4.1.22)—is given in Exercise 4.8.

After taking the partial derivative with respect to x in both sides of (4.1.8) and evaluating at $x = x^*(s)$, we have

$$-V_{tx}(s,x^*(s)) = c_x(s,x^*(s),a^*(s)) + \lambda(s)F_x(s,x^*(s),a^*(s)) + V_{xx}(s,x^*(s))\cdot \dot{x}^*(s)$$

for each s in $[0,T)$. Then

$$c_x(s,x^*(s),a^*(s)) + \lambda(s)F_x(s,x^*(s),a^*(s)) = -V_{tx}(s,x^*) - V_{xx}(s,x^*(s))\cdot \dot{x}^*(s) = -\frac{d}{dt}V_x(s,x^*(s)),$$

that is,

$$-\dot{\lambda}(s) = c_x(s,x^*(s),a^*(s)) + \lambda(s)F_x(s,x^*(s),a^*(s)), \quad 0\le s<T. \tag{4.1.23}$$

Besides, the terminal condition

$$\lambda(T) = C_x(x^*(T)) \tag{4.1.24}$$

follows from the definition of λ and (4.1.9).

Finally, the so-called *Minimum Principle* can be stated as follows.

The Minimum Principle. Under the assumptions made above, if $(x^*(\cdot), a^*(\cdot))$ minimizes the cost functional (4.1.1) subject to the dynamics (4.0.1), then there exists a function $\lambda : [0, T] \to \mathbb{R}^n$ that satisfies the *minimum condition* (4.1.20) and the *adjoint equation* (4.1.23)–(4.1.24).

Remark 4.9. (a) The way we obtained the Minimum Principle is similar to that in Bertsekas (2005). For a general treatment about the relationship between the Minimum Principle and the HJB equation, see Yong and Zhou (1999). In the latter reference, the stochastic case is also considered.

(b) If we consider maximization problems, then (4.1.20) is replaced by a *maximum condition*. This is one reason why the Minimum Principle is also known as *Maximum Principle*. Another name for the principle is *Pontryagin's Maximum Principle* because the Soviet mathematician L. S. Pontryagin started the research on a class of OCPs and formulated a first version of the above equations. According to Gamkrelidze (1999), the complete formulation of the Maximum Principle as well as a full proof were accomplished however by Pontryagin and two of his close collaborators V. Boltyanski and R. V. Gamkrelidze.

(c) In the optimal control literature, the adjoint equation (4.1.23) is also named *costate* equation whereas the variable λ is called adjoint or costate variable. In particular, in economics, λ can be interpreted as a *shadow price*; see, for instance, Sect. 2.2.4 in Sethi (2021). Likewise, the terminal condition (4.1.24) is also called *transversality condition*. ◊

The Minimum Principle provides only *necessary* conditions for optimality. However, the Minimum Principle along with some additional conditions are sufficient to find solutions in a class of OCPs. The following result is a simplified version of a theorem in Mangasarian (1966) (see Remark 4.11 below).

Theorem 4.10. *Let $x^*(\cdot)$, $a^*(\cdot)$, and $\lambda(\cdot)$ satisfy (4.1.20), (4.1.23), and (4.1.24). Suppose the following*

(a) A and X are convex sets,
(b) the functions $C(\cdot)$ and $H(t,\cdot,\cdot,\lambda(t))$ are convex and continuously differentiable for each t, and
(c) $a^*(t)$ is an interior point of A for each t.

Then $(x^*(\cdot), a^*(\cdot))$ minimizes the cost functional (4.1.1) subject to (4.0.1).

Proof. Consider a control function $a(\cdot) \in \mathcal{A}[0,T]$, and let $x(\cdot)$ be the corresponding solution to (4.0.1). Put

$$D := \int_0^T c(t, x^*(t), a^*(t))\,dt + C(x^*(T)) - \int_0^T c(t, x(t), u(t))\,dt - C(x(T)),$$

$$\begin{aligned}
\mathcal{H}^*(t) &:= H(t, x^*(t), a^*(t), \lambda(t)), \qquad 0 \le t < T,\\
\mathcal{H}_x^*(t) &:= H_x(t, x^*(t), a^*(t), \lambda(t)), \qquad 0 \le t < T,\\
\mathcal{H}_a^*(t) &:= H_a(t, x^*(t), a^*(t), \lambda(t)), \qquad 0 \le t < T,
\end{aligned}$$

and

$$\mathcal{H}(t) := H(t, x(t), a(t), \lambda(t)), \qquad 0 \le t < T.$$

By assumptions (b)–(c) and (4.1.23),

$$\begin{aligned}
\mathcal{H}^*(t) - \mathcal{H}(t) &\le \mathcal{H}_x(t) \cdot [x^*(t) - x(t)] + \mathcal{H}_a^* \cdot [a^*(t) - a(t)]\\
&= \mathcal{H}_x(t) \cdot [x^*(t) - x(t)] + 0\\
&= -\dot{\lambda}(t) \cdot [x^*(t) - x(t)]
\end{aligned} \tag{4.1.25}$$

and

$$C(x^*(T)) - C(x(T)) \le C_x(x^*(T)) \cdot [x^*(T) - x(T)].$$

Then

$$
\begin{aligned}
D &= \int_0^T [\mathcal{H}^*(t) - \lambda(t)\dot{x}^*(t)]dt + C(x^*(T)) \\
&\qquad - \int_0^T [\mathcal{H}(t) - \lambda(t)\dot{x}(t)]dt - C(x(T)) \\
&= \int_0^T [\mathcal{H}^*(t) - \mathcal{H}(t)]dt + \int_0^T \lambda(t)[\dot{x}(t) - \dot{x}^*(t)]dt \\
&\qquad + C(x^*(T)) - C(x(T)) \\
&\leq \int_0^T \dot{\lambda}(t)[x(t) - x^*(t)]dt + \int_0^T \lambda(t)[\dot{x}(t) - \dot{x}^*(t)]dt \\
&\qquad + C_x(x^*(T)) \cdot [x^*(T) - x(T)] \\
&= \int_0^T \left(\frac{d}{dt}\Big\{ \lambda(t)[x(t) - x^*(t)] \Big\} \right) dt + C_x(x^*(T)) \cdot [x^*(T) - x(T)] \\
&= \lambda(T)[x(T) - x^*(T)] - 0 + C_x(x^*(T)) \cdot [x^*(T) - x(T)] \\
&= 0.
\end{aligned}
$$

That is, $D \leq 0$ so

$$
\int_0^T c(t, x^*(t), a^*(t))\, dt + C(x^*(T)) \leq \int_0^T c(t, x(t), u(t))\, dt + C(x(T)),
$$

which yields the required conclusion. □

Remark 4.11. We point out that Theorem 4.10 is still valid without the assumption (c). Indeed, from (4.1.25), we notice that the inequality

$$
\mathcal{H}_a^* \cdot [a^*(t) - a(t)] \leq 0
$$

(instead of the equality) is enough to prove the theorem. The latter inequality holds under assumptions (a) and (b) of the theorem—see Proposition 2.1.1 in Borwein and Lewis (2006). ◊

Theorem 4.10 can be used to solve OCPs according to the following procedure:

Step 1. From (4.1.20), find a^* (as a function of x^* and λ) that minimizes the Hamiltonian.

Step 2. Substitute a^* in (4.0.1) and (4.1.23) to obtain a system of two ordinary differential equations in the variables x^* and λ.

Step 3. Determine x^* and λ by using the initial condition $x^*(0) = x_0$ and the terminal condition (4.1.24).
Step 4. Go to *Step 1* and find the open-loop policy a^*.
Step 5. Verify the assumptions (a), (b), and (c) in Theorem 4.10.

The following example illustrates the above procedure to find an optimal policy of a scalar *LQ* system, by means of the Minimum Principle.

Example 4.12. Let $X = A = \mathbb{R}$, $\gamma, \beta \in \mathbb{R}$, $r > 0$, and $q \geq 0$. Consider the linear system

$$\dot{x}(t) = \gamma x(t) + \beta a(t), \qquad x(0) = x_0, \tag{4.1.26}$$

and the functional cost

$$\frac{1}{2}\int_0^T [qx^2(t) + a^2(t)]dt + \frac{r}{2}x^2(T).$$

For notational ease, we simply write $a(t)$ and $x(t)$ instead of $a^*(t)$ and $x^*(t)$. We observe that

$$a(t) = -\beta\lambda(t) \tag{4.1.27}$$

minimizes the Hamiltonian

$$H(t, x(t), a, \lambda(t)) = \frac{1}{2}[qx^2(t) + a^2] + \lambda(t)[\gamma x(t) + \beta a].$$

Thus the dynamics (4.1.26) and the adjoint equation (4.1.23) become

$$\begin{cases} \dot{x}(t) = \gamma x(t) + \beta a(t), & x(0) = x_0 \\ \dot{\lambda}(t) = -qx(t) - \gamma\lambda(t), & \lambda(T) = rx(T). \end{cases}$$

The solution $(x(\cdot), \lambda(\cdot))$ to this (linear and homogeneous) system of ordinary differential equations can be found by standard methods. Finally, notice that assumptions (a), (b), and (c) in Theorem 4.10 hold. Therefore, (4.1.27) gives an open–loop optimal policy. ◊

Remark 4.13. The Minimum Principle holds, with the corresponding changes, for other classes of OCPs. For instance, when

there is a constraint for the terminal state, say $g(x(T)) = 0$ or $h(x(T)) \geq 0$; see Chap. 5 in Li and Yong (1995). Another class consists of OCPs with infinite horizon; see Halkin (1974).

4.2 The Discounted Case

Consider the OCP at the beginning of this chapter, except that (4.1.1) is replaced with

$$J(a(\cdot)) := \int_0^T e^{-rt} c(t, x(t), a(t)) dt + e^{-rT} C(x(T)), \qquad (4.2.1)$$

where $r > 0$ is a given discount factor. In this case, the HJB equation (4.1.8)–(4.1.9) becomes

$$V_t + \inf_{a \in A} [e^{-rt} c(t, x, a) + V_x \cdot F(t, x, a)] = 0$$

for $(t, x) \in [0, T) \times X$, with terminal condition $V(T, x) = e^{-rT} C(x)$. With the change of variable $v(t, x) := e^{-rt} V(t, x)$, it can be seen that the HJB in the discounted case becomes

$$v_t + \inf_{a \in A} [c(t, x, a) + v_x \cdot F(t, x, a)] = rv \qquad (4.2.2)$$

for all $(t, x) \in [0, T) \times X$, and

$$v(T, x) = C(x) \quad \forall \ x \in X. \qquad (4.2.3)$$

Note that if $r = 0$, then (4.2.1) and (4.2.2) reduce to (4.1.1) and (4.1.8), respectively. Similarly, (4.1.11) and (4.1.13) become

$$v_t + \inf_{a \in A} H(t, x, a, v_x) = rv \qquad (4.2.4)$$

and

$$\inf_{a \in A} [c(t, x, a) + L^a v(t, x)] = rv \qquad (4.2.5)$$

for $(t, x) \in [0, T) \times X$, with the terminal condition (4.2.3). ◊

Example 4.14 (The discounted LQ case). In the general LQ problem (also known as the *linear regulator problem*) the state and

control spaces are general finite–dimensional spaces, say, $X = \mathbb{R}^n$ and $A = \mathbb{R}^m$, respectively. Here, however, to simplify the presentation we shall consider the scalar case $n = m = 1$, so $X = A = \mathbb{R}$. The state process is given by

$$\dot{x}(t) = \gamma(t)x(t) + \beta(t)a(t), \quad \text{for} \quad t \in [0, T], \tag{4.2.6}$$

with continuously differentiable coefficients $\gamma(\cdot)$ and $\beta(\cdot)$, and a given initial condition $x(0) = x_0$. The Eq. (4.2.6) is of course of the form (4.0.1) with

$$F(t, x, a) = \gamma(t)x + \beta(t)a.$$

The associated cost functional is a discounted cost, as in (4.2.1), with

$$c(t, x, a) := Q(t)x^2 + R(t)a^2 \quad \text{and} \quad C(x) := x^2$$

with coefficients $Q(\cdot) \geq 0$ and $R(\cdot) > 0$. Hence, the HJB equation (4.2.2) becomes

$$\begin{aligned} rv &= \inf_a [c(t, x, a) + v_t + v_x \cdot F(t, x, a)] \\ &= Qx^2 + v_t + v_x \cdot \gamma x + \inf_a [Ra^2 + v_x \cdot \beta a], \end{aligned} \tag{4.2.7}$$

with terminal condition $v(T, x) = x^2$. The minimum in (4.2.7) is attained at

$$a^*(t, x) = -(2R(t))^{-1}\beta(t) \cdot v_x. \tag{4.2.8}$$

Inserting this value of a^* in (4.2.7) we obtain the partial differential equation

$$Qx^2 + v_t + \gamma x v_x - (\beta v_x)^2/4Q = rv \tag{4.2.9}$$

for $(t, x) \in [0, T) \times X$, and $v(T, x) = x^2$. The problem now is how to obtain a solution to (4.2.9). By the form of the LQ problem (or by analogy with the discrete-time case—see Example 2.4), we may try a solution of the form

$$v(t, x) = k(t)x^2 + h(t) \quad \textit{for} \quad 0 \leq t < T, \tag{4.2.10}$$

with continuously differentiable coefficients $k(t), h(t)$, and $k(\cdot) \geq 0$. Note that the terminal condition $v(T,x) = k(T)x^2 + h(T) = x^2$ gives that

$$k(T) = 1,\ h(T) = 0. \tag{4.2.11}$$

From (4.2.10), $v_t = \dot{k}(t)x^2 + \dot{h}(t)$ and $v_x = 2k(t)x$. Replacing these values in (4.2.9) we obtain the equation

$$[\dot{k} + (2\gamma - r)k + (k\beta)^2/R + Q]x^2 + (\dot{h}(t) - rh(t)) = 0,$$

which in turn yields two ordinary differential equations:

$$\dot{k} + (2\gamma - r)k + (k\beta)^2/R + Q = 0, \tag{4.2.12}$$

and the linear equation $\dot{h}(t) = rh(t)$. In view of the terminal condition $h(T) = 0$ in (4.2.11), it can be seen that $h(t) = 0$ for all $t \in [0,T]$. Therefore, the function v in (4.2.10) becomes

$$v(t,x) = k(t)x^2,$$

where $k(t)$ is the solution of the *Riccati equation* (4.2.12). (The existence of this solution is ensured in our present context. See, for instance, Theorem 5.2 in Fleming and Rishel (1975), Sect. IV.5.) Finally, since $v_x = 2k(t)x$, from (4.2.8) we obtain that the optimal control for the LQ problem is $a^*(t,x) = -R(t)^{-1}\beta(t)k(t)x$, which is a linear function in the state x. ♢

4.3 Infinite–Horizon Discounted Cost

We consider again the system (4.0.1) except that now it is defined for all $t \geq 0$, i.e.,

$$\dot{x}(t) = F(t, x(t), a(t)) \quad \text{for} \quad t \geq 0, \tag{4.3.1}$$

with the same initial condition $x(0) = x_0$. Recall that the *state space* is $X := \mathbb{R}^n$. Moreover, instead of the discounted cost (4.2.1) we consider the infinite–horizon discounted cost functional

$$V(x_0; a(\cdot)) := \int_0^\infty e^{-rt} c(t, x(t), a(t)) dt.$$

We will assume that the running (or instantaneous) cost $c(t, x, a)$ is *nonnegative* and, in addition, the set $\mathcal{A}_{x_0}$ of piecewise–continuous controls $a(\cdot) : [0, \infty) \to A$ such that $V(x_0; a(\cdot)) < \infty$ is nonempty. In this context, the verification Theorem 4.6, concerning the HJB equation (4.2.2) (or (4.2.5)) becomes as follows.

Theorem 4.15. *Let $v : [0, \infty) \times X \to \mathbb{R}$ be a continuously differentiable function that satisfies the equation*

$$rv(s, x) = \inf_{a \in A} [c(s, x, a) + L^a v(s, x)] \tag{4.3.2}$$

for all $(s, x) \in [0, \infty) \times X$, with L^a as in (4.1.12). Then:

(a) $v(s, x) \leq V(s, x; a(\cdot))$ for each control $a(\cdot) \in \mathcal{A}_{x_0}$ such that, as $t \to \infty$,

$$e^{-rt} v(t, x(t)) \to 0, \tag{4.3.3}$$

where

$$V(s, x; a(\cdot)) := \int_s^\infty e^{-r(t-s)} c(t, x(t), a(t)) dt.$$

(b) If $a^* \equiv a^*(s, x) \in A$ attains the minimum in (4.3.2), i.e.

$$rv(s, x) = c(s, x, a^*) + L^{a^*} v(s, x) \quad \forall \ (s, x), \tag{4.3.4}$$

then $a^*(\cdot) \equiv a^*(\cdot, x(\cdot))$ is optimal within the class of controls in $\mathcal{A}_{x_0}$ that satisfy the condition (4.3.3).

Proof. (a) By (4.3.2),

$$rv(s, x) \leq c(s, x, a) + L^a v(s, x) \quad \forall \ (s, x, a). \tag{4.3.5}$$

Let $u(s, x) := e^{-rs} v(s, x)$ and note that, by (4.1.12),

$$\begin{aligned} L^a u(s,x) &:= u_s + u_x \cdot F(s,x,a) \\ &= e^{-rs}[v_s(s,x) - rv(s,x) + v_x(s,x) \cdot F(s,x,a)] \\ &= e^{-rs}[L^a v(s,x) - rv(s,x)] \\ &\geq -e^{-rs}c(s,x,a), \qquad \text{[by (4.3.5)]} \end{aligned}$$

that is, $L^a u(s,x) \geq -e^{-rs}c(s,x,a)$ for all (s,x,a). Integrating both sides of the latter inequality, and recalling the Remark 4.7(c), we obtain

$$u(T,x(T)) - u(s,x) \geq \int_s^T e^{-rt}c(t,x(t),a(t))dt. \tag{4.3.6}$$

Equivalently, by definition of u,

$$v(s,x) \leq \int_s^T e^{-r(t-s)}c(t,x(t),a(t))dt + e^{-r(T-s)}v(T,x(T)).$$

Finally, letting $T \to \infty$, (4.3.3) yields the desired conclusion in part (a).

(b) On the other hand, if (4.3.4) holds, we have equalities throught (4.3.5)–(4.3.6) with $a = a^*$, and part (b) follows. □

Example 4.16. Consider the OCP: Minimize, over all nonnegative controls $a(\cdot) \in \mathcal{A}_{x_0}$, the objective function

$$V(x_0; a(\cdot)) = \int_0^\infty e^{-rt}[x(t) + ka(t)^2]dt$$

subject to

$$\dot{x}(t) = p - a(t)x(t)^{1/2} \quad \forall \ t > 0,$$

with k and $x(0) = x_0$ both positive. Note that this is an *autonomous* or *stationary* or *time–invariant* OCP because both the instantaneous cost $c(t,x,a) \equiv c(x,a) = x + ka^2$ and the state transition function $F(t,x,a) \equiv F(x,a) = p - ax^{1/2}$ are independent of the time variable t, in which case the value function v depends only on the state variable x. Therefore, the HJB equation (4.3.2) becomes

$$rv = \inf_{a \geq 0}[x + ka^2 + v_x \cdot (p - ax^{1/2})], \tag{4.3.7}$$

that attains its minimum at $a^* \equiv a^*(x) = (2k)^{-1} v_x \cdot x^{1/2}$. Substituting this value in (4.3.7) we obtain

$$rv(x) = x + p \cdot v_x - (4k)^{-1} x \cdot v_x^2. \tag{4.3.8}$$

Guessing that a possible solution of this equation could be of the form $v(x) = Px + Q$, for some constants P, Q, substitution of this function $v(x)$ in (4.3.8) gives that it indeed solves (4.3.8) provided that P and Q satisfy the equations $rQ = pP$ and

$$P^2 + 4krP - 4k = 0. \tag{4.3.9}$$

Hence $v(x) = Px + Q$ solves (4.3.8) if $Q = pP/r$ and P is the positive solution of (4.3.9). Further, since $v_x = P$, the optimal control is $a^*(x) = (2k)^{-1} P x^{1/2}$. ◊

Remark 4.17. For OCPs in an infinite horizon there are several optimality concepts, in addition to the discounted cost in Sect. 4.3. For instance, Carlson et al. (1991), Sect. 1.5, introduce the following concepts: (a) Strong optimality; (b) Overtaking optimality; (c) Weak overtaking optimality; (d) Finite optimality. They also show the following chain of implications: (a) ⇒ (b) ⇒ (c) ⇒ (d). ◊

Example 4.18 (An infinite-horizon discounted LQ problem). Consider an infinite-horizon scalar (i.e., $X = A = \mathbb{R}$) LQ problem in which we wish to minimize

$$V(x; a(\cdot)) = \int_0^\infty e^{-rt} [Qx^2(t) + Ra^2(t)] dt \tag{4.3.10}$$

subject to

$$\dot{x}(t) = \delta x(t) + \eta a(t), \quad t \geq 0, \quad x(0) = x. \tag{4.3.11}$$

The coefficients Q and R are both positive, and $\eta \neq 0$. Since the system (4.3.10)–(4.3.11) is time-homogeneous or time-invariant (in particular, all the coefficients are constant), the HJB equation (4.2.2) or (4.3.2) for $v(x)$ becomes

$$rv = \inf_{a \in A} \left\{ Qx^2 + Ra^2 + v_x \cdot (\delta x + \eta a) \right\}. \tag{4.3.12}$$

As in similar LQ problems (see Examples 2.46 or 4.14, for instance), we seek a solution of (4.3.12) of the form $v(x) = kx^2 + h$, where $k > 0$ and h are constants to be determined. With this value of $v(\cdot)$, (4.3.12) can be expressed as

$$(rk - Q - 2\delta k)x^2 = \min_a \{Ra^2 + 2\eta kxa\}.$$

The minimum is attained at the stationary Markov control

$$f^*(x) = -R^{-1}\eta kx \quad \forall x \in \mathbb{R},$$

and the value function is $v_r(x) = V(x, f^*) = kx^2 + h$, where $h = 0$ and k is the positive solution of the quadratic equation

$$\eta^2 k^2 + (r - 2\delta)Rk - RQ = 0. \tag{4.3.13}$$

Finally, observe that f^* satisfies (4.3.3), i.e.,

$$e^{-rt}v(x^*(t)) \to 0 \quad \text{as} \quad t \to \infty.$$

Therefore Theorem 4.15 gives that f^* is indeed optimal in the class $\mathcal{A}_{x_0}$. ◊

4.4 Long-Run Average Cost Problems

Consider the infinite-horizon control system

$$\dot{x}(t) = F(x(t), a(t)), \quad t \geq 0, \quad x(0) = x, \tag{4.4.1}$$

with state space $X = \mathbb{R}^n$ and control (or action) set $A \subset \mathbb{R}^m$. We denote by $\mathcal{A}$ the family of piecewise–continuous control functions $a(\cdot) : [0, \infty) \to A$.

For each control $a(\cdot) \in \mathcal{A}$ and each $T > 0$, let

$$J_T(x, a(\cdot)) := \int_0^T c(x(t), a(t))dt, \tag{4.4.2}$$

where the running cost c is nonnegative. In this section we wish to minimize the long–run average cost (AC)

$$J(x, a(\cdot)) := \limsup_{T\to\infty} \frac{1}{T} J_T(x, a(\cdot)) \tag{4.4.3}$$

subject to (4.4.1). The AC-value function is

$$J^*(x) := \inf_{a(\cdot)} J(x, a(\cdot)) \tag{4.4.4}$$

for all $x \in X$. As usual, a control function a^* is said to be AC–optimal if

$$J(x, a^*(\cdot)) = J^*(x) \quad \text{for all} \quad x \in X.$$

We will assume the existence of a control function $a(\cdot)$ such that $J(x, a(\cdot)) < \infty$ for every $x \in X$. This condition ensures that $J^*(\cdot)$ is a finite–valued function.

The OCP (4.4.1)–(4.4.4) is, of course, a deterministic continuous–time version of the discrete–time AC problems in Sects. 2.5 and 3.7. Hence it is no surprise that some of the techniques for discrete–time problems are also applicable, with obvious changes, to the continuous–time case. These techniques include the *average cost optimality equation* (ACOE), the *steady state* approach, and the *vanishing discount* approach, which we introduce in the remainder of this section.

4.4.1 The Average Cost Optimality Equation (ACOE)

In addition to the AC control problem (4.4.1)–(4.4.3), consider the operator L^a in the Remark 4.7(d), i.e,

$$L^a v(x) = v_x(x) \cdot F(x, a) \tag{4.4.5}$$

for $v \in C^1(X)$. Note that, by (4.4.1) and the chain rule,

$$L^a v(x) = \frac{d}{dt} v(x(t))|_{(x,a)}. \tag{4.4.6}$$

The ACOE approach to the AC problem is based on the so–called *Poisson equation* in the following lemma.

Lemma 4.19. Let $j \in \mathbb{R}$ and $h(\cdot) \in C^1(X)$ be such that the pair (j, h) satisfies the Poisson equation

$$j = c(x, a) + L^a h(x) \ \ \forall\, x, a. \tag{4.4.7}$$

Moreover, let $a(\cdot) \in \mathcal{A}$ be a control function such that, as $t \to \infty$,

$$h(x^a(t))/t \to 0 \tag{4.4.8}$$

for every initial state $x(0) = x$, where x^a is the solution of (4.4.1) when using the control $a(\cdot)$. Then

(a) $j = J(x, a(\cdot))$ for all $x \in X$.
(b) If the equality in (4.4.7) is replaced by $\geq$ (resp. $\leq$), then in (a) we have $j \geq J(x, a(\cdot))$ (resp. $\leq$) for all x.

Proof.

(a) For notational ease, we will write $x^a(\cdot)$ as $x(\cdot)$. Then, by the Remark 4.7(d), the Poisson equation (4.4.7) yields

$$j = c(x(t), a(t)) + \frac{d}{dt} h(x(t)) \ \ \forall\, t \geq 0. \tag{4.4.9}$$

It follows that, for all $T > 0$,

$$Tj = \int_0^T c(x(t), a(t))dt + h(x(T)) - h(x). \tag{4.4.10}$$

Multiplying both sides by $1/T$ and then letting $T \to \infty$, from (4.4.8) and (4.4.3) we obtain part (a).

(b) In (4.4.7) replace $=$ with either $\geq$ or $\leq$. Then in (4.4.9)–(4.4.10) we obtain $\geq$ or $\leq$, respectively, in lieu of $=$. □

Remark. Observe that Lemma 4.19 tacitly assumes the existence of a solution (j, h) to the Poisson equation (4.4.7). In other words, the lemma itself does not guarantee the existence of such a solution. If, however, that solution exists, then necessarily j is unique and h is unique up to additive constants. A similar result holds for the ACOE (4.4.12) below. (See Exercise 4.14.) ◊

From the Poisson equation (4.4.7) we obtain the *average cost optimality equation* (ACOE) in Theorem 4.20, below, where we use the following notation: Given a function $h : X \to \mathbb{R}$, we denote by $\mathcal{A}_h$ the family of controls $a(\cdot) \in \mathcal{A}$ such that

$$\frac{1}{t} h(x^a(t)) \to 0 \quad \text{as} \quad t \to \infty \tag{4.4.11}$$

where x^a is the solution of (4.4.1) when using the control a, for any initial state $x(0) = x$.

The ACOE (4.4.12) is also known as the HJB (or the dynamic programming or simply the Bellman) equation for the AC control problem (4.4.1)–(4.4.3).

Theorem 4.20. *Let us assume that $j \in \mathbb{R}$ and $h \in C^1(X)$ form a solution to the ACOE*

$$j = \inf_{a \in A} [c(x, a) + L^a h(x)] \quad \forall\, x \in X. \tag{4.4.12}$$

Then, for every initial state $x(0) = x$,

(a) $j \le J(x, a(\cdot))$ for all $a(\cdot) \in \mathcal{A}_h$; hence
(b) $j \le J^*(x)$ if $\mathcal{A} = \mathcal{A}_h$.

In addition, let us suppose that there exists a control $a^*(\cdot) \in \mathcal{A}_h$ that attains the minimum in the right–hand side of (4.4.12), i.e., for every $x \in X$, $a^*(x) \in A$ is such that

$$j = c(x, a^*(x)) + L^{a^*(x)} h(x) \quad \forall\, x \in X. \tag{4.4.13}$$

Then, for all $x \in X$,

(c) $j = J(x, a^*(\cdot)) \le J(x, a(\cdot))$ for all $a(\cdot) \in \mathcal{A}_h$; hence
(d) $a^*(\cdot)$ is AC–optimal and $J(\cdot, a^*(\cdot)) \equiv J^*(\cdot) \equiv j$ if $\mathcal{A} = \mathcal{A}_h$.

Proof. If (4.4.12) holds, then

$$j \le c(x, a) + L^a h(x) \quad \forall\, (x, a) \in X \times A.$$

Consequently, (a) follows from Lemma 4.19(b). On the other hand, if $\mathcal{A} = \mathcal{A}_h$, then (b) follows from (a).

Suppose now that the Poisson equation (4.4.13) holds. Then Lemma 4.19(a) yields the equality in (c), whereas the inequality is obtained from (a). Finally, (d) follows from (c). □

Remark 4.21. If the pair (j, h) is a solution to the ACOE (4.4.12), then it is also called a *canonical pair*. Similarly, if (4.4.13) holds, then (j, h, a^*) is said to be a *canonical triplet*. Unfortunately, Theorem 4.20 does not say how to solve the ACOE, and all the known results (see, for instance, Arisawa (1997)) impose restrictive conditions such as *compact* state space and/or *compact* control set and/or *bounded* running cost. However, none of these conditions is satisfied in the following LQ example but still we do obtain a canonical triplet. ◊

Example 4.22. We again consider the scalar LQ system in Example 4.18 with state equation in (4.3.11), i.e.,

$$\dot{x}(t) = \delta x(t) + \eta a(t),\ t \geq 0, \quad x(0) = x,$$

and running cost $c(x, a) = Qx^2 + Ra^2$. In this case the ACOE (4.4.12) is

$$j = \inf_a [Qx^2 + Ra^2 + h'(x) \cdot (\delta x + \eta a)]. \tag{4.4.14}$$

In view of previous LQ examples, we conjecture that the function h is of the form $h(x) = bx^2$ for some constant $b > 0$. To verify that this is indeed the case, we insert h in (4.4.14) and obtain that, for each $x \in X$, the minimum is attained at

$$a^*(x) = -Bx \quad \text{with} \quad B := R^{-1}b\eta. \tag{4.4.15}$$

Therefore, (4.4.14) can be expressed as

$$j = (Q + 2b\delta - R^{-1}b^2\eta^2)x^2 \quad \forall\, x \in X.$$

This equation holds provided that $j = 0$ and b solves the quadratic equation

$$R^{-1}\eta^2 b^2 - 2\delta b - Q = 0. \tag{4.4.16}$$

To proceed further observe that, with a^* as in (4.4.15), the state equation becomes

$$\dot{x}(t) = \Delta x(t), \quad \text{with} \quad \Delta := R^{-1}(R\delta - b\eta^2).$$

Now, let us assume that b is such that $\Delta < 0$. Then the corresponding state trajectory

$$x^*(t) = x_0 e^{\Delta t} \to 0 \quad \text{as} \quad t \to \infty \tag{4.4.17}$$

for every initial state x_0. Moreover, $J(x, a^*(\cdot)) = 0 = j$ for all $x_0 = x$.

Finally, from (4.4.15)–(4.4.16) we conclude that, if b is the positive solution of (4.4.16), then $(j, h(\cdot), a^*(\cdot))$ is a canonical triplet for the ACOE and, by Theorem 4.20, a^* is AC–optimal in the class of controls $\mathcal{A}_h$.

It should be noted that the results in this "scalar" LQ example are valid in the general vector case, with state and control spaces $X = \mathbb{R}^n$ and $A = \mathbb{R}^m$, respectively, except that now we require special concepts from linear systems theory, say "stabilizability", "detectability", and so forth. For details, see Theorem 5.4.4 in Davis (1977), for instance. ◊

Example 4.23 (Example 4.16 cont'd.). Let us consider the AC problem (4.4.1)–(4.4.3) with transition function $F(x, a)$ and running cost $c(x, a)$ as in the Example 4.16, i.e.,

$$c(x,a){=}x{+}ka^2 \quad \text{and} \quad F(x,a){=}p - ax^{1/2}, \quad \text{with } x(0) = x > 0, \tag{4.4.18}$$

where k and p are given positive constants. The OCP is to minimize the average cost $J(x, a(\cdot))$ in (4.4.3) over all the controls $a(\cdot) \geq 0$ in $\mathcal{A}$. In this case, the ACOE (4.4.12) becomes

$$j = \inf_{a \geq 0}[x + ka^2 + h'(x) \cdot (p - ax^{1/2})], \tag{4.4.19}$$

where $h'(x) = dh(x)/dx$. By comparison with the r-discount HJB equation (4.3.7), we will propose $h(x) = Rx$ as a possible solution of (4.4.19), with the coefficient R to be determined. Replacing $h(\cdot)$ in (4.4.19) we obtain

$$j = x + pR + \inf_{a}[ka^2 - Rx^{1/2}a], \tag{4.4.20}$$

which attains the minimum at

$$a^*(x) = (2k)^{-1}Rx^{1/2} \quad \forall x \geq 0. \tag{4.4.21}$$

With this value of $a = a^*$, (4.4.20) becomes

$$j = pR + (1 - R^2/4k)x \quad \forall x \geq 0,$$

which yields $j = pR$ and $R^2/4k = 1$, i.e., $R = 2k^{1/2}$. Hence, we already have a canonical triplet (j, h, a^*), that is, a triplet satisfying (4.4.13). Finally, to use Theorem 4.20 we will identify the family $\mathcal{A}_h$ of controls that satisfy (4.4.11).

Let $F(x, a) = p - ax^{1/2}$ be as in (4.4.18), and $a(\cdot) = a^*(\cdot)$ as in (4.4.21). Then the corresponding state trajectory $x^*(\cdot)$ is the solution of the linear equation

$$\dot{x}(t) = -\delta x(t) + p, \quad \text{with} \quad \delta := R/2k = k^{-1/2},$$

for each initial $x(0) = x_0 > 0$. Hence, for some constant C,

$$x^*(t) = Ce^{-\delta t} + p/\delta,$$

which yields (4.4.11), that is, since $h(x) = Rx$,

$$\frac{1}{t}h(x^*(t)) \to 0 \text{ as } t \to \infty. \tag{4.4.22}$$

Therefore, by Theorem 4.20(c) we conclude that $a^*(\cdot)$ in (4.4.21) is AC-optimal within the class of controls that satisfy (4.4.22). ◊

4.4.2 The Steady-State Approach

Let us consider again the AC control problem (4.4.1)–(4.4.3). As in the discrete-time case (see (2.5.17)), the steady-state approach to the AC problem hinges on the existence of state-action pairs $(\bar{x}, \bar{a})$ which are "steady" in the sense that $F(\bar{x}, \bar{a}) = 0$. Let $\mathcal{K} \subset X \times A$ be the set of all such pairs. Then a pair $(\bar{x}, \bar{a})$ in $\mathcal{K}$ is said to be a *minimum steady pair* if it is a solution of the constrained optimization problem:

$$\text{minimize } c(x,a) \text{ subject to } F(x,a)=0. \tag{4.4.23}$$

We denote by $\mathcal{K}^*$ the family of minimum steady pairs. We wish to find conditions under which a minimum steady pair gives the AC-value function J^* in (4.4.4). In particular, the following Theorem 4.24 mimics the discrete-time result in Theorem 2.53.

Theorem 4.24. *Suppose that the running cost c is continuous and, in addition, the AC control problem (4.4.1)–(4.4.3) is such that:*

- *(a_1) There exists a minimum steady pair $(x^*,a^*)\in\mathcal{K}^*$.*
- *(a_2) The control system is dissipative with respect to the pair (x^*,a^*) in (a_1), in the following sense: There exists a real-valued function $l\in C^1(X)$ such that*

$$c(x^*,a^*)\le c(x,a)+L^a l(x) \tag{4.4.24}$$

  for all $x\in X, a\in A$.
- *(a_3) The control system is stabilizable, that is, there exists a control $\bar a(\cdot)\in\mathcal{A}_l$ such that, as $t\to\infty$,*

$$(\bar x(t),\bar a(t))\to(x^*,a^*) \tag{4.4.25}$$

  with $\mathcal{A}_l$ as in (4.4.11).

Then $j^*:=c(x^*,a^*)$ is such that, for all $x\in X$,

- *(b_1) $J(x,\bar a(\cdot))=j^*$ and, moreover, $j^*\le J(x,a(\cdot))$ for all $a(\cdot)\in\mathcal{A}_l$;*
- *(b_2) The control $\bar a(\cdot)$ is AC-optimal and the AC-value function is $J^*(\cdot)\equiv j^*$ if $\mathcal{A}=\mathcal{A}_l$.*

Proof. (b_1) Since c is continuous, (4.4.25) yields

$$c(\bar x(t),\bar a(t))\to c(x^*,a^*)=:j^*$$

as $t\to\infty$. This implies that $J(\cdot,\bar a(\cdot))\equiv j^*$. Moreover, the inequality $j^*\le J(\cdot,a(\cdot))$ follows from (4.4.24) and Lemma 4.19(b) (with the inequality $\le$).

On the other hand, if $\mathcal{A}=\mathcal{A}_l$, then ($b_2$) follows from ($b_1$). □

Remark 4.25. (a) Under the hypotheses of Theorem 4.24, (x^*, a^*) is a so-called *minimum pair* in the sense that

$$j^* := c(x^*, a^*) = \inf_x \inf_{a(\cdot)} J(x, a(\cdot)).$$

(b) Recalling (4.4.6), we can rewrite (4.4.24) as

$$l(x(t)) - l(x) + \int_0^t [c(x(s), a(s)) - c(x^*, a^*)]ds \geq 0$$

for any solution $(x(\cdot), a(\cdot))$ of (4.4.1). (Reader beware: The definition of "dissipativity" is not standard; that is, different authors may use different definitions.)

Example 4.26. Consider the LQ system in Example 4.22, in which the system function and the running cost are

$$F(x, a) = \delta x + \eta a \quad \text{and} \quad c(x, a) = Qx^2 + Ra^2,$$

respectively. Hence, (x, a) is a steady state-action pair if $F(x, a) = 0$, which holds if $a = -\delta x/\eta$. Replacing this value in $c(x, a)$ we see that

$$c(x, a) = (Q + R\delta^2/\eta^2)x^2.$$

Therefore, we have the minimum steady pair $(x^*, a^*) = (0, 0)$, that is, the hypothesis (a_1) in Theorem 4.24. The hypotheses (a_2) and (a_3) can be obtained from (4.4.14)–(4.4.17). ◊

Example 4.27. We continue Example 4.23. From (4.4.18), $F(x, a) = 0$ if $a = px^{-1/2}$. With this value of a, we obtain

$$c(x, a) = x + ka^2 = x + kp^2/x$$

which is minimized at $x = pk^{1/2} > 0$. Thus, we have the minimum steady pair $(x^*, a^*) = (pk^{1/2}, p^{1/2}k^{-1/4})$, with corresponding minimum AC cost $j^* = c(x^*, a^*) = 2pk^{1/2}$ as in Example 4.23. The latter example also yields the remaining parts of Theorem 4.24. ◊

Remark 4.28. In Example 4.29, below, we wish to *maximize* over $\mathcal{A}$ a long-run *average reward* (AR) defined as

$$J_{AR}(x, a(\cdot)) := \liminf_{T\to\infty} \frac{1}{T} \int_0^T r(x(t), a(t))dt$$

subject to (4.4.1), where $r(x, a)$ is the given running (or instantaneous) reward function. In this case, the ACOE (4.4.12) is replaced by the average reward optimality equation (AROE)

$$\rho = \sup_{a\in A}[r(x, a) + L^a h(x)] \quad \forall x \in X \tag{4.4.26}$$

for some function $h \in C^1(X)$ and some constant ρ. Theorems 4.20 and 4.24 are modified accordingly. In particular, the existence of a minimum steady pair is replaced by a *maximum steady pair* (x^*, a^*) that solves the constrained optimization problem:

$$\text{Maximize} \quad r(x, a) \quad \text{subject to} \quad F(x, a) = 0, \tag{4.4.27}$$

whereas the dissipativity inequality (4.4.24) is replaced by

$$r(x^*, a^*) \geq r(x, a) + L^a h(x)$$

for all $x \in X, a \in A$. ◊

Example 4.29 (Control of pollution accumulation). The control of pollution accumulation is a standard OCP in environmental economics since the years 1970s. Here we consider a special case of an application by Kawaguchi (2003) in which he wishes to obtain a consumption strategy $a(\cdot)$ that maximizes the long-run average welfare

$$J_{AR}(a(\cdot)) = \liminf_{T\to\infty} \frac{1}{T} \int_0^T [U(a(t)) - D(x(t))]dt \tag{4.4.28}$$

where $x(t)$ is the stock of pollution at time t associated to $a(\cdot)$. Furthermore, $U : A \to [0, \infty)$ is a social utility function and $D : X \to [0, \infty)$ is a disutility function, with state and control sets $A = X = [0, \infty)$. Both functions are assumed to be continuously differentiable functions and satisfying suitable concavity/convexity conditions. The stock of pollution evolves according to the equation

$$\dot{x}(t) = a(t) - \varphi(x(t)), \quad \text{with} \quad x(0) = x_0 > 0 \tag{4.4.29}$$

where $\varphi(\cdot)$ denotes the rate of pollution decay.

Kawaguchi gives general conditions ensuring that the HJB equation associated to (4.4.28)–(4.4.29) has a "classical" solution. Here, however, to illustrate the steady state approach we will suppose that the utility and disutility functions and the pollution decay are of the form

$$U(a) = 2a^{1/2}, \quad D(x) = d_1 x, \quad \varphi(x) = d_0 x \tag{4.4.30}$$

for some positive constants d_0, d_1. Hence, to obtain a maximum steady state-action pair first note that, from (4.4.29)–(4.4.30),

$$F(x, a) = a - \varphi(x) = a - d_0 x = 0$$

holds if $a^* = a^*(x) = d_0 x$. With this value of a^* the running reward $r(x, a) := U(a) - D(x)$ becomes $r(x, a) = 2(d_0 x)^{1/2} - d_1 x$, which is maximized at $x^* = d_0/d_1^2$. Therefore, we have obtained a maximum steady pair $(x^*, a^*) = (d_0/d_1^2, (d_0/d_1)^2)$, as in the hypothesis (a_1) of Theorem 4.24. Hence

$$r^* := r(x^*, a^*) = U(a^*) - D(x^*) = d_0/d_1. \tag{4.4.31}$$

To verify the hypotheses (a_2)-(a_3), let us try to obtain the AROE in (4.4.26), i.e., from (4.4.28)–(4.4.30),

$$\begin{aligned} r^* &= \sup_{a>0}[r(x, a) + h'(x)F(x, a)] \\ &= \sup_{a>0}[U(a) - D(x) + h'(x)(a - d_0 x)] \\ &= -(d_1 + d_0 h'(x))x + \sup_{a>0}[2a^{1/2} + h'(x)a]. \end{aligned} \tag{4.4.32}$$

Clearly, the maximum at the right-hand side of (4.4.32) is attained at

$$a^*(x) = 1/(h'(x))^2. \tag{4.4.33}$$

Finally, in (4.4.32) let us try a solution $h(\cdot)$ of the form $h(x) = h_0 x$ for a constant h_0 to be determined. We then see that (4.4.32) indeed holds with $h(x) = h_0 x$ if $h_0 = d_1/d_0$. Moreover, (4.4.33)

shows that the AR-optimal control is the constant $a^*(\cdot) \equiv (d_0/d_1)^2$, which is the same as a^* in (4.4.31). ◊

4.4.3 The Vanishing Discount Approach

As in Sect. 2.5 for discrete-time OCPs, we can relate AC and discounted cost problems directly from the definition of a discounted cost functional. To this end, consider the problem of minimizing the discounted cost

$$V_r(x, a(\cdot)) = \int_0^\infty e^{-rt} c(x(t), a(t)) dt \tag{4.4.34}$$

for a given discount factor $r > 0$, subject to the dynamics (4.4.1). Inside the integral in (4.4.34) replace the cost $c(x, a)$ by $c(x, a) \pm M$ for some constant M. Then (4.4.34) can be expressed as

$$V_r(x, a(\cdot)) = \int_0^\infty e^{-rt} [c(x(t), a(t)) - M] dt + \frac{M}{r};$$

that is,

$$rV_r(x, a(\cdot)) = M + r \int_0^\infty e^{-rt} [c(x(t), a(t)) - M] dt.$$

In particular, taking M as the average cost $j(x) := J(x, a(\cdot))$, we obtain

$$rV_r(x, a(\cdot)) = j(x) + r \int_0^\infty e^{-rt} [c(x(t), a(t)) - j(x)] dt. \tag{4.4.35}$$

This relation obviously suggests that, as $r \to 0^+$,

$$rV_r(x, a(\cdot)) \to j(x) \tag{4.4.36}$$

provided that the rightmost term in (4.4.35) tends to 0 as $r \downarrow 0$. An example of this situation is in the context of Theorem 4.24, as in the following Proposition 4.30. (The proof is left to the reader: Exercise 4.11.)

Proposition 4.30. Suppose that c is continuous and that the hypotheses (a_1) and (a_3) of Theorem 4.24 hold; that is, there exists

a minimum steady pair (x^*, a^*) and a control $\bar{a}(\cdot)$ that satisfy (4.4.25). Let $j^* := c(x^*, a^*)$. Then

$$\lim_{r \downarrow 0} r V_r(x, \bar{a}(\cdot)) = J(x, \bar{a}(\cdot)) = j^*$$

for every initial state $x(0) = x$.

In general, (4.4.36) can be obtained by means of an Abelian theorem as in parts (b)–(c) of the next lemma, where we use essentially the same terminology as in Lemma 2.56.

Lemma 4.31. For $t \geq 0$, let $\psi(t)$ be a nondecreasing continuous function with $\psi(0) = 0$. Define the upper and lower (Cesàro) limits

$$C^L := \liminf_{t \to \infty} \psi(t)/t, \quad C^U := \limsup_{t \to \infty} \psi(t)/t,$$

and the lower and upper Abelian limits

$$A^L := \liminf_{r \downarrow 0} r \int_0^\infty e^{-rt} d\psi(t), \quad A_U := \limsup_{r \downarrow 0} r \int_0^\infty e^{-rt} d\psi(t).$$

Suppose that

$$C^U < \infty. \tag{4.4.37}$$

Then:

(a) $\int_0^\infty e^{-rt} d\psi(t) = r \int_0^\infty e^{-rt} \psi(t) dt$ for every $r > 0$.
(b) $C^L \leq A^L \leq A^U \leq C^U$.
(c) If the limit $j := \lim_{t \to \infty} \psi(t)/t$ exists, then

$$\lim_{r \downarrow 0} r \int_0^\infty e^{-rt} d\psi(t) = j.$$

(In other words, if $C^L = C^U = j$, then $A^L = A^U = j$.)

Proof. Part (a) follows from the integration-by-parts formula

$$\int_0^t e^{-rs} d\psi(s) = e^{-rt} \psi(t) + r \int_0^t e^{-rs} \psi(s) dt, \tag{4.4.38}$$

and noting that (4.4.37) implies

$$\limsup_{t \to \infty} e^{-rt} \psi(t) = 0. \tag{4.4.39}$$

To prove the third inequality in (b), we use (4.4.37) again and choose an arbitrary $\epsilon > 0$ and let $T = T(\epsilon)$ be such that

$$\sup_{s \geq t} \psi(s)/s \leq C^U + \epsilon \quad \forall t \geq T.$$

Hence

$$\begin{aligned} r^2 \int_T^t e^{-rs} \psi(s) ds &= r^2 \int_T^t s e^{-rs} [\psi(s)/s] ds \\ &\leq (C^U + \epsilon) r^2 \int_T^t s e^{-rs} ds \\ &\leq C^U + \epsilon, \end{aligned}$$

because $r^2 \int_0^t s e^{-rs} ds = 1 - re^{-rt}(1/r + t) \leq 1$. Thus, for $t \geq T$, (4.4.38) gives

$$r \int_0^t e^{-rs} d\psi(s) \leq re^{-rt} \psi(t) + r^2 \int_0^T e^{-rs} \psi(s) ds + C^U + \epsilon.$$

Letting $t \to \infty$ and then $r \downarrow 0$, from (4.4.39) we obtain the third inequality in (b), since ϵ was arbitrary. The first inequality is proved similarly, and the second one is obvious.

Finally, Part (c) follows from (b). □

Lemma 4.31 is a well-known result in Laplace transform theory (see, for instance, Widder (1941), pp. 181–182).

Let us now suppose that the instantaneous cost function $c(x, a)$ is continuous and nonnegative. Let $a(\cdot) \in \mathcal{A}$ be any given control function and $x(\cdot)$ the corresponding solution of (4.4.1). In addition, let

$$\psi(t) := \int_0^t c(x(s), a(s)) ds, t \geq 0.$$

Then (4.4.2)–(4.4.3) yield that the average cost $J(x, a(\cdot))$ can be expressed as

$$J(x, a(\cdot)) = \limsup_{t \to \infty} \psi(t)/t$$

whereas the r-discounted cost becomes

$$V_r(x, a(\cdot)) := \int_0^\infty e^{-rt} c(x(t), a(t)) dt$$
$$= \int_0^\infty e^{-rt} d\psi(t).$$

Hence, from the third inequality in Lemma 4.31(b),

$$\limsup_{r \downarrow 0} r V_r(x, a(\cdot)) \leq J(x, a(\cdot)).$$

Consequently, since the control function $a(\cdot)$ was arbitrary, we conclude the following.

Proposition 4.32. The AC-value function $J^*(\cdot)$ and the r-discount value function $V_r(x) := \inf_{a(\cdot)} V_r(x, a(\cdot))$ satisfy that

$$\limsup_{r \downarrow 0} r V_r(x) \leq J^*(x) \tag{4.4.40}$$

for all $x \in X$.

In words, (4.4.40) states that, for r sufficiently small, $rV_r(\cdot)$ is a lower bound for $J^*(\cdot)$. We are actually interested in the "convergence", in some sense, of rV_r to some particular value of J^* as $r \downarrow 0$. We next explain this.

Suppose that the value function V_r is in $C^1(X)$ and satisfies the HJB equation (4.3.2) in the autonomous (or time-homogeneous) case, i.e.,

$$r V_r(x) = \min_{a \in A} [c(x, a) + L^a V_r(x)], \quad x \in X. \tag{4.4.41}$$

Now pick (and fix) a state $\bar{x} \in X$, and let

$$m_r := r V_r(\bar{x}) \quad \text{and} \quad l_r(x) := V_r(x) - V_r(\bar{x}) \tag{4.4.42}$$

for $x \in X$. Then we can rewrite (4.4.41) as

$$m_r + r l_r(x) = \min_{a \in A} [c(x, a) + L^a l_r(x)].$$

Finally, the key step is to find a sequence $r_n \downarrow 0$ and a pair $(j, l(\cdot))$ in $\mathbb{R} \times C^1(X)$ such, as $n \to \infty$,

$$m_{r_n} \to j, \quad l_{r_n}(\cdot) \to l(\cdot), \tag{4.4.43}$$

and $(j, l(\cdot))$ satisfies either the ACOE (4.4.12) or the *AC optimality inequality* (ACOI)

$$j \geq \inf_{a \in A}[c(x,a) + L^a l(x)]. \tag{4.4.44}$$

Hence, if there exists a control $a^*(\cdot) \in \mathcal{A}_h$ that attains the minimum in (4.4.44), i.e.,

$$j \geq c(x, a^*(x)) + L^{a^*(x)} l(x) \quad \forall x \in X, \tag{4.4.45}$$

then $a^*(\cdot)$ is AC-optimal and j is the AC-value function, i.e., $J^*(\cdot) \equiv j$.

The good news is that the procedure (4.4.43) works in many *particular* AC control problems. The bad news however is that, to the best of our knowledge, there are no general results ensuring the existence of such a pair $(j, l(\cdot))$. (The existing results require restrictive hypotheses. See Bardi and Capuzzo-Dolcetta (1997), Sect. VII.1, for instance.) We will next show some particular cases.

Example 4.33 (Example 4.16 cont'd.). Consider the AC control problem (4.4.1)–(4.4.3) with system function and running cost as in Example 4.16, that is,

$$F(x,a) = p - ax^{1/2}, \quad c(x,a) = x + ka^2$$

with k and $x(0) = x_0$ both positive. In the r-discounted case, Example 4.16 shows that the OCP value function and the r-optimal control are

$$V_r(x) = P(r)x + Q(r) \quad \text{and} \quad a^*(x) = (2k)^{-1}P(r)x^{1/2},$$

where $Q(r) = pP(r)/r$ and and $P = P(r)$ is the positive solution of (4.3.9). Hence, for any given (fixed) state $\bar{x} \geq 0$ (4.4.42) gives

$$m_r = rV_r(\bar{x}) = r[P(r)\bar{x} + Q(r)], \quad l_r(x) = V_r(x) - V_r(\bar{x}) = P(r)(x - \bar{x}).$$

It follows that, as $r \downarrow 0$,

$$m_r \to pP(0) \quad \text{and} \quad l_r(x) \to P(0)(x - \bar{x}).$$

In other words, (4.4.43) holds with $j = pP(0)$ and $l(x) = P(0)(x - \bar{x})$, where $P(0) = 2k^{1/2}$ is the positive solution of (4.3.9). Moreover, the AC-optimal control is $a^*(x) = (2k)^{-1} P(0)x^{1/2}$. ◊

Example 4.34 (Example 4.18 cont'd.). Consider again the r-discounted LQ problem in Example 4.18, in which the optimal control is $f_r^*(x) = -R^{-1}\eta k x$ with R and η as in (4.3.10)–(4.3.11), and $k = k(r)$ is the unique positive solution of (4.3.13). The corresponding value function is $v_r(x) = k(r)x^2$. In (4.4.42) we can take an arbitrary state $\bar{x}$. However, to simplify the presentation we take $\bar{x} = 0$. Therefore, $m_r = 0$ and

$$\begin{aligned} l_r(x) &= v_r(x) \\ &= k(r)x^2 \\ &\to k(0)x^2 \end{aligned}$$

as $r \downarrow 0$, where $k(0) = (RQ)^{1/2}/\eta$ is the positive solution of (4.3.13) with $r = 0$.

4.5 The Policy Improvement Algorithm

We will now introduce the policy improvement (or policy iteration) algorithm (PIA) for the discounted and the average cost problems in Sects. 4.3 and 4.4. The general ideas are, of course, similar to the discrete-time cases in Sects. 2.4 and 3.6 for deterministic and stochastic problems, respectively. Namely, we wish to find a sequence of control functions $a_n \in \mathcal{A}$ such that, for each n, a_{n+1} *improves* a_n in the sense that the cost v_{n+1} when using a_{n+1} is "better" than the cost v_n when using a_n, because $v_{n+1} \leq v_n$. Therefore, the cost functions v_n form a monotone *nonincreasing* sequence.

4.5.1 The PIA: Discounted Cost Problems

Given a discount factor $r > 0$, the OCP is to minimize the discounted cost

$$V(x, a(\cdot)) := \int_0^\infty e^{-rt} c(x(t), a(t)) dt \tag{4.5.1}$$

subject to

$$\dot{x}(t) = F(x(t), a(t)), \quad t \geq 0, \quad x(0) = x. \tag{4.5.2}$$

The running cost c is supposed to be nonnegative, and the controls are restricted to the class $\mathcal{A}_{SM}^* \subset \mathcal{A}$ of stationary Markov controls for which

$$e^{-rt} V(x(t), a(\cdot)) \to 0 \quad \text{as} \quad t \to \infty. \tag{4.5.3}$$

Recall from Chap. 1 that a stationary Markov control (also known as a feedback or closed-loop control) is a function $f(\cdot) : X \to A$ such that, at any time $t \geq 0$, the control action is $f(x) \in A$ if $x(t) = x$. We will denote by $\mathcal{A}_{SM}$ the family of stationary Markov controls, and by $\mathcal{A}_{SM}^*$ the subset of controls that satisfy (4.5.3).

For notational convenience we will write Markov controls either as $f(\cdot)$ or $a(\cdot)$. Moreover, we will write $c(x, f(x))$ and $F(x, f(x))$ as $c(x, f)$ and $F(x, f)$, respectively.

Before describing the PIA let us note the following.

Initialization. Let $f_0 \in \mathcal{A}_{SM}^*$ be a control with discounted cost $v^0(\cdot) := V(\cdot, f_0) \in C^1(X)$, so that

$$r v^0(x) = c(x, f_0) + v_x^0(x) F(x, f_0) \tag{4.5.4}$$

for all $x \in X$. The latter equation yields

$$r v^0(x) \geq \inf_{a \in A} [c(x, a) + v_x^0(x) F(x, a)]. \tag{4.5.5}$$

Let us now assume that there exists $f_1 \in \mathcal{A}_{SM}^*$ that attains the minimum in (4.5.5); that is, for each $x \in X$, $f_1(x) \in A$ is such that

$$c(x, f_1) + v_x^0(x)F(x, f_1) = \min_{a\in A}[c(x, a) + v_x^0 F(x, a)].$$

Therefore, we can rewrite the inequality (4.5.5) as

$$rv^0(x) \geq c(x, f_1) + v_x^0(x)F(x, f_1) \quad \forall x \in X. \tag{4.5.6}$$

A key step in the PIA is to show that (4.5.6) implies that f_1 *improves* f_0 in the sense that $v^0(\cdot) \geq v^1(\cdot)$, where $v^1(x) := V(x, f_1)$. More precisely, we have the following.

Lemma 4.35. For any two controls $f \equiv f_0$ and $g \equiv f_1$ in $\mathcal{A}_{SM}^*$ that satisfy (4.5.4) and (4.5.6) we have $V(x, f) \geq V(x, g)$ for all $x \in X$.

Proof. Let $u(t, x) := e^{-rt}v^0(x)$. Then, from (4.1.12),

$$\begin{aligned} L^{f_1}u(t, x) &= u_t + u_x \cdot F(x, f_1) \\ &= e^{-rt}[v_x^0(x) \cdot F(x, f_1) - rv^0(x)] \\ &\leq -e^{-rt}c(x, f_1). \quad \text{[by (4.5.6)]} \end{aligned}$$

Thus, recalling from Remark 4.7(c) that $L^a u(t, x) = du(t, x(t))/dt|_{(x,a)}$, integration of both sides of the latter inequality from $t = 0$ to $t = T$ gives

$$e^{-rT}v^0(x(T)) - v^0(x(0)) \leq -\int_0^T e^{-rt}c(x(t), f^1)dt.$$

Finally, letting $T \to \infty$ we see that $-v^0(x) \leq -v^1(x)$, which yields the desired result. □

Having the Initialization step and Lemma 4.35 we can proceed with the PIA as follows, where f_n is a control in $\mathcal{A}_{SM}^*$ with r-discounted cost $v^n(\cdot) := V(\cdot, f_n) \in C^1(X)$.

(PI$_1$) Given f_n $(n = 0, 1, \ldots)$ compute the corresponding discounted cost $v^n(\cdot)$ so, for every $x \in X$,

$$rv^n(x) = c(x, f_n)) + v_x^n(x)F(x, f_n).$$

Hence

$$rv^n(x) \geq \inf_{a\in A}[c(x, a) + v_x^n(x)F(x, a)]. \tag{4.5.7}$$

(PI_2) *Policy improvement.* Assume that there exists $f_{n+1} \in \mathcal{A}^*_{SM}$ such that, for every $x \in X$, $f_{n+1}(x) \in A$ attains the minimum in (4.5.7), that is,

$$c(x, f_{n+1}) + v^n_x(x)F(x, f_{n+1}) = \min_{a\in A}[c(x, a) + v^n_x(x)F(x, a)]. \tag{4.5.8}$$

If $v^{n+1}(\cdot) \equiv v^n(\cdot)$, then stop the algorithm because v^n is the optimal discounted cost (see Proposition 4.36(a) below). Otherwise, replace n by $n + 1$ and go back to (PI_1).

The existence of f_{n+1} as in (PI_2) is ensured by well known results. See, for instance, Lemma 2.16(a) above or Theorems B.8 and B.9 in Appendix B.

A first step on the convergence of the PIA is the following.

Proposition 4.36. Let f_n and v^n $(n = 0, 1, \ldots)$ be as in (PI_1) and (PI_2).

(a) If for some n we have $v^n(x) = v^{n+1}(x)$ for all $x \in X$, then $v^n(\cdot) \equiv V(\cdot)$ is the discounted value function, and f_n and f_{n+1} are optimal controls.

(b) In general, there exists a function $v \geq 0$ such that, for every $x \in X$, $v^n(x) \downarrow v(x)$.

Proof. (a) If $v^n(\cdot) = v^{n+1}(\cdot)$, then in the right-hand side of (4.5.7) we can replace v^n with v^{n+1}, which combined with (4.5.8) gives

$$rv^n(x) \geq \inf_a[c(x, a) + v^{n+1}_x(x)F(x, a)] = rv^{n+1}(x)$$

for all $x \in X$. Therefore, $v(\cdot) := v^n(\cdot) = v^{n+1}(\cdot)$ satisfies the r-discounted cost HJB equation

$$rv(x) = \min_{a\in A}[c(x, a) + v_x(x) \cdot F(x, a)], \tag{4.5.9}$$

and f_n and f_{n+1} are optimal controls.

(b) This part is a consequence of the monotonicity of $\{v^n\}$. □

Remark 4.37. Unfortunately, the conditions we have so far on the OCP (4.5.1)–(4.5.2) are not enough to guarantee two key steps in the PIA, namely:

(I) *Convergence to the value function*; that is, it remains to show that the limiting function v in Proposition 4.36(b) is in $C^1(X)$ and that it satisfies the HJB equation (4.5.9).

(II) *Convergence of controls*; that is, convergence of f_n (or a subsequence thereof) to a control $f \in \mathcal{A}^*_{SM}$ that attains the minimum in (4.5.9), and so it is optimal for (4.5.1)–(4.5.2).

To deal with (I), there are three usual options:

I_1. Impose suitable assumptions on the OCP, as in Doshi (1976a) or Jacka and Mijatović (2017), for instance. Typically, the idea is to impose conditions ensuring that the Arzela-Ascoli Theorem in Remark 2.59 is applicable, and then one shows that v preserves some of the properties of v^n, such as $v \in C^1(X)$. (As an example of this approach see Kawaguchi (2003)). In general, however, the assumptions are so restrictive that are not applicable to, for instance, our Example 4.38 below. (Indeed, the conditions on the PIA usually require compact state space X and/or compact control set A and/or bounded running cost $c(x, a)$. None of these conditions, however, is satisfied in the Example 4.38.)

I_2. Use a numerical approach, as in Alla et al. (2015) or Wei et al. (2020).

I_3. The direct approach: Verify the convergence of v^n to a function $v \in C^1(X)$, and then show that v indeed satisfies (4.5.9). (See Example 4.38.)

Concerning (II), we can try a direct approach, as in I_3 above, or (if possible) use a general result such as Theorem B.10 or Proposition B.12 in the Appendix B. ◊

Example 4.38. Consider the scalar LQ problem in Example 4.18, where

$$V(x, a(\cdot)) = \int_0^\infty e^{-rt}[Qx^2(t) + Ra^2(t)]dt \tag{4.5.10}$$

and

$$\dot{x}(t) = \delta x(t) + \eta a(t), \quad t \geq 0, \quad x(0) = x. \tag{4.5.11}$$

Recall that Example 4.18 requires $\eta \neq 0$ and $R > 0$. Let f_0 be the linear Markov control $f_0(x) := C_0 x$ for some constant C_0. In this

case (4.5.11) becomes

$$\dot{x}(t) = D_0 x(t), \quad \text{with} \quad D_0 := \delta + \eta C_0,$$

so $x(t) = xe^{D_0 t}$ for all $t \geq 0$. Hence, with $a(t) := f_0(x(t)) = C_0 x(t)$ in (4.5.10), we see that $v^0(x) := V(x, f_0)$ is given by

$$\begin{aligned} v^0(x) &= (Q + RC_0^2)x^2 \int_0^\infty e^{-rt} e^{2D_0 t} dt \\ &= F_0 x^2 \qquad \text{with} \quad F_0 := \frac{Q + RC_0^2}{r - 2D_0} \end{aligned} \tag{4.5.12}$$

if $r - 2D_0 > 0$, which is assumed hereafter. This implies that, in particular,

$$e^{-rt} v^0(x(t)) = F_0 x^2 e^{-(r-2D_0)t} \to 0$$

as $t \to \infty$, so f_0 is indeed in the class of stationary Markov controls that satisfy (4.5.3). Moreover, v^0 is in $C^1(X)$ and it can be directly verified that (4.5.4) holds, i.e.,

$$r v^0(x) = c(x, f_0) + 2F_0 x \cdot F(x, f_0) \quad \forall x \in X.$$

The latter equality yields that, for all $x \in X$,

$$r v^0(x) \geq \inf_a [c(x, a) + 2F_0 x \cdot (\delta x + \eta a)], \tag{4.5.13}$$

and the minimum at the right-hand side is attained at

$$f_1(x) = C_1 x \quad \text{with} \quad C_1 := -F_0 \eta / R.$$

With this value of $a = f_1(x)$ in (4.5.11) we obtain

$$\dot{x}(t) = D_1 x(t) \quad \text{with} \quad D_1 := \delta + \eta C_1,$$

so $x(t) = xe^{D_1 t}$ for all $t \geq 0$. Similarly, from (4.5.10) we see that $v^1(x) := V(x, f_1)$ is given by

$$v^1(x) = F_1 x^2 \quad \text{with} \quad F_1 := \frac{Q + RC_1^2}{r - 2D_1}$$

if $r - 2D_1 > 0$.

In general, if for some $n = 0, 1, \ldots$ we are given that $f_n(x) := C_n x$ for some C_n, then

$$\dot{x}(t) = D_n x(t) \quad \text{with} \quad D_n := \delta + \eta C_n,$$

so $x(t) = xe^{D_n t}$ for all $t \geq 0$, and

$$v^n(x) := V(x, f_n) = F_n x^2, \quad \text{with} \quad F_n := \frac{Q + RC_n^2}{r - 2D_n}$$

provided that $r - 2D_n > 0$. The latter condition yields that f_n satisfies (4.5.3) and, on the other hand, one can directly verify that (4.5.4) holds with v^n and f_n. Furthermore, as in (4.5.13), we can see that, for all $x \in X$,

$$f_{n+1}(x) = C_{n+1} x \quad \text{and} \quad v^{n+1}(x) = F_{n+1} x^2$$

with

$$C_{n+1} := -F_n \eta / R, \qquad D_{n+1} := \delta + \eta C_{n+1}, \tag{4.5.14}$$

and

$$F_{n+1} := \frac{Q + RC_{n+1}^2}{r - 2D_{n+1}} \tag{4.5.15}$$

if $r - 2D_{n+1} > 0$. Now, in (4.5.15) replace C_{n+1} and D_{n+1} by their values in (4.5.14) to obtain that

$$F_{n+1} = \frac{QR + (F_n \eta)^2}{(r - 2\delta)R - 2F_n \eta^2}. \tag{4.5.16}$$

Then a direct calculation gives that $F_{n+1} \leq F_n$ for all $n = 0, 1, \ldots$; that is, the sequence $\{F_n\}$ is nonincreasing. Therefore, there exists a number $k \geq 0$ such that $F_n \downarrow k$. In fact, letting $n \to \infty$ in (4.5.16) we see that k is the same as the positive solution of the quadratic equation (4.3.13).

In other words, we conclude that the *direct approach* in I_3 above works very nicely in this example, that is, the controls f_n and the value functions v^n converge to the values f^* and v_r in Example 4.18. ◊

Remark 4.39. In Chap. 6, below, we will study *stochastic* differential equations (SDEs) of the form

$$dx(t) = F(x(t), a(t))dt + \sigma(x(t))dW(t), \tag{4.5.17}$$

which of course is an "extension" of the deterministic equation (4.5.2). An obvious question is if results for (4.5.17), such as the PIA, are applicable to the deterministic case. The answer, in general, is negative!

Indeed, one should be careful because some results for SDEs require $\sigma(x)$ to be *nonzero*. For instance, a typical condition is that $\sigma(0)$ may or may not be 0, but $|\sigma(x)| > 0$ for all $x \neq 0$. (See Eq. (6.4.19), for instance.) ♢

4.5.2 The PIA: Average Cost Problems

We will now consider the long-run average cost (AC) problem in (4.4.1)–(4.4.3). Hence, we wish to minimize over $a(\cdot) \in \mathcal{A}$ the AC defined as

$$J(x, a(\cdot)) := \limsup_{T\to\infty} \frac{1}{T} \int_0^T c(x(t), a(t))dt \tag{4.5.18}$$

subject to

$$\dot{x}(t) = F(x(t), a(t)), \quad t \geq 0, \quad x(0) = x. \tag{4.5.19}$$

Recall from Sect. 4.4 that the cost $c(x, a)$ is assumed to be nonnegative.

Given a function $h \in C^1(X)$, we will denote by $\mathcal{A}^h_{SM}$ the family of stationary Markov controls $a \in \mathcal{A}_{SM}$ that satisfy (4.4.8), i.e.,

$$h(x^a(t))/t \to 0 \quad \text{as} \quad t \to \infty, \tag{4.5.20}$$

where $x^a(\cdot)$ stands for the state process in (4.5.19) when using the control function $a(\cdot)$.

Let $L^a h(x)$ be as in (4.4.5)–(4.4.6). The PIA in the AC case hinges on the Poisson equation (4.4.7) and an AC-analogue of Lemma 4.35, which is the following.

Lemma 4.40. Given a stationary Markov control f, let us suppose that there exists a constant $j(f)$ and a function $h^f \in C^1(X)$ that satisfy the Poisson equation

$$j(f) = c(x, f) + h_x^f(x) \cdot F(x, f) \quad \forall x \in X, \tag{4.5.21}$$

where we are using the notation introduced in Sect. 4.5.1, namely, $c(x, f) := c(x, f(x))$ and $F(x, f) := F(x, f(x))$. We assume that $a(t) = f(x(t))$ satisfies (4.5.20). By (4.5.21),

$$j(f) \geq \inf_{a \in A} [c(x, a) + h_x^f(x) \cdot F(x, a)] \tag{4.5.22}$$

for all $x \in X$, and we suppose the existence of a Markov control $g \in \mathcal{A}_{SM}^{h^f}$ that attains the minimum in (4.5.22), so

$$j(f) \geq c(x, g) + h_x^f(x) \cdot F(x, g) \quad \forall x. \tag{4.5.23}$$

Then g *improves* f in the sense that the AC $j(g) \leq j(f)$, with $j(g) := J(x, g)$ for all $x \in X$.

Proof. By (4.4.6), we can express the rightmost term in (4.5.23) as

$$h_x^f(x) \cdot F(x, g) = \frac{d}{dt} h^f(x(t))|_{(x, g(x))}.$$

Therefore, integration of (4.5.23) from $t = 0$ to $t = T$ gives

$$j(f)T \geq \int_0^T c(x(t), g)dt + h^f(x(T)) - h^f(x).$$

Finally, multiply both sides of the latter inequality by $1/T$ and then let $T \to \infty$ to obtain the desired conclusion. □

We now introduce the PIA for the average cost OCP. Consider a sequence of Markov controls f_n as follows.

(PI$_1$) Given f_n, for some $n = 0, 1, \ldots$, suppose that there exists a solution $(j(f_n), h^n(\cdot)) \in \mathbb{R} \times C^1(X)$ to the Poisson equation

$$j(f_n) = c(x, f_n) + h_x^n(x) \cdot F(x, f_n),$$

where $f_n \in \mathcal{A}_{SM}^{h_n}$, so (by Lemma 4.19) $j(f_n) \equiv J(\cdot, f_n)$. It follows that

$$j(f_n) \geq \inf_{a\in A}[c(x,a) + h_x^n(x)\cdot F(x,a)]. \tag{4.5.24}$$

(PI_2)*Policy improvement.* Assume the existence of $f_{n+1} \in \mathcal{A}_{SM}^{h^n}$ such that, for every $x \in X$, $f_{n+1}(x)$ attains the minimum in (4.5.23), i.e.,

$$c(x, f_{n+1}) + h_x^n(x)\cdot F(x, f_{n+1}) = \min_{a\in A}[c(x,a) + h_x^n(x)\cdot F(x,a)], \tag{4.5.25}$$

and then find the solution $(j(f_{n+1}), h^{n+1}(\cdot))$ of the Poisson equation corresponding to f_{n+1}.

Proposition 4.41. For $n = 0, 1, \ldots$, let f_n and h^n be as in (PI_1)-(PI_2).

(a) If for some n, $j(f_n) = j(f_{n+1})$, then stop the PIA: f_n and f_{n+1} are both AC-optimal in the class $\mathcal{A}_{SM}^{h_n}$, that is,

$$j(f_n) = j(f_{n+1}) \leq J(\cdot, f) \quad \forall f \in \mathcal{A}_{SM}^{h_n}.$$

(b) Otherwise, if $j(f_n) > j(f_{n+1})$ for all $n = 0, 1, \ldots$, then there exists a number $j^* \geq 0$ such that $j(f_n) \downarrow j^*$.

Proof. (a) By Lemma 4.40 and (4.5.25),

$$\begin{aligned} j(f_{n+1}) &= \min_{a\in A}[c(x,a) + h_x^n(x)\cdot F(x,a)] \\ &\leq j(f_n) \\ &= j(f_{n+1}). \end{aligned}$$

Therefore, both f_n and f_{n+1} satisfy the ACOE (4.4.12), and so the desired result follows from Theorem 4.20(c).

Part (b) is a consequence of Lemma 4.40. □

Remark 4.42. (a) Open problems: The bad news about Proposition 4.41 is that (as in the Remark 4.37 concerning the *discounted cost*) the proposition does not guarantee the "convergence" of the PIA, that is, it does not state the convergence of the functions $h_n(\cdot)$ nor that j^* is in fact the AC-value function. Moreover, in Remark 4.37 we mentioned three options to prove the convergence of the PIA for discounted problems, but two of them are

not applicable in the AC case; namely, the option I_1 (about imposing "suitable assumptions" on the OCP) and the option I_2 (about using a numerical approach) are, to the best of our knowledge, completely unexplored in the AC case. This situation suggests some interesting research problems. On the other hand, the good news is that the option I_3 (the direct approach) in the Remark 4.37 is—at least sometimes—applicable to AC problems. See the following Example 4.43.

(b) The following comment is similar to Remark 4.39: Since there are some results on the PIA for AC problems related to *stochastic* differential equations (see Sect. 6.5) of the form

$$dx(t) = F(x(t), a(t))dt + \sigma(x(t))dW(t), \tag{4.5.26}$$

an obvious question is if one can deduce results for the *deterministic* problem (4.4.1)–(4.4.3) simply by taking $\sigma(\cdot) \equiv 0$ in (4.5.26). The answer, in general, is **no**. The reason is that (as can be seen in the references in Sect. 6.5) the AC results in the stochastic case require $|\sigma(x)| > 0$ *for all* $x \in X$, except perhaps at $x = 0$. (The latter fact is required even in the one-dimensional case. See for instance assumption (A1) in Anulova et al. (2020).) ♢

Example 4.43. Let us consider again the AC control problem in Example 4.23 with transition function and running cost given by

$$F(x, a) := p - ax^{1/2} \quad \text{and} \quad c(x, a) := x + ka^2. \tag{4.5.27}$$

The constants p and k, and the initial state $x(0) = x$ are all positive. To initialize the PIA, take the Markov control $f_0(x) = C_0 x^{1/2}$ with $C_0 > 0$. We selected this value of f_0 because then the right-hand side of F is linear in x and similarly for the cost c. More explicitly, with $a = f_0(x)$ the system equation becomes

$$\dot{x}(t) = p - C_0 x(t), \quad t \geq 0, \tag{4.5.28}$$

so $x(t) = D_0 e^{-C_0 t} + p/C_0$ with $D_0 := x(0) - p/C_0$. Therefore, by definition of $c(x, a)$ in (4.5.27), the corresponding average cost $j_0 = J(f_0)$ is given by

$$j_0 = \limsup_{T\to\infty} \int_0^T (1 + kC_0^2)x(t)dt,$$

where the integrand

$$(1 + kC_0^2)x(t) = (1 + kC_0^2)D_0 e^{-C_0 t} + (1 + kC_0^2)p/C_0.$$

Observe that, in the right-hand side, the first term $\to 0$ as $t \to \infty$. Consequently,

$$j_0 = (1 + kC_0^2)p/C_0. \tag{4.5.29}$$

We now wish to find the Poisson equation associated to f_0, i.e.,

$$j_0 = c(x, f_0) + h_0'(x) \cdot F(x, f_0) \tag{4.5.30}$$

for some function h_0, where $h_0'(x) = dh_0(x)/dx$. From the definition (4.5.27) of F and c we obtain

$$j_0 = (kC_0^2 - h_0'(x)C_0 + 1)x + h_0'(x) \cdot p.$$

This equation is satisfied if $h_0'(x) \cdot p = j_0$, or $h_0'(x) = j_0/p$, and $kC_0^2 - j_0C_0/p + 1 = 0$ or

$$pkC_0^2 - j_0C_0 + p = 0.$$

(Note that the latter equation is the same as (4.5.29).) This concludes the initialization step in the PIA.

To proceed with the policy improvement procedure, instead of (4.5.30) we now consider the inequality

$$j_0 \geq \inf_{a\geq 0}[c(x, a) + h_0'(x)F(x, a)].$$

Thus, proceeding as usual, the minimum in the right-hand side is attained at $f_1(x) = C_1 x^{1/2}$, with $C_1 = h_0'(x)/2k = j_0/(2kp)$. Hence, the system equation (4.5.28) now becomes $\dot{x}(t) = p - C_1 x(t)$ for $t \geq 0$, so

$$x(t) = D_1 e^{-C_1 t} + p/C_1 \quad \text{with} \quad D_1 := x(0) - p/C_1.$$

Continuing this process we see that the average cost $j_1 = J(f_1)$ is

$$j_1 = (1 + kC_1^2)p/C_1 \tag{4.5.31}$$

and the Poisson equation

$$j_1 = c(x, f_1) + h_1'(x) \cdot F(x, f_1)$$

is satisfied with $h_1'(x) = j_1/p$. Moreover, from (4.5.29) and (4.5.31) we see that $j_1 < j_0$.

In general, given the Markov control $f_n(x) := C_n x^{1/2}$, with $C_n > 0$, we obtain the average cost

$$j_n := J(f_n) = (1 + kC_n^2)p/C_n, \quad n = 0, 1, \ldots$$

and the associated Poisson equation holds with a function h_n such $h_n'(x) = j_n/p$, whereas C_n is the positive root of the quadratic equation

$$pkC_n^2 - j_n C_n + p = 0.$$

One can also show, as in (4.5.29) and (4.5.31), that the j_n form a monotone decreasing sequence and, in fact,

$$j_n \downarrow j^* := (1 + kC^2)p/C \tag{4.5.32}$$

where C is the positive solution of

$$pkC^2 - j^*C + p = 0. \tag{4.5.33}$$

To conclude, one can verify that the Markov controls $f_n(x)$ converge to the optimal AC control $f^*(x) = Cx^{1/2}$ and that j^* is the optimal average cost, with $C > 0$ as in (4.5.33). ◊

Exercises

4.1. Solve the following OCP: Minimize

$$J(a(\cdot)) = \int_0^1 c(x(t), a(t))dt$$

subject to $\dot{x}(t) = a(t)^2$, with $x(0) = x(1) = 0$, and control set $A = [-1, 1]$.
Hint. Note that $x(\cdot) = 0$.

4.2. Consider the OCP: minimize

$$\int_0^T e^{-rt}[x(t) + ka(t)^2]dt + e^{-rT}qx(T)$$

over all control functions $a(\cdot) \geq 0$, subject to

$$\dot{x}(t) = b - a(t)x(t)^{1/2}, \text{ for } 0 \leq t \leq T, x(0) = x_0.$$

Propose a solution of the form $v(t,x) = P(t)x + Q(t)$ for the corresponding HJB equation. Show that, in this case, the coefficients $P(\cdot)$ and $Q(\cdot)$ should satisfy that

$$\dot{P}(t) = rP(t) + P(t)^2/4k - 1, \ \dot{Q}(t) = rQ(t) - bP(t)$$

with $P(T) = q, Q(T) = 0$, and the optimal control is

$$a^*(t,x) = (2k)^{-1}v_x \cdot x^{1/2}.$$

4.3. The goal of this exercise is to prove that, under Assumption 4.2, there is a unique solution to (4.1.2) for every control $a \in \mathcal{A}$.

(a) Fix $a \in \mathcal{A}$, and let $\delta > 0$ be such $s + \delta \leq T$. Let $\mathcal{C} \equiv C([s, s+\delta], \mathbb{R}^n)$ be the complete linear space of continuous functions $x : [s, s+\delta] \to \mathbb{R}^n$, with the supremum norm

$$\|x\| := \sup\{|x(t)| : s \leq t \leq s + \delta\}.$$

For each $x \in \mathcal{C}$, define the mapping $t \to R[x](t)$, for every $t \in [s, s+\delta]$ and some $y \in \mathbb{R}^n$, as

$$R[x](t) := y + \int_s^t F(r, x(r), a(r))dr.$$

Prove that

$$\|R[x_1] - R[x_2]\| \leq L\delta\|x_1 - x_2\| \quad \forall \quad x_1, x_2 \in \mathcal{C},$$

where L is the constant in Assumption 4.2.

(b) Show that, for δ small enough, (4.1.2) has a unique solution in $[s, s+\delta]$, and explain how to extend such a solution to the whole interval $[s, T]$.

4.4. (Dockner et al. 2000, p. 44) Let $X = \mathbb{R}$ and $A = [-1, 1]$ be the state and action spaces, respectively. Suppose that we wish to maximize

$$J(t, x; a(\cdot)) := \int_t^T x(s)a(s)ds \quad \text{for} \quad 0 \le t \le T$$

over all $a(\cdot) \in \mathcal{A}[t, T]$, subject to $\dot{x}(s) = a(s)$ and initial condition $x(t) = x \in \mathbb{R}$.

(a) Let x_a be the state path associated to the control $a(\cdot)$. Verify that

$$-(T - t) \le x_a(T) - x \le T - t,$$

and $J(t, x; a(\cdot)) = (x_a(T)^2 - x^2)/2$ for all $t \le T$.

(b) Show that the control policy a^* defined, for all $s \in [t, T]$, as

$$a^*(s) = \begin{cases} 1 & \text{if} \quad x \ge 0, \\ -1 & \text{if} \quad x < 0 \end{cases}$$

is optimal.

(c) Show that the value function can be expressed as $V(t, x) = \frac{(T-t)^2}{2} + (T - t)|x|$ for $t \le T, x \in \mathbb{R}$. Note that V is not differentiable at $(t, 0)$ for $t < T$.

4.5. Let $X = A = \mathbb{R}$. Consider the OCP

$$\min_{a(\cdot)} \left[\int_t^T [1 + a(s)]^{1/2} ds + |y(T) - b| \right]$$

subject to

$$\dot{y}(s) = a(s) \quad \forall \quad t \le s \le T, \quad \text{with} \quad y(t) = y.$$

Show that $V(t, y) = [(T - t)^2 + (b - y)^2]^{1/2}$ is a solution to the corresponding HJB equation. Give a geometric interpretation.

4.6. (Managing investment income) Consider an optimal investment problem at a rate of interest $\beta > 0$, with the state space (capital) $X = [0, \infty)$, the set of feasible actions (consumptions) at the state x given by $A(x) = [0, \beta x]$, and the performance index

$$\int_0^T e^{-rt}\sqrt{a(t)}dt,$$

subject to $\dot{x}(t) = \beta x(t) - a(t)$ and $x(T) = 0$. Find a non negative function $v(t)$ such that the value function is given by $V(t,x) := e^{-rt}\sqrt{v(t)x}$.

Hint: The HJB equation (4.2.2) has an optimal action $a(t) = x(t)/v(t)$, and becomes in $\dot{v}(t) - (2r - \beta)v(t) + 1 = 0$ with boundary condition $v(T) = 0$.

4.7. (A heuristic derivation of the HJB equation.) Consider the time–invariant OCP:

$$\min_{a(\cdot)} \left[\int_0^T c(x(t), a(t))dt + C(x(T)) \right]$$

subject to

$$\dot{x}(t) = F(x(t), a(t)) \quad \forall \quad 0 \le t \le T, \quad \text{with } x(0) = x.$$

(a) For $N = 1, 2, ...$ and $t \in [0, T)$, let $\delta := \frac{T-t}{N}$. Assume that the following discrete–time *approximating problem*

$$\min \left[\sum_{k=0}^{N-1} \delta c(x_k, a_k) + C(x_N) \right]$$

subject to

$$x_{k+1} = x_k + \delta F(x_k, a_k) \quad \forall \quad k = 0, 1, ..., N-1, \quad \text{and} \quad x_0 = x$$

has a solution. Show that there is a function J that satisfies

$$J(t,x) = \min_a \left[\delta c(x,a) + J(t+\delta, x + \delta F(x,a)) \right].$$

Hint. Fix $t < T$. Let $J(t + N\delta, x) := C(x)$, and then go backwards.

(b) Assume that J is a well–defined C^1 function in a neighborhood of (t,x). Consider the Taylor's expansion

$$J(t+\delta, x + \delta F(x,a)) = J(t,x) + J_t(t,x) \cdot \delta + J_x(t,x) \cdot \delta F(x,a) + o(\delta),$$

where $o(\delta)/\delta \to 0$ as $\delta \to 0$. Show that

$$\min_a [c(x,a) + J_t(t,x) + J_x(t,x) \cdot F(x,a)] = 0.$$

4.8. (Cruz-Suárez and Montes-de Oca (2008)) Let $X \subset \mathbb{R}^n$ and, for each $x \in X$, let $A(x) \subset \mathbb{R}^m$ be a set with nonempty interior. Moreover, let $K = \{(x,a) \mid x \in X,\ a \in A(x)\}$, and $c : K \to \mathbb{R}$. The purpose of this exercise is to provide some sufficient conditions for the differentiability of the function

$$w(x) := \inf\{c(x,a) \mid a \in A(x)\}.$$

Assume c is of class C^2 in the interior of K and, for each x, the Hessian matrix $c_{aa}(x,a)$ is nonsingular for every a. Further, suppose there is a function $f : X \to \mathbb{R}^m$ such that

$$w(x) = c(x, f(x))$$

and $f(x)$ is an interior point of $A(x)$ for each $x \in X$. Justify the equality

$$c_a(x, f(x)) = 0.$$

Use the Implicit Function Theorem (see for instance Theorem 9.28 in Rudin (1976)) to show that f is differentiable and find $f_x(x)$. Show also that

$$w_x(x) = c_x(x, f(x)) \qquad x \in X. \tag{4.5.34}$$

4.9. Consider the dynamics (4.0.1) and the cost functional

$$\int_0^T c(t, x(t), a(t))dt + C(T, x(T)) \tag{4.5.35}$$

which is said to be in the *Bolza form*. The so-called *Lagrange form* of the functional happens when $C \equiv 0$, whereas $c \equiv 0$ corresponds to the *Mayer form*. Suppose that Assumption 4.2 holds with C as a function of (t,x).

(a) Show that any cost C in the Mayer form can be put in the Lagrange form.

Hint. $C(T,x(T)) - C(0,x(0)) = \int_0^T [\frac{d}{dt} C(t,x(t))]dt$.

(b) Consider an additional state x_{n+1} to the system (4.0.1) given by

$$\dot{x}_{n+1}(t) = c(t,x(t),a(t)), \qquad x_{n+1}(0) = 0.$$

Prove that the Bolza form (4.5.35) can be put in the Mayer form.

Answer. $x_{n+1}(T) + C(T,x(T))$ where $x = (x_1,\dots,x_n)$.

4.10. Let a^*, x^*, and λ satisfy the Minimum Principle for the OCP with cost (4.2.1) and dynamics (4.0.1). Let $\mu(t) := e^{rt}\lambda(t)$ and

$$\overline{H}(t,x,a,\mu) = c(t,x,a) + \mu \cdot F(t,x,a), \qquad t \in [0,T].$$

Show that the minimum condition and the adjoint equation can be respectively written as

$$\overline{H}(t,x^*(t),a^*(t),\mu(t)) = \min_{a\in A} \overline{H}(t,x^*(t),a,\mu(t))$$

and

$$-\dot{\mu}(t) + r\mu(t) = \overline{H}(t,x^*(t),a^*(t),\mu(t)), \qquad \mu(T) = C_x(x^*(T))$$

for $0 \le t < T$.

4.11. Prove Proposition 4.30.

4.12. Consider the r-discount OCP (4.5.1)–(4.5.2) and let $f \in \mathcal{A}^*_{SM}$ be a stationary Markov control that satisfies (4.5.3).

(a) Assuming that $v(\cdot) := V(\cdot, f(\cdot))$ is in $C^1(X)$, show that v is the unique solution of the equation

$$rv(x) = c(x,f) + L^{f(x)}v(x) \quad \forall x \in X.$$

(b) Show that the r-discount value function $v^*(x) := \inf_{a\in\mathcal{A}} V(x,a(\cdot))$ is the unique solution in $C^1(X)$ of the HJB equation

$$rv^*(x) = \inf_{a\in A}[c(x,a) + L^a v^*(x)].$$

Hint. In both cases (a) and (b) consider a function of the form $u(t,x) = e^{-rt}v(x)$ as in the proof of Lemma 4.35.

4.13. (The Ramsey model). Consider the maximization problem of the discounted utility functional

$$V(k_0, c(\cdot)) = \int_0^\infty u(c(t))e^{-\rho t}dt$$

over all consumption strategies $c(\cdot)\geq 0$ subject to $\dot{k}(t)=F(k(t))-c(t)$, $k(0) = k_0$, where the utility and system functions are given, respectively, by

$$u(c) = \frac{c^{1-\sigma}}{1-\sigma} \quad \text{and} \quad F(k) = Ak^\alpha, \quad \alpha, A > 0, 0 < \sigma < 1.$$

For the particular case $\alpha = \sigma$, solve the HJB equation by conjecturing the value function as $v(k) = B_0 + B_1 k^{1-\sigma}$, where B_0 and B_1 are undetermined coefficients.

Answer.

$$v(k) = \left(\frac{\sigma}{\rho}\right)^\sigma \left(\frac{A}{\rho} + \frac{1}{1-\sigma}k^{1-\sigma}\right), \quad c^*(t) = \frac{\rho}{\sigma}k^*(t),$$

$$k^*(t) = \left[\frac{A\sigma}{\rho} + \left(k_0^{1-\sigma} - \frac{A\sigma}{\rho}\right)e^{-(1-\sigma)\frac{\rho}{\sigma}t}\right]^{\frac{1}{1-\sigma}}.$$

4.14. Let $(j, h(\cdot))$ be a solution to the Poisson equation (4.4.7)–(4.4.8). Prove that j is unique, and h is unique up to additive constants.
Hint. Recall (4.4.6) and (4.4.10).

Chapter 5

Continuous–Time Markov Control Processes

As noted in Remark 4.7(b), the solution $x(\cdot)$ of the (deterministic) ordinary differential equation (4.0.1) can be interpreted as a *Markov control process* (MCP), also known as a *controlled Markov process*. In this chapter we introduce some facts on general continuous–time MCPs, which allows us to make a unified presentation of related control problems. We will begin below with some comments on (noncontrolled) continuous–time Markov processes. (We only wish to motivate some concepts, so our presentation is not very precise. For further details, see the bibliographical notes at the end of this chapter.)

For notational convenience, sometimes we write $x(t)$ as x_t.

5.1 Markov Processes

Consider a continuous–time stochastic process $\mathcal{X} = \{x(t) : t \geq 0\}$ with values in a set $X \subset \mathbb{R}^n$ for some positive integer n. (In most applications, X is an open set; perhaps $\mathbb{R}^n$ itself. There are, however, other important cases. For instance, if $\mathcal{X}$ is a so–called *jump process*, then X is usually a countable set.) Accordingly, there is a probability space $(\Omega, \mathcal{F}, P)$ such that, for each $t \geq 0$, $x(t)$ is a random variable (or measurable function) from Ω to $\mathbb{R}^n$. Hence, strictly speaking, we have a function of two variables $(t, \omega) \mapsto x(t)(\omega) \equiv x(t, \omega)$.

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*, Texts in Applied Mathematics 76,
https://doi.org/10.1007/978-3-031-21139-3_5

Summarizing, for each $t \geq 0$, we have a random variable $\omega \mapsto x(t, \omega)$ on Ω; and, for each $\omega \in \Omega$, we have a function $t \mapsto x(t, \omega)$, for $t \geq 0$, which is called a *trajectory* or *sample path* of $\mathcal{X}$.

By a standard convention, the variable ω is omitted and we write $x(t)$ rather than $x(t, \omega)$.

The continuous–time process $\mathcal{X}$ is said to satisfy the *Markov property* if, informally, given the "present" state, the future behavior of the process is independent of its past. To be a little more precise, fix an arbitrary time $s \geq 0$, and let $x(s)$ be the "present" state. Then we can express the Markov property as follows: for any "future time" $t > s$ and $B \subset X$,

$$P[x(t) \in B | x(r) \; \forall \; r \leq s] = P[x(t) \in B | x(s)]. \tag{5.1.1}$$

In words, (5.1.1) states that the distribution (or "behavior") of the process at any "future" state $x(t)$, for $t > s$, given the "past history" $\{x(r), r \leq s\}$, depends only on the present state $x(s)$.

From the right–hand side of (5.1.1) we obtain the *transition probabilities*

$$P(s, x, t, B) := P[x(t) \in B | x(s) = x] \tag{5.1.2}$$

for every $0 \leq s \leq t, x \in X$, and $B \subset X$. If $t = s$, then (5.1.2) becomes the Dirac (or unit) measure concentrated at $x(s) = x$, i.e.,

$$P(s, x, s, B) = \delta_x(B),$$

which is defined as $\delta_x(B) := 1$ if $x \in B$, and $:= 0$ if $x \notin B$. Alternatively, $\delta_x(B) = I_B(x)$ where I_B denotes the *indicator function* of the set B, $I_B(x) := 1$ if $x \in B$, and $:= 0$ if $x \notin B$.

Remark 5.1. The transition probabilities are called *stationary* or *time–homogeneous* if they depend only on the time difference $t - s$, that is,

$$P(s, x, t, B) \equiv P(t - s, x, B).$$

In this case, (5.1.2) becomes

$$P(t, x, B) := P[x(t) \in B | x(0) = x] \text{ for } t \geq 0,$$

and the Markov process itself is said to be stationary or time–homogeneous. ◊

Example 5.2. *A deterministic system.* Consider the ordinary differential equation

$$\dot{x}(t) = F(t, x(t)) \quad \text{for } t \geq 0, \tag{5.1.3}$$

with a given initial condition $x(0) = x_0$. Let us suppose that F is continuous and has continuous first partial derivatives with respect to the components of $x \in X$, where $X \subset \mathbb{R}^n$. In this case, (5.1.3) has a unique solution

$$x(t) = x_0 + \int_0^t F(r, x(r))dr \quad \forall\, t \geq 0,$$

which can be expressed as

$$x(t) = x(s) + \int_s^t F(r, x(r))dr \quad \forall\, 0 \leq s \leq t. \tag{5.1.4}$$

If we interpret $x(s)$ as the "present" state, and $x(t)$, for $t \geq s$, as the "future", then it follows that (5.1.4) is the "deterministic version" of the Markov property (5.1.1).

Observe that the *deterministic function* $t \mapsto x(t)$ in (5.1.3) or (5.1.4) can be seen as a "degenerate" stochastic process in the sense that, for each $t \geq 0$, we can interpret $x(t)$ as a *constant* random variable, that is, $\omega \to x(t, \omega) \equiv x(t)$. Consequently, the transition probability in (5.1.2) is a Dirac measure

$$P(s, x, t, B) = \delta_{x(t;s,x)}(B), \tag{5.1.5}$$

where $x(t; s, x)$ is given by (5.1.4) when the "initial condition" is $x(s) = x$.

For future reference note that integration with respect to this Dirac measure (5.1.5) yields

$$\int_X P(s, x, t, dy)v(y) = v[x(t; s, x)] \tag{5.1.6}$$

for any bounded measurable function v on X. ◊

Example 5.3 (Wiener process.). A Wiener process, also known as a *Brownian motion*, is a real–valued stochastic process $\{w(t), t \geq 0\}$ that plays an important role in pure and applied mathematics, and also in physics, astronomy, economics, mathematical finance, and many other fields. It satisfies that $w(0) = 0$ and, furthermore, by definition,

(a) it has *independent increments*, which means that if $t_0 = 0 < t_1 < \cdots < t_m$, then the "increments"

$$w(t_1) - w(t_0), w(t_2) - w(t_1), ..., w(t_m) - w(t_{m-1})$$

are independent random variables; and

(b) it has Gaussian *stationary increments*, that is, for any $t \geq 0$ and $h > 0$, the distribution of the increment $w(t + h) - w(t)$ is Gaussian (or normal) with 0 mean and variance h. (This increment is "stationary" in the sense that its distribution depends on the "time increment" h only, not on t.)

Remark. A continuous–time stochastic process with independent increments is Markov. (See Ash and Gardner (1975), Theorem 4.6.5) ♢

Hence, by this remark, the Wiener process $w(\cdot)$ is Markov. Moreover, from (a) and (b) above, for any $t > 0$ and initial state $w(0) = x$, the stationary transition probability is

$$P(t, x, B) = \int_B n_{x,t}(y)dy, \tag{5.1.7}$$

where $n_{x,t}(\cdot)$ denotes the Gaussian (or normal) density with mean x and variance t, i.e.,

$$n_{x,t}(y) = (2\pi t)^{-1/2} \exp(-|y - x|^2/2t) \quad \forall\, y \in \mathbb{R}.$$

Among the many properties of a Wiener process it is the fact that it has continuous sample paths $t \mapsto w(t)$ that are nowhere differentiable!

A process $(w_1(t), \ldots, w_n(t)) \in \mathbb{R}^n$, for $t \geq 0$, is called an n–dimensional Wiener process (or Brownian motion) if $w_1, ..., w_n$ are independent 1–dimensional Wiener processes. ♢

Example 5.4. Stochastic differential equations. In Chap. 6, below, we consider n–dimensional stochastic differential equations (SDEs) of the form

$$dx(t) = b(t, x(t))dt + \sigma(t, x(t))dw(t), \ t \geq 0, \tag{5.1.8}$$

with a given initial condition $x(0) = x_0$, where $w(\cdot)$ is a Wiener process. The functions b and σ in (5.1.8) are called the SDE's *drift coefficient* and the *diffusion coefficient*, respectively. As noted in Example 5.3, the sample paths $t \mapsto w(t)$ are not differentiable. Hence, strictly speaking we should express (5.1.8), for any $0 \leq s \leq t$, in the integral form

$$x(t) = x(s) + \int_s^t b(r, x(r))dr + \int_s^t \sigma(r, x(r))dw(r), \tag{5.1.9}$$

where the second integral in the right–hand side is well defined as a so–called *Itô integral*.

For our present purposes, it suffices to note that, under suitable conditions (see Assumption 6.1), the solution of (5.1.8) is a Markov process with transition probabilities

$$P(s, x, t, B) = P[x(t) \in B | x(s) = x] = P[x(t; s, x) \in B]$$

for all $0 \leq s \leq t$, $x \in \mathbb{R}^n$, and $B \subset \mathbb{R}^n$, where $x(t; s, x)$ is given by (5.1.9) when the initial condition is $x(s) = s$. Moreover, with some additional mild condition (Assumption 6.2) the solutions of SDEs form a class of so–called *diffusion processes*, and, therefore, by an abuse of terminology, sometimes one uses the latter term, diffusion processes, to refer to the SDEs (5.1.8)–(5.1.9).

If the coefficients $b(t, x) \equiv b(x)$ and $\sigma(t, x) \equiv \sigma(x)$ in (5.1.8) do not depend on the time parameter, then the Markov process $x(\cdot)$ is time–homogeneous with transition probabilities

$$P(t, x, B) = P[x(t) \in B | x(0) = x].$$

Finally, observe that an ordinary differential equation as in (5.1.3) can be seen, of course, as a special case of a SDE with diffusion coefficient $\sigma(\cdot) \equiv 0$.

$\Diamond$

We conclude this section with another example of a continuous-time Markov process, namely, the *Poisson process*, which is very useful in some applications, for instance, in the control of queues and other systems in which the state space is a *denumerable* (or countable) set. First, we recall the following definition from elementary probability.

A random variable N with values in the set of nonnegative integers $\mathbb{N} = \{0, 1, ...\}$ is said to have a *Poisson distribution* with rate $\lambda > 0$ if

$$P(N = k) := \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k = 0, 1, \ldots.$$

On the other hand, a continuous-time stochastic process $\{N(t), t \geq 0\}$ with values in $\mathbb{N}$ is called a *counting process* if, for any $t \geq 0$ and $h > 0$, the increment $N(t + h) - N(t)$ equals the number of "events" that have occurred in the interval $(t, t + h]$. (Compare the following definition with Example 5.3 above.)

Definition 5.5. A counting process $\{N(t),\ t \geq 0\}$ with $N(0) = 0$ is called a *Poisson process* with rate $\lambda > 0$ if it has

(a) *independent increments*, and
(b) *Poisson stationary increments* with rate λ, that is, for each $t \geq 0$ and $h > 0$, the increment $N(t + h) - N(t)$ is a Poisson random variable with parameter λh, i.e.,

$$P(N(t + h) - N(t) = k) = \frac{e^{-\lambda h}(\lambda h)^k}{k!}$$

for $k = 0, 1, ...$

By property (a), a Poisson process is a Markov process. (See the Remark in Example 5.3)

5.2 The Infinitesimal Generator

As in the previous section, we consider a continuous–time Markov process $\mathcal{X} = \{x(t) : t \geq 0\}$ in $X \subset \mathbb{R}^n$, with transition probabilities (5.1.2). In this section we introduce the *infinitesimal generator*

(or simply the *generator*) of the Markov process $\mathcal{X}$, which is a key tool to study different aspects of $\mathcal{X}$.

An important property of the transition probabilities is expressed by the *Chapman–Kolmogorov equation*:

$$P(s,x,r,B) = \int_X P(s,x,t,dy)P(t,y,r,B) \tag{5.2.1}$$

for $0 \le s \le t \le r$.

We will now introduce three families $M \supset M_0 \supset \mathcal{D}$ of real–valued measurable functions on $X_\infty := [0,\infty) \times X$.

Definition 5.6. Let M be the linear space of real–valued measurable functions v on X_∞ such that

$$\int_X P(s,x,t,dy)|v(t,y)| < \infty$$

for each $0 \le s \le t$, $x \in X$.

For each $t \ge 0$, and $v \in M$, let $T_t v$ be the function such that, for each $(s,x) \in X_\infty$, the expected value

$$T_t v(s,x) := E_{s,x}[v(s+t,x(s+t))] = \int_X P(s,x,s+t,dy)v(s+t,y) \tag{5.2.2}$$

is well defined (and finite), where $E_{s,x}[\cdots]$ denotes the conditional expectation given the initial condition $x(s) = x$. The operators $T_t, t \ge 0$, form a *semigroup of operators* on M, that is, T_0 =Identity, each T_t maps M into itself, and

$$T_{t+r} = T_t T_r \quad \forall\, t,r \ge 0,$$

where the latter equality follows from the right-hand side of (5.2.2) and the Chapman–Kolmogorov equation (5.2.1). (See Exercise 5.1.)

Let M_0 be the subfamily of M consisting of those functions $v \in M$ such that:

(a) $\lim_{t \downarrow 0} T_t v(s,x) = v(s,x)$ for every $(s,x) \in X_\infty$, and
(b) there exists $t_0 > 0$ and $u \in M$ such that

$$T_t|v|(s,x) \le u(s,x) \quad \forall\ (s,x) \in X_\infty, 0 \le t \le t_0.$$

The next definition introduces the *infinitesimal generator* of the semigroup T_t.

Definition 5.7. Let $\mathcal{D}(L)$ be the subset of functions $v \in M_0$ that satisfy the following conditions:

(a) The limit

$$Lv(s,x) := \lim_{t \downarrow 0} t^{-1}[T_t v(s,x) - v(s,x)] \tag{5.2.3}$$

exists for all $(s,x) \in X_\infty$, and

(b) Lv is in M_0.

The operator L in (5.2.3) is called the *infinitesimal generator* of the semigroup T_t, and is also known as the infinitesimal generator of the Markov process $\mathcal{X}$. The set $\mathcal{D}(L)$ is called the *domain of* L.

Example 5.8. (a) Consider the deterministic system in (5.1.3), and suppose that $(t,x) \mapsto v(t,x)$ is a real–valued mapping on X_∞ such that (for instance) it is continuously differentiable in x with bounded derivatives, uniformly in $t \ge 0$. Then, from (5.1.6) and (5.2.2),

$$T_t v(s,x) = v(s+t, x(s+t;s,x))$$

and (5.2.3) becomes

$$Lv(s,x) = v_s(s,x) + v_x(s,x)F(s,x), \tag{5.2.4}$$

where v_s denotes the partial derivative of v with respect to s, and v_x is the gradient of v (in the x–variables), that is, the row vector of partial derivatives $v_{x_1}, ..., v_{x_n}$. Hence, more explicitly, we can express (5.2.4) as

$$Lv(s,x) = v_s(s,x) + \sum_{i=1}^{n} F_i(s,x) v_{x_i}(s,x).$$

Compare the expressions (5.2.4) and (4.1.12). Omitting the control variable $a \in A$ in (4.1.12), these expressions are essentially the same. See also Remark 4.7(d) or (4.4.5)–(4.4.6) for the time-homogeneous deterministic case.

(b) Let $w(\cdot)$ be the 1–dimensional Wiener process in Example 5.3, and let $x \mapsto v(x)$ be a twice continuously differentiable function. Since $w(\cdot)$ is a time–homogeneous Markov process with transition probability (5.1.8), suitable calculations yield that (5.2.3) becomes

$$Lv(x) = (1/2)v_{xx}(x),$$

where v_{xx} denotes the second derivative of v with respect to x. In the n–dimensional case $w = (w_1, ..., w_n)$,

$$Lv(x) = \frac{1}{2}\sum_{i=1}^{n} v_{x_i x_i}(x),$$

where $v_{x_i x_i}$ denotes the second partial derivative of v with respect to x_i.

For the SDE in Example 5.4, the corresponding generator Lv is given in (6.1.4), below. ◊

The infinitesimal generator L is in fact an extension of the "weak infinitesimal generator" of a semigroup, defined in Chap. 1 of Dynkin (1965), and it has essentially the same properties. For instance, straightforward modifications of the proofs in the latter reference give the following results.

Lemma 5.9. If $v \in \mathcal{D}(L)$, then, for all $(s, x) \in X_\infty := [0, \infty) \times X$,

(a) $\frac{d^+}{dt}T_t v := \lim_{h \downarrow 0} h^{-1}[T_{t+h}v - T_t v] = T_t Lv$;

(b) $T_t v(s, x) - v(s, x) = \int_0^t T_r(Lv)(s, x)dr$.

Moreover, if $\rho \geq 0$ and $v_\rho(s, x) := e^{-\rho s}v(s, x)$, then v_ρ is also in $\mathcal{D}(L)$ and

(c) $Lv_\rho(s, x) = e^{-\rho s}[Lv(s, x) - \rho v(s, x)]$.

(d) v is a constant if, and only if, $Lv(s, x) = 0$ for all (s, x).

The proof of Lemma 5.9 is left to the reader. (See Exercise 5.2.)

Remark 5.10. (a) The expression in Lemma 5.9(b) is called *Dynkin's formula* and using (5.2.2) can be rewritten as

$$E_{s,x}v(s+t,x(s+t)) - v(s,x) = E_{s,x}\int_s^{s+t} Lv(r,x(r))dr \tag{5.2.5}$$

for $v \in \mathcal{D}(L)$. In the *deterministic* case, Dynkin's formula can be obtained from Remark 4.7(c):

$$v(T,x(T)) - v(s,x) = \int_s^T Lv(r,x(r))dr.$$

(b) Under suitable assumptions, (5.2.5) holds if t is replaced by a (random) *stopping time* τ. (See, for instance, the "corollary" in Dynkin (1965), p. 133.)

$\Diamond$

The proof of the following proposition illustrates the use of Dynkin's formula (5.2.5).

Proposition 5.11. Let c and K be nonnegative functions on $X_T := [0,T] \times X$, for some $T > 0$. Suppose that c is in M_0, and let $\rho \geq 0$.

(a) If $v \in \mathcal{D}(L)$ satisfies the equation

$$\rho v(s,x) = c(s,x) + Lv(s,x) \quad \forall\ (s,x) \in X_T \tag{5.2.6}$$

with the "terminal" condition

$$v(T,x) = K(T,x) \quad \forall\ x \in X, \tag{5.2.7}$$

then, for all $(s,x) \in X_T$,

$$v(s,x) = E_{s,x}[\int_s^T e^{-\rho(t-s)}c(t,x(t))dt + e^{-\rho(T-s)}K(T,x(T))]. \tag{5.2.8}$$

(a') If instead of (5.2.6) we have the *inequality* $\rho v \leq c + Lv$, then in (5.2.8) we replace "=" with "$\leq$". Similarly, if in (5.2.6) we replace the equality with "$\geq$", then in (5.2.8) we replace "=" with "$\geq$".

(b) Suppose that c is as above, but $\rho > 0$ and $K \equiv 0$. Suppose that $v \in \mathcal{D}(L)$ satisfies (5.2.6) for all $(s,x) \in X_\infty := [0,\infty) \times X$ and the condition

$$e^{-\rho t}E_{s,x}v(s+t,x(s+t)) = e^{-\rho t}T_t v(s,x) \to 0 \quad \text{as} \quad t \to \infty. \tag{5.2.9}$$

Then $v \equiv v^\rho$ is given by

$$\begin{aligned} v^\rho(s,x) &= E_{s,x}\int_s^\infty e^{-\rho(t-s)}c(t,x(t))dt \\ &= \int_0^\infty e^{-\rho t}T_t c(s,x)dt. \end{aligned} \tag{5.2.10}$$

(b') If in (5.2.6) we replace the equality "=" with either "$\leq$" or "$\geq$", then in (5.2.10) we replace the equality with "$\leq$" or "$\geq$", respectively.

Proof.

(a) As in Lemma 5.9(c), let $v_\rho(s,x) := e^{-\rho s}v(s,x)$ and note that (5.2.6) can be expressed as $Lv(s,x) - \rho v(s,x) = -c(s,c)$. Hence, Lemma 5.9(c) yields

$$Lv_\rho(s,x) = e^{-\rho s}[Lv(s,x) - \rho v(s,x)] = -e^{-\rho s}c(s,x). \tag{5.2.11}$$

Therefore, applying Dynkin's formula (5.2.5) to v_ρ, we obtain

$$e^{-\rho(s+t)}v(s+t,x(s+t)) - e^{-\rho s}v(s,x) = -\int_s^{s+t} e^{-\rho r}c(r,x(r))dr.$$

Multiplying both sides of the latter expression by $e^{\rho s}$ and taking $t = T - s$ it follows that

$$e^{-\rho(T-s)}E_{s,x}v(T,x(T)) - v(s,x) = -\int_s^T e^{-\rho(r-s)}c(r,x(r))dr.$$

Finally, rearranging terms and using (5.2.7) we obtain (5.2.8).

(a') If $\rho v \leq c + Lv$, then instead of (5.2.11) we obtain $Lv_\rho(s,x) \geq -e^{-\rho s}c(s,x)$. Thus, with the obvious changes, the proof of (a) gives also (a').

(b) If $K \equiv 0$, (5.2.9) and (5.2.8) give (5.2.10) as $T \to \infty$. The proof of (b') is left as an exercise. □

Proposition 5.11 will be very useful to obtain the dynamic programming (or Hamilton–Jacobi–Bellman) equation associated to some stochastic control problems.

If the function $c \in M_0$ in Proposition 5.11 is a "cost rate", that is, a cost per unit time, then $v(s,x)$ in (5.2.8) can be interpreted as a *total expected cost* during the time interval $[s,T]$, with initial state $x(s) = x \in X$ and terminal cost $K(T, x(T))$. Similarly, (5.2.10) can be seen as an *infinite–horizon expected discounted cost* from time s onward, with *discount factor* $\rho > 0$ and initial condition $x(s) = x$. In the following Proposition 5.13 we present a result for the *long–run expected average cost* defined as follows.

First, given a "cost rate" $c \in M_0$ and $t > 0$, let

$$\begin{aligned} v_t(s,x) &:= E_{s,x}\left[\int_s^{s+t} c(r, x(r))dr\right] \\ &= \int_0^t T_r c(s,x)dr \end{aligned} \tag{5.2.12}$$

be the total expected cost in the interval $[s, s+t]$, given the initial condition $x(s) = x \in X$ at time $s \geq 0$. Then

$$J_t(s,x) := \frac{v_t(s,x)}{t} \tag{5.2.13}$$

denotes the *expected average cost* during the time interval $[s, s+t]$, with initial condition $x(s) = x$. To define the "long–run expected average cost" we would like to take the limit as $t \to \infty$ in (5.2.13). A priori, however, we do not know if such a limit exists; hence, we can take instead either the "lim sup", i.e.,

$$J^{\sup}(s,x) := \limsup_{t\to\infty} J_t(s,x), \tag{5.2.14}$$

or the "lim inf",

$$J_{\inf}(s,x) := \liminf_{t\to\infty} J_t(s,x). \tag{5.2.15}$$

For theoretical reasons, taking the *lim sup* is more convenient, so we will take (5.2.14) as the **definition** of the **long–run expected average cost**. (As an example, taking the average cost as $J^{\sup}$ in

(5.2.14) simplifies the calculations to obtain results as in Exercise 5.8(b),(c).)

Remark 5.12. (a) As a rule of thumb, if we wish to *minimize* a long–run expected average *cost*, we use the lim sup as in (5.2.14). (We thus take a "conservative" or "minimax" attitude—we wish to minimize a "maximum" or "lim sup".) Nevertheless, if we wish to *maximize* a long–run average *reward* (or utility or income), then we use the lim inf in (5.2.15). (This is again a "conservative", in fact, "maximin" attitude.)

(b) Since the long-run expected average cost (5.2.14) concerns the convergence of time averages (5.2.13), it is also known as an *ergodic cost*. (See, for instance, Arapostathis et al. (2012) or Arisawa (1997).) ♢

The following Proposition 5.13 is a general version of Lemma 4.19 for deterministic continuous-time systems.

Proposition 5.13. Let $c \in M_0$ be a given function, and suppose that there exists a number $j(c) \in \mathbb{R}$ and a function $h_c \in \mathcal{D}(L)$ such the pair $(j(c), h_c)$ satisfies, for all $(s,x) \in X_\infty := [0,\infty) \times X$, the so–called *Poisson equation*

$$j(c) = c(s,x) + Lh_c(s,x). \tag{5.2.16}$$

Moreover, suppose that, for all $(s,x) \in X_\infty$, h_c is such that

$$\lim_{t\to\infty} T_t h_c(s,x)/t = 0. \tag{5.2.17}$$

Then:
(a) The constant $j(c) = J^{\sup}(s,x)$ for all $(s,x) \in X_\infty$.
(b) If the equality in (5.2.16) is replaced with "$\leq$", then the equality in (a) is replaced with "$\leq$"; that is, if

$$j(c) \leq c(s,x) + Lh_c(s,x) \quad \forall\ (s,x),$$

then $j(c) \leq J^{\sup}(s,x)$ for all (s,x). This result is also true if we have "$\geq$" in lieu of "$\leq$".

Proof. (a) By Dynkin's formula in Lemma 5.9(b) (or (5.2.5)),

$$T_t h_c(s,x) - h_c(s,x) = \int_0^t T_r(Lh_c)(s,x)dr$$
$$= E_{s,x}\int_0^t Lh_c(s+r, x(s+r))dr$$

[by (5.2.17)]

$$= tj(c) - E_{s,x}\int_0^t c(s+r, x(s+r))dr$$
$$= tj(c) - \int_0^t T_r c(s,x)dr.$$

Hence, rearranging terms and multiplying by t^{-1},

$$j(c) = t^{-1}\int_0^t T_r c(s,x)dr + t^{-1}[T_t h_c(s,x) - h_c(s,x)].$$

Finally, letting $t \to \infty$, (5.2.17) yields (a).

The proof of part (b) is similar. □

Remark 5.14. (a) The pair $(j(c), h_c)$ in Proposition 5.13 is called a *canonical pair* or a *solution* of the Poisson equation (5.2.16) corresponding to the function $c \in M_0$. There are several approaches to obtain such a pair. Some of these approaches are briefly introduced in part (d) below, in Sect. 5.5, and also in the exercise section. See also Sect. 4.4 for the *deterministic* case.

(b) If $(j(c), h_c)$ is a solution to (5.2.16), then $j(c)$ is unique, but h_c is unique up to additive constants only. More precisely, suppose that, for $i = 1, 2$, the pair $(j^i, h^i) \in \mathbb{R} \times \mathcal{D}(L)$ is a solution to (5.2.16), i.e.,

$$j^i = c(s,x) + Lh^i(s,x) \quad \forall\ (s,x),$$

and it satisfies (5.2.17). Then $j^1 = j^2$, and h^1, h^2 differ by a constant: $h^1 = h^2 +$ constant. (See Exercise 5.6.)

(c) For a stationary (or time–homogeneous) Markov process, the notation and results in this section simplify in the obvious manner. For instance, the linear space M in Definition 5.6

becomes the space of measurable functions $v : X \to \mathbb{R}$ such that

$$\int_X P(t,x,dy)|v(y)| < \infty \quad \forall\, t \geq 0, x \in X,$$

and (5.2.2) becomes

$$T_t v(x) = E_x v(x(t)) = \int_X P(t,x,dy)v(y) \quad \forall\, t \geq 0, x \in X.$$

(d) As in part (c), above, consider a time–homogeneous Markov process $\mathcal{X}$ with transition probabilities $P(t,x,\cdot)$, and a function $c \in M_0$. Let μ be a probability measure on X, which is an *invariant probability measure*[1] for $\mathcal{X}$; that is, for each Borel set $B \subset X$, $t \geq 0$, and $x \in X$, we have

$$\mu(B) = \int_X P(t,x,B)\mu(dx).$$

Let $\|\cdot\|^*$ be a norm on the linear space of finite signed measures on X. We assume that our Markov process is *uniformly ergodic* (or *geometrically ergodic*) with respect to the norm $\|\cdot\|^*$, that is, there exist positive constants θ and γ such that, for all $t \geq 0$ and $x \in X$,

$$\|P(t,x,\cdot) - \mu(\cdot)\|^* \leq \theta e^{-\gamma t}. \tag{5.2.18}$$

In this case, as $t \to \infty$, $P(t,x,\cdot)$ converges geometrically fast to $\mu(\cdot)$ for any initial state x. (For conditions ensuring geometric ergodicity, see the references in Exercise 5.9.) Finally, suppose that $c \in M_0$ is bounded[2] by some constant $\bar{c}$: $|c(x)| \leq \bar{c}$. Let

$$j(c) := \int_X c(x)\mu(dx), \quad \text{and} \quad h_c(x) := \int_0^\infty [T_t c(x) - j(c)]dt \tag{5.2.19}$$

[1] The probability measure μ is said to be "invariant" or a "stationary measure" for $\mathcal{X}$ because if the initial state $x(0)$ has distribution μ, then the state $x(t)$ has distribution μ *for all* $t \geq 0$. A Markov process is called *ergodic* if it has a unique invariant probability measure.

[2] The requirement that c is bounded simplifies the presentation, but it is *not* necessary. (See the references in Exercise 5.9.)

for $x \in X$. Then h_c is in $\mathcal{D}(L)$, and the pair $(j(c), h_c)$ is a solution to the Poisson equation

$$j(c) = c(x) + Lh_c(x) \quad \text{for all} \quad x \in X. \tag{5.2.20}$$

(See Exercise 5.9.) ◊

5.3 Markov Control Processes

For our present purposes, a continuous–time *Markov control process* (MCP) is specified by:

(a) Two sets $X \subset \mathbb{R}^n$ and $A \subset \mathbb{R}^m$, called the *state space* and the *action set* (or *action space*), respectively;

(b) *The law of motion*: Corresponding to each action $a \in A$, there exists a linear operator L^a that is the infinitesimal generator of a X–valued Markov process with transition probabilities

$$P^a(s, x, t, B).$$

(c) A *cost rate function* $c(s, x, a)$, which is a real–valued measurable function defined on $[0, \infty) \times X \times A$. We assume that c is nonnegative.

The quadruple (X, A, L^a, c) in (a), (b), (c) expresses in a compact form a continuous–time MCP. (In a more general context, X and A can be complete and separable metric spaces, also known as *Polish spaces*. Moreover, the condition that the cost rate c is nonnegative simplifies some theoretical and computational aspects, but strictly speaking it is not necessary.)

Example 5.15. The (deterministic) differential system (4.0.1) defines a continuous–time MCP with state and action spaces X and A as in Chap. 4, and infinitesimal generator L^a in (4.1.12), that is,

$$L^a v(s, x) := v_s(s, x) + v_x(s, x) \cdot F(s, x, a). \tag{5.3.1}$$

The situation is essentially the same as in Examples 5.2 and 5.8(a) except that now the system function F depends also on the control variable $a \in A$. Compare (5.3.1) and (5.2.4).

On the other hand, the cost rate function $c(s, x, a)$ is as in (4.1.1), and in some cases—as in the finite–horizon case (4.1.1)—we also need to specify a terminal cost function $C(x)$. $\Diamond$

Given a MCP, we will only consider *Markov control policies* (also known as *closed–loop* or *feedback* controls), that is, measurable functions $\pi : [0, \infty) \times X \to A$ such that $a^\pi := \pi(s, x)$ denotes the control action prescribed by π when the state $x \in X$ is observed at time s. A Markov policy is said to be *stationary* if it is independent of the time parameter s, that is, $\pi(s, x) \equiv \pi(x)$ for all (s, x). We will denote by Π the set of all Markov policies, and by Π_S the subset of stationary policies. Moreover, for technical reasons, we restrict ourselves to the class Π of Markov policies for which the corresponding state (Markov) processes are nicely behaved, in the following sense.

Assumption 5.16. For each $\pi \in \Pi$ there exists a continuous–time Markov process $x^\pi(\cdot) = x(\cdot)$ such that:

(a) Almost all the sample paths of $x(\cdot)$ are right–continuous, with left–hand limits, and have only finitely many discontinuities in any finite interval of time.

(b) $x(\cdot)$ is a Markov process with transition probability denoted by $P^\pi(s, x, t, B)$ and associated semigroup T^π; see (5.2.2).

(c) *The substitution property.* The infinitesimal generator L^π of $x(\cdot)$ satisfies that

$$L^\pi = L^a \quad \text{if} \quad \pi(s, a) = a.$$

(d) The process $x(\cdot)$ is *conservative* in the sense that if $v(s, x) \equiv 1$ for all (s, x), then $L^\pi v = 0$.

(e) There is a nonempty subfamily Π_0 of Π such that $\mathcal{D}(L^\pi)$ is nonempty for all $\pi \in \Pi_0$, and, furthermore, the cost rate $c(s, x, a)$ is such that the function is in M_0 for all $\pi \in \Pi_0$, where $c^\pi(s, x) := c(s, x, \pi(s, x))$ for all $\pi \in \Pi_0$.

To state our final assumption in this chapter we note that, for each $\pi \in \Pi$, the function sets $M \supset M_0 \supset \mathcal{D}(L)$ in Sect. 5.2 depend

on the policy π being used, so they now will be written as M^π, M_0^π, and $\mathcal{D}(L^\pi)$, respectively. With this notation and the family Π_0 in Assumption 5.16(e), we have the following.

Assumption 5.17. There exist nonempty sets $\mathcal{M} \supset \mathcal{M}_0 \supset \mathcal{D}$ such that, for all $\pi \in \Pi_0$,

$$\mathcal{M} \subset M^\pi, \ \mathcal{M}_0 \subset M_0^\pi, \text{ and } \ \mathcal{D} \subset \mathcal{D}(L^\pi).$$

Remark 5.18 (Notation). Given a policy $\pi \in \Pi_0$ and the cost rate $c(s, x, a)$ we will use the notation:

$$c^\pi(s, x) = c(s, x, \pi) = c(s, x, \pi(s, x)) \quad \text{if} \quad x(s) = x. \tag{5.3.2}$$

In particular, if $c(s, x, a) = c(x, a)$ is *independent of the time parameter* s, then (5.3.2) means:

$$c^\pi(s, x) = c(x, \pi(s, x)) \quad \text{if} \quad x(s) = x. \tag{5.3.3}$$

If, in addition, π is stationary, so $\pi(s, x) \equiv \pi(x)$, then (5.3.3) becomes

$$c^\pi(x) = c(x, \pi) = c(x, \pi(x)). \tag{5.3.4}$$

Similarly, when using a policy $\pi \in \Pi_0$, expectations such as (5.2.2) will be written as

$$T_t^\pi v(s, x) = E_{s,x}^\pi v(s + t, x(s + t)) = \int_X P^\pi(s, x, s+t, dy) v(s + t, y) \tag{5.3.5}$$

or, by the *substitution property* in Assumption 5.16(c),

$$T_t^a v(s, x) = E_{s,x}^a v(s + t, x(s + t)) = \int_X P^a(s, x, s + t, dy) v(s + t, y) \tag{5.3.6}$$

if $\pi(s, x) = a$. In particular, for a time–homogeneous MCP and a function $v(s, x) \equiv v(x)$ in $\mathcal{D}$, if $\pi(s, x) = a$, then

$$T_t^\pi v(s, x) = T_t^a v(x) \quad \text{and} \quad L^\pi v(s, x) = L^a v(x). \tag{5.3.7}$$

$\Diamond$

5.4 The Dynamic Programming Approach

As noted in previous chapters, when using the dynamic programming (DP) approach to study a given optimal control problem (OCP) the idea is to obtain an equation—the so–called *Bellman equation* or *dynamic programming equation* (DPE)—from which we can obtain, under appropriate conditions, the OCP's value function and also optimal control policies.

Remark 5.19. For continuous–time MCPs, the dynamic programming (or Bellman) equation is also known as the *Hamilton–Jacobi–Bellman* (HJB) equation. ◊

In the remainder of this chapter we use the DP approach to analyze some OCPs associated to a general continuous–time MCP (X, A, L^a, c) that satisfies the conditions in Sect. 5.3, in particular, Assumptions 5.16 and 5.17. We consider, first, a finite–horizon OCP.

Fix $\rho \geq 0$ and $T > 0$. Consider the cost functional

$$V(s,x,\pi) := E^{\pi}_{s,x}[\int_s^T e^{-\rho(t-s)}c^{\pi}(t,x(t))dt + e^{-\rho(T-s)}K(T,x(T)) \tag{5.4.1}$$

with $0 \leq s \leq T, x \in X$, and $\pi \in \Pi_0$, where $K \in \mathcal{M}$ is a given nonnegative function representing a "terminal cost" at the terminal time T. (Recall that $\mathcal{M}$ is the set in Assumption 5.17.)

For future reference, compare (5.4.1) and (5.2.8): they are essentially the same except that (5.4.1) *depends on* π.

If $\rho = 0$ in (5.4.1), then $V(s,x,\pi)$ is called the *expected total cost* during the interval $[s,T]$ when using the policy π. If, on the other hand, $\rho > 0$ then V is the *discounted cost* during $[s,T]$ when using π. In either case, the OCP is to find a policy π^* such

$$V(s,x,\pi^*) = \inf_{\pi} V(s,x,\pi) =: V^*(s,x) \;\; \forall\, (s,x) \in X_T. \tag{5.4.2}$$

If this is the case, then we say that π^* is an *optimal policy*, and the function V^* is called the OCP's *value function* or *optimal*

cost function. The corresponding dynamic programming theorem is Theorem 5.21 below.

Remark 5.20. Compare Theorem 4.6 and the following Theorem 5.21 with $\rho = 0$: the theorems are the same except that $L^a v$ in (5.4.3) is written in the form (4.1.12), and the terminal cost $K(T,x)$ in (5.4.4) takes the form $C(x)$ in (4.1.9). In other words, Theorem 4.6 is a *special case* of Theorem 5.21 when the Markov control problem is given by the deterministic system (4.0.1)–(4.1.1). Similarly, in the following chapter we will specialize Theorem 5.21 (and also Theorem 5.23) to controlled diffusion processes, which is a class of controlled stochastic differential equations. ◊

Recall that $\mathcal{D}$ is the set in Assumption 5.17.

Theorem 5.21. *Suppose that $v \in \mathcal{D}$ satisfies the equation*

$$\rho v(s,x) = \inf_{a \in A}[c(s,x,a) + L^a v(s,x)] \ \ \forall\ (s,x) \in X_T \tag{5.4.3}$$

with the boundary (or "terminal") condition

$$v(T,x) = K(T,x) \ \ \forall\ x \in X. \tag{5.4.4}$$

Then:

(a) $v(s,x) \le V(s,x,\pi)$ *for all* $(s,x) \in X_T$ *and* $\pi \in \Pi_0$.

(b) If $\pi^* \in \Pi_0$ *is an admissible Markov policy such that* $\pi^*(s,x)$ *attains the minimum in the right-hand side of (5.4.3), that is (using the notation in Remark 5.18),*

$$\rho v(s,x) = c^{\pi^*}(s,x) + L^{\pi^*} v(s,x) \ \ \forall\ (s,x) \in X_T, \tag{5.4.5}$$

then $v(s,x) = V(s,x,\pi^*)$, *and so (by part (a))* π^* *is an optimal policy and* $v = V^*$ *is the optimal cost function in (5.4.2).*

Proof.

(a) Suppose that v satisfies (5.4.3). Then, for all $(s,x) \in X_T$ and $a \in A$,

$$\rho v(s,x) \le c(s,x,a) + L^a v(s,x).$$

Therefore, for any $\pi \in \Pi_0$,

$$\rho v(s,x) \leq c^{\pi}(s,x) + L^{\pi}v(s,x).$$

Note that this inequality can be written as in Proposition 5.11(a'), that is, $\rho v \leq c^{\pi} + L^{\pi}v$. Hence, this inequality and (5.4.4) yield (by Proposition 5.11(a') and comparing (5.4.1) with (5.2.8))

$$v(s,x) \leq V(s,x,\pi) \ \ \forall\ (s,x) \in X_T.$$

This completes the proof of part (a).

(b) If (5.4.5) and (5.4.4) hold, then Proposition 5.11(a) and (5.4.1) give

$$v(s,x) = V(s,x,\pi^*) \ \ \forall\ (s,x) \in X_T.$$

Thus, part (a) and (5.4.2) yield the desired conclusion. □

To conclude this section, we consider the *infinite–horizon* version of (5.4.1), with $\rho > 0$, a given discount factor, and $K \equiv 0$. Hence, consider

$$V_{\infty}(s,x,\pi) := E^{\pi}_{s,x} \int_s^{\infty} e^{-\rho(t-s)} c^{\pi}(t,x(t))dt \tag{5.4.6}$$

$$= \lim_{T\to\infty} E^{\pi}_{s,x} \int_s^{T} e^{-\rho(t-s)} c^{\pi}(t,x(t))dt$$

for all (s,x) in $X_{\infty} := [0,\infty) \times X$, and $\pi \in \Pi_0$. (Compare (5.4.6) and (5.2.10).) The corresponding value (or optimal cost) function is

$$V^*_{\infty}(s,x) := \inf_{\pi} V_{\infty}(s,x,\pi).$$

As usual, a policy $\pi^* \in \Pi_0$ is said to be *optimal* if $V_{\infty}(s,x,\pi^*) = V^*_{\infty}(s,x)$ for all $(s,x) \in X_{\infty}$. On the other hand, to have a non-trivial OCP we assume the following.

Assumption 5.22. There exists an admissible policy $\pi \in \Pi_0$ such that $V_{\infty}(s,x,\pi) < \infty$ for every $(s,x) \in X_{\infty}$.

Assumption 5.22 ensures that $V^*_{\infty}(s,x) < \infty$ for every initial condition (s,x). As an example, Assumption 5.22 trivially holds

if $c(s, x.a)$ is *bounded*, that is, for some positive constant K, $0 \leq c(s, x, a) \leq K$ for all $(s, x, a) \in X_\infty \times A$. In this case, (5.4.6) yields

$$V_\infty(s, x, \pi) \leq K/\rho \quad \forall\, s, x, \pi.$$

In the *infinite–horizon* case (5.4.6), the DP Theorem 5.21 becomes as follows. (Note that Theorem 4.15, in Sect. 4.3, is a *deterministic* version of Theorem 5.23. In particular, compare (5.4.8) and (4.3.3).)

Theorem 5.23. *Suppose that $v \in \mathcal{D}$ satisfies the equation*

$$\rho v(s, x) = \inf_{a \in A}[c(s, x, a) + L^a v(s, x)] \tag{5.4.7}$$

for all $(s, x) \in X_\infty$. Then:

(a) $v(s, x) \leq V_\infty(s, x, \pi)$ for every policy $\pi \in \Pi_0$ such that

$$e^{-\rho t} T_t^\pi v(s, x) \to 0 \quad \text{as} \quad t \to \infty. \tag{5.4.8}$$

(b) If $\pi^* \in \Pi_0$ is such that $\pi^*(s, x) \in A$ attains the minimum in (5.4.7), that is,

$$\rho v(s, x) = c^{\pi^*}(s, x) + L^{\pi^*} v(s, x) \tag{5.4.9}$$

for all $(s, x) \in X_\infty$, then π^* is optimal within the class of policies $\pi \in \Pi_0$ that satisfy (5.4.8) and, moreover, $v(s, x) = V_\infty^*(s, x) = V_\infty(s, x, \pi^*)$ for all (s, x).

Proof. As in the proof of Theorem 5.21(a), the relation (5.4.7) implies that

$$\rho v \leq c^\pi + L^\pi \quad \forall\, \pi \in \Pi_0.$$

If, in addition, $\pi \in \Pi_0$ satisfies the condition (5.4.8), then Proposition 5.11(b') yields that

$$v(\cdot, \cdot) \leq V(\cdot, \cdot, \pi).$$

This proves part (a). Similarly, if π^* satisfies (5.4.9) and (5.4.8), then Proposition 5.11(b) gives that

$$v(\cdot,\cdot) = V_\infty(\cdot,\cdot,\pi^*).$$

Therefore, the desired conclusion in (b) follows from (a). □

5.5 Long–Run Average Cost Problems

In this section we consider a time-homogeneous continuous–time MCP (X, A, L^a, c), with a nonnegative cost function c. We will use the notation (5.3.3), (5.3.4), and (5.3.7).

For each $t \geq 0, x \in X$, and $\pi \in \Pi$, let

$$J_t(x,\pi) := E_x^\pi \int_0^t c^\pi(r, x(r))dr \tag{5.5.1}$$

be the total expected cost in $[0,t]$, when using the control policy π, given the initial state $x(0) = x$. (In (5.5.1) we are using the notation (5.3.3), according to which $c^\pi(r,x) := c(x, \pi(r,x))$ if $x(r) = x$.)

As in (5.2.13)–(5.2.14), we now consider the long-run expected average cost, or simply the *average cost* (AC), when using $\pi \in \Pi$, defined as

$$J(x,\pi) := \limsup_{t\to\infty} J_t(x,\pi)/t \tag{5.5.2}$$

for each initial state x. (As a particular case, see the *deterministic* problem (4.4.1)–(4.4.3).)

Assumption 5.24. There exists a policy $\pi \in \Pi$ such that $J(x,\pi) < \infty$ for every $x \in X$.

For instance, if c is bounded (say, there is a constant $\bar{c}$ such that $0 \leq c(x,a) \leq \bar{c}$ for all $x \in X$ and $a \in A$), then Assumption 5.24 holds with $J(x,\pi) \leq \bar{c}$ for all x, π.

Under our current assumptions, the AC–value function

$$J^*(x) := \inf_{\pi\in\Pi} J(x,\pi),\ x \in X, \tag{5.5.3}$$

is finite–valued. As usual, a policy $\pi^* \in \Pi$ is said to be optimal with respect to (5.5.2), or *AC–optimal*, if

$$J(x,\pi^*) = J^*(x) \quad \forall\, x \in X.$$

To analyze the AC–optimal control problem, we will first use Proposition 5.13 to obtain a characterization of $J(x,\pi)$.

Remark 5.25. In (5.5.4)–(5.5.5) below we use the notation (5.3.3) and (5.3.7). ◊

Proposition 5.26. (a) Let $\pi \in \Pi$ be a policy for which the following holds: There exists a number j^π and a function $h^\pi \in \mathcal{D}$ such that the pair (j^π, h^π) satisfies the *Poisson equation*

$$j^\pi = c^\pi(s,x) + L^\pi h^\pi(s,x) \quad \forall\, s,x \tag{5.5.4}$$

and, furthermore,

$$\lim_{t\to\infty} T_t^\pi h^\pi(s,x)/t = 0. \tag{5.5.5}$$

Then $J(\cdot,\pi)$ is the constant j^π, i.e.,

$$j^\pi = J(x,\pi) \quad \forall\, x \in X. \tag{5.5.6}$$

(b) If in (5.5.4) we replace the equality by either $\leq$ or $\geq$, then in (5.5.6) the equality is replaced by $\leq$ or $\geq$, respectively.

We will omit the proof of Proposition 5.26 because it is the same as that of Proposition 5.13. On the other hand, observe that (5.5.5) is obviously true if h^π is a bounded function.

We will next consider the AC optimal control problem in which we wish to minimize the function $\pi \mapsto J(x,\pi)$ for every $x \in X$. To this end we introduce the following definition.

Definition 5.27. Consider a pair (j^*, h^*) that consists of a real number j^* and a function $h^* \in \mathcal{D}$. The pair (j^*, h^*) is called
(a) a solution to the *average cost optimality equation* (ACOE) if

$$j^* = \inf_{a\in A}[c(x,a) + L^a h^*(x)] \quad \forall\, x \in X; \tag{5.5.7}$$

(b) a solution to the *average cost optimality inequality* (ACOI) if

$$j^* \geq \inf_{a\in A}[c(x,a) + L^a h^*(x)] \quad \forall\, x \in X. \tag{5.5.8}$$

A large part of the analysis of AC problems concerns either the ACOE (5.5.7) or the ACOI (5.5.8). This is mainly due to the following theorem.

Theorem 5.28. *Let $(j^*, h^*) \in \mathbb{R} \times \mathcal{D}$ be a solution to the ACOE (5.5.7). Let $\Pi^{AC} \subset \Pi$ be family of policies $\pi \in \Pi$ such that*

$$\lim_{t\to\infty} T_t^{\pi} h^*(s,x)/t = 0 \quad \forall\, s, x. \tag{5.5.9}$$

Then for every $x \in X$:
(a) $j^* \le \inf_{\pi\in\Pi^{AC}} J(x,\pi)$; hence
(a') $j^* \le J^*(x)$ if $\Pi^{AC} = \Pi$, where $J^*(\cdot)$ is the AC–value function in (5.5.3). Moreover, let $\Pi_S^{AC} \subset \Pi_S$ be the family of stationary policies that satisfy (5.5.9). If $\pi^* \in \Pi_S^{AC}$ is such that, for every $x \in X$, $\pi^*(x) \in A$ minimizes the right–hand side of (5.5.7), i.e.,

$$j^* = c^{\pi^*}(x) + L^{\pi^*} h^*(x) \quad \forall\, x \in X, \tag{5.5.10}$$

then, for all $x \in X$,
(b) $j^* = J(x,\pi^*) = \inf_{\pi\in\Pi_S^{AC}} J(x,\pi)$; hence
(b') $j^* = J(x,\pi^*) = J^*(x)$ if $\Pi^{AC} = \Pi$. In words, if $\pi^* \in \Pi_S$ is a stationary policy that satisfies (5.5.10) and, in addition, (5.5.9) holds for every $\pi \in \Pi$, then π^* is AC–optimal, and the optimal cost is the constant $J(\cdot,\pi^*) \equiv j^*$.
(c) Suppose that, instead of (5.5.10), $\pi^* \in \Pi_S^{AC}$ minimizes the right–hand side of the ACOI (5.5.8), i.e.,

$$j^* \ge c^{\pi^*}(x) + L^{\pi^*} h^*(x) \quad \forall\, x \in X. \tag{5.5.11}$$

Then

$$j^* \ge J(x,\pi^*) \ge J^*(x) \quad \forall\, x \in X. \tag{5.5.12}$$

(c') Part (b') holds, that is, $j^* = J(\cdot,\pi^*) = J^*(\cdot)$ if $\Pi^{AC} = \Pi$.

Proof. (a) By (5.5.7),

$$j^* \le c(x,a) + L^a h^*(x) \quad \forall\, x \in X, a \in A,$$

and so

$$j^* \le c^{\pi}(s,x) + L^{\pi} h^*(s,x) \quad \forall\, \pi \in \Pi.$$

Therefore, by Proposition 5.26(b) and (5.5.9),

$$j^* \leq J(x,\pi) \quad \forall\, x \in X, \pi \in \Pi^{AC}.$$

This implies (a), and also (a') if $\Pi^{AC} = \Pi$.

(b) From Proposition (5.2.14)(a), $j^* = J(x,\pi^*)$ for all $x \in X$. Hence, (b) follows from part (a). Clearly, (b) implies (b').

(c) The first inequality in (5.5.12) is a consequence of (5.5.11) and Proposition 5.26(b). The second inequality follows from the definition (5.5.3) of J^*. Finally, parts (c) and (a) give (c'). □

Remark 5.29. In results such as Theorem 5.28(a') or (c), we require conditions ensuring the existence of measurable mappings $\pi^* : X \to A$ such that, for every $x \in X$, $\pi^*(x) \in A$ attains the minimum in the right–hand side of (5.5.7) or (5.5.8); that is, if

$$v(x,a) := c(x,a) + L^a h^*(x), \tag{5.5.13}$$

then

$$\inf_{a \in A} v(x,a) = v^{\pi^*}(x) := v(x,\pi^*(x)) \tag{5.5.14}$$

for all $x \in X$. These conditions can be obtained from results as those in Appendix B. For example, suppose that A is a compact metric space, and in Theorem B.3 consider the "constant" multifunction $\Phi(\cdot) \equiv A$. Suppose, in addition, that v in (5.5.13) is such that $a \mapsto v(x,a)$ is l.s.c. on A for each $x \in X$. Then Theorem B.3 gives the existence of π^* that satisfies (5.5.14). If A is not compact, we can try to use Theorems B.8 or B.9, for instance. ◊

Theorem 5.28 shows that the ACOE (5.5.7) and the ACOI (5.5.8) give a lower bound or an upper bound, respectively, for the optimal AC function $J^*(\cdot)$. They also give means to obtain an AC–optimal policy $\pi^* \in \Pi_S^{AC}$. Then the obvious question is, of course, *how to obtain a solution* (j^*, h^*) to (5.5.7) or (5.5.8)? There are several ways to answer this question, depending on the underlying assumptions, such as the *ergodicity approach*, *the vanishing discount approach*, the *infinite–dimensional linear programming*

approach,..., etc. In fact, to the best of our knowledge, none of these approaches has been developed for the *general MCPs introduced in this chapter*. The first two, however, can be naturally extended to our current context—see the following Sects. 5.5.1 and 5.5.2. The infinite-dimensional linear programming approach requires a more technical background, but the general ideas are as in the discrete-time case in Hernández-Lerma and Lasserre (1996) and (1999) Chaps. 6 and 12, respectively.

5.5.1 The Ergodicity Approach

In Remark 5.14(d) and Proposition 5.26, let us suppose that for each stationary policy $\pi \in \Pi_S$, the corresponding Markov process $x(\cdot)$ is *uniformly geometrically ergodic* in the sense of (5.2.18), that is, for every $\pi \in \Pi_S, t \geq 0$, and $x \in X$,

$$\|P^\pi(t, x, \cdot) - \mu^\pi(\cdot)\|^* \leq \theta e^{-\gamma t}, \tag{5.5.15}$$

where θ and γ are positive constants. Let us suppose, in addition, that the cost function c is bounded. Then defining $j^\pi \in \mathbb{R}$ and $h^\pi(\cdot) \in \mathcal{D}$ as in (5.2.19), we obtain a solution (j^π, h^π) to the Poisson equation (5.5.4), i.e., for each $\pi \in \Pi_S$,

$$j^\pi = c^\pi(x) + L^\pi h^\pi(x) \quad \forall\, x \in X. \tag{5.5.16}$$

Moreover, since c is bounded, then so is h^π and hence (5.5.5) holds. Therefore, from (5.5.6), for every $\pi \in \Pi_S$ and $x \in X$,

$$j^\pi = J(x, \pi) \geq \inf_{\pi \in \Pi_S} J(x, \pi). \tag{5.5.17}$$

For examples and further comments on the ergodicity approach, see Sect. 6.5. In the meantime, note that results such as (5.5.15) are well known in the literature on Markov processes; see, for instance, Down et al. (1995) or Lund et al. (1996).

5.5.2 The Vanishing Discount Approach

The vanishing discount approach to the AC problem refers to the analysis of the ρ–discounted cost (5.4.6) for a time–homogeneous MCP, say,

$$V^{\rho}(x,\pi) := E_x^{\pi}\int_0^{\infty} e^{-\rho t}c^{\pi}(x(t))dt \tag{5.5.18}$$

as "the discount ρ vanishes", that is, as $\rho \downarrow 0$. (For *deterministic* continuous-time AC problems, the vanishing discount approach is studied in Sect. 4.4.3.)

There are several way to see the connection between (5.5.18) and the AC (5.5.2). For instance, from the Abelian theorems in Exercises 5.7 and 5.8(c) it can be seen that if (5.5.2) holds with "lim sup" replaced by "limit", i.e.,

$$J(x,\pi) = \lim_{t\to\infty} J_t(x,\pi)/t,$$

then

$$\lim_{\rho\downarrow 0} \rho V^{\rho}(x,\pi) = J(x,\pi). \tag{5.5.19}$$

(See Exercise 5.8(b) or (c).)

Alternatively, (5.5.19) can be obtained in the context of (5.5.15)–(5.5.17). Indeed, inside the integral in (5.5.18) replace $c^{\pi}(\cdot)$ with $c^{\pi}(\cdot) - j^{\pi} + j^{\pi}$, with $j^{\pi} = J(x,\pi)$ as in (5.5.17). Then $V^{\rho}(x,\pi)$ can be expressed as

$$V^{\rho}(x,\pi) = E_x^{\pi}\int_0^{\infty} e^{-\rho t}[c^{\pi}(x(t)) - j^{\pi}]dt + j^{\pi}/\rho,$$

so, multiplying both sides by ρ, we obtain

$$\rho V^{\rho}(x,\pi) = j^{\pi} + \rho E_x^{\pi}\int_0^{\infty} e^{-\rho t}[c^{\pi}(x(t)) - j^{\pi}]dt. \tag{5.5.20}$$

Therefore, this fact yields again (5.5.19) provided that the right-most term in (5.5.20) tends to zero as $\rho \downarrow 0$, i.e.,

$$\lim_{\rho\downarrow 0} \rho E_x^{\pi}\int_0^{\infty} e^{-\rho t}\,[c^{\pi}(x(t)) - j^{\pi}]\,dt = 0. \tag{5.5.21}$$

As an example, this is true if (5.5.15) holds. (See Exercise 5.10.)

The starting point of the so–called "vanishing discount approach" in stochastic control theory is the ρ–discount dynamic programming equation (5.4.7) for a time–homogeneous MCP, that is, with $v_\rho(\cdot) \equiv v(\cdot)$,

$$\rho v_\rho(x) = \inf_{a \in A}[c(x,a) + L^a v_\rho(x)], x \in X. \tag{5.5.22}$$

Recall that, in the time-homogeneous case,

$$v_\rho(x) := \inf_\pi v_\rho(x,\pi),$$

with $v_\rho(x,\pi) := E_x^\pi \int_0^\infty e^{-\rho t} c^\pi(x(t))dt$.

For applications of the vanishing discount approach to controlled diffusion processes, see Sect. 6.5.

Notes—Chapter 5

1. Most of the material in this chapter comes from Hernández-Lerma (1994) but, in fact, general continuous–time MCPs are a standard subject; see Doshi (1976a, b, 1979), Fleming (1984), Gihman and Skorohod (1979), Hernández-Lerma and Govindan (2001), Rishel (1990), and their references. For *noncontrolled* continuous-time Markov processes, as in Sects. 5.1 and 5.2, above, there are many excellent textbooks; see, for instance, Evans (2013) or Mikosch (1998) or the introductory chapters in Arnold (1974), Hanson (2007), Øksendal (2003), ...

2. In these notes, as examples of continuous–time MCPs, we only consider the deterministic systems in Chap. 4, and the controlled diffusion processes in Chap. 6. There are, however, many other important classes of continuous–time MCPs, such as controlled jump-Markov processes with a countable state space (see, for instance, Guo and Hernández-Lerma (2009) or Prieto-Rumeau and Hernández-Lerma (2012)) or an uncountable (Borel) state space (as in Piunovskiy and Zhang (2020)), and controlled jump–diffusion processes (as in Hanson (2007) or Øksendal and Sulem (2007)).

Exercises

5.1. Let T_t be as in (5.2.2). Prove that $T_t, t \geq 0$, is a semigroup of operators on M, that is, T_0 =Identity, each T_t maps M into itself, and $T_{t+r} = T_t T_r$ for all $t, r \geq 0$.

5.2. Prove Lemma 5.9.

5.3. Prove Proposition 5.11(b').

5.4. Show directly that Propositions 5.11(a) and (a') hold when $\rho = 0$. More explicitly, suppose that the functions c and K are as in Proposition 5.11. Then:
(a) If $v \in \mathcal{D}(L)$ satisfies that

$$c(s,x) + Lv(s,x) = 0 \quad \forall\, (s,x) \in X_T,$$

with the terminal condition (5.2.7), then

$$v(s,x) = E_{s,x}[\int_s^T c(t,x(t))dt + K(T,x(T))] \quad \forall\, (s,x) \in X_T.$$

Similarly for (a').

Remark. Let bM_0 be the class of functions $c \in M_0$ that are *bounded* in the supremum norm, $\|c\| := \sup_{s,x} |c(s,x)|$. For fixed $\rho > 0$, the operator R_ρ on bM_0 defined by (5.2.10), i.e.,

$$R_\rho c(s,x) := \int_0^\infty e^{-\rho t} T_t c(s,x) dt \tag{5.5.23}$$

is called the *resolvent* of the semigroup T_t. The following exercise shows that the resolvent is the *unique* solution of (5.2.6) in $\mathcal{D}(L)$ if c is in bM_0. ♢

5.5. Show that if c is in bM_0, then the resolvent $v := R_\rho c$ in (5.5.23) is the unique function in $\mathcal{D}(L)$ that satisfies (5.2.6) for all $(s,x) \in X_\infty$.

5.6. Prove the statement in Remark 5.14(b).

The result in the following Exercise 5.7 is a so–called *Abelian theorem*, well known in Laplace transform theory. (See Widder

(1941), pp. 181–182, for instance.) As shown in Exercise 5.8, Abelian theorems establish a connection between *discounted costs* (as in (5.2.10)) and *long–run average costs* (as in (5.2.14)). Exercises 5.7 and 5.8 extend to general MCPs the results in Lemma 4.31 and Proposition 4.32.

5.7. For $t \geq 0$, let $t \mapsto \alpha(t)$ be a nondecreasing function with $\alpha(0) = 0$. Define

$$\alpha_{\inf} := \liminf_{t\to\infty} \alpha(t)/t, \ \alpha^{\sup} := \limsup_{t\to\infty} \alpha(t)/t$$

and suppose $\alpha^{\sup} < \infty$. Prove that, for every $\rho > 0$,
(a) $\int_0^\infty e^{-\rho t} d\alpha(t) = \rho \int_0^\infty e^{-\rho t} \alpha(t) dt$;
(b) $\alpha_{\inf} \leq \liminf_{\rho\downarrow 0} \rho \int_0^\infty e^{-\rho t} d\alpha(t) \leq \limsup_{\rho\downarrow 0} \rho \int_0^\infty e^{-\rho t} d\alpha(t) \leq \alpha^{\sup}$;
(c) If the limit $\alpha(t)/t \to \alpha^*$ exists, then

$$\alpha^* = \lim_{\rho\downarrow 0} \rho \int_0^\infty e^{-\rho t} d\alpha(t).$$

5.8. Let $c \in M_0$ be nonnegative. In Exercise 5.7, let

$$\alpha(t) := \int_0^t T_r c(s,x) dr,$$

so the discounted cost v^ρ in (5.2.10) and the long-run average cost $J^{\sup}$ in (5.2.14) become

$$\begin{aligned} v^\rho(s,x) &= \int_0^\infty e^{-\rho t} T_t c(s,x) dt \\ &= \int_0^\infty e^{-\rho t} d\alpha(t) \end{aligned}$$

and $J^{\sup}(s,x) = \alpha^{\sup}$, respectively, with $\alpha^{\sup} < \infty$ as in Exercise 5.7. Then
(a) for every $\rho > 0$,

$$v^\rho(s,x) = \rho \int_0^\infty e^{-\rho t} \left[\int_0^t T_r c(s,x) dr \right] dt;$$

(b) $J_{\inf}(s,x) \leq \liminf_{\rho\downarrow 0} \rho v^\rho(s,x)$

$$\leq \limsup_{\rho\downarrow 0} \rho v^{\rho}(s,x) \leq J^{\sup}(x,s),$$

with $J^{\sup}$ and $J_{\inf}$ as in (5.2.14) and (5.2.15), respectively.
(c) If $J_t(s,x) \to j^*$ as $t \to \infty$, then $\lim_{\rho\downarrow 0} \rho v^{\rho}(s,x) = j^*$.

5.9. Prove the statement in Remark 5.14(d). More explicitly, let $\mathcal{X} = \{x(t), t \geq 0\}$ be a time–homogeneous Markov process with an invariant probability measure μ and geometrically (or uniformly) ergodic in the sense of (5.2.18). Let $c \in M_0$ be a bounded function, say, $|c(x)| \leq \bar{c}$ for all $x \in X$. Then the function h_c in (5.2.19) is in $\mathcal{D}(L)$, and the pair $(j(c), h_c)$ in (5.2.19) is a solution to the Poisson equation (5.2.20). (The uniform ergodicity condition (5.2.18) is well known for several norms $\|\cdot\|^*$ and different Markov processes. See for instance (5.3.1))
Solution. Observe that, for all $r \geq 0$ and $x \in X$,

$$\begin{aligned} |T_r c(x) - j(c)| &= |\int_X [P(r,x,dy) - \mu(dy)]|c(y)| \\ &\leq \bar{c}\|P(r,x,\cdot) - \mu(\cdot)\|^* \\ &\leq \bar{c}\theta e^{-\gamma r}. \end{aligned} \tag{5.5.24}$$

In the latter inequality we used (5.2.18). Note that h_c is bounded, because (from (5.2.19) and (5.2.20))

$$\begin{aligned} |h_c(x)| &\leq \int_0^{\infty} |T_r c(x) - j(c)| dr \\ &\leq \bar{c}\theta/\gamma \quad \forall\, x \in X. \end{aligned}$$

Moreover (since (5.5.24) allows the interchange of integrals),

$$\begin{aligned} T_s h_c(x) &= \int_0^{\infty} [T_{r+s} c(x) - j(c)] dr \\ &= \int_s^{\infty} [T_r c(x) - j(c)] dr \\ &= h_c(x) - \int_0^{s} [T_r c(x) - j(c)] dr; \end{aligned}$$

that is,

$$T_s h_c(x) - h_c(x) = -\int_0^{s} [T_r c(x) - j(c)] dr.$$

Finally, multiplying by $1/s$ and then letting $s \downarrow 0$, the Poisson equation (5.2.20) follows. □

5.10. Prove: (5.5.15) implies (5.5.21).
Solution. As in (5.5.24), for every $t \geq 0$ and $x \in X$,

$$|E_x^\pi c^\pi(x(t)) - j^\pi| \leq \bar{c}\theta e^{-\gamma t},$$

where $\bar{c}$ is an upper bound for $c(x, a)$. Consequently, for every $x \in X$,

$$|\rho E_x^\pi \int_0^\infty [c^\pi(x(t)) - j^\pi] dt| \leq \rho \bar{c}\theta/\gamma.$$

This fact yields (5.5.21). □

Chapter 6

Controlled Diffusion Processes

6.1 Diffusion Processes

In the remainder of these notes we consider a class of $\mathbb{R}^d$–valued Markov processes $\{x(t), t \geq 0\}$ called (Markov) *diffusion processes*. These are processes that are characterized in a suitable sense by a function $b : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}^d$ called the *drift vector*, and a $d \times d$ matrix D on $[0, \infty) \times \mathbb{R}^d$ called the *diffusion matrix*, which is assumed to be symmetric and nonnegative definite. In the extreme case in which $D \equiv 0$, the zero matrix, the process $x(\cdot)$ is the solution of an ordinary differential equation $\dot{x}(t) = b(t, x(t)), t \geq 0$. At the other extreme, if $b \equiv 0$ and $D \equiv I$ the identity matrix, then $x(\cdot)$ is a Markov process called *Wiener process* or *Brownian motion*. (See Example 5.3, above, or any introductory book on *stochastic analysis* or *stochastic differential equations*, for instance, Arnold (1974), Evans (2013), Mikosch (1998), Øksendal (2003),...)

More precisely, we will consider so–called *Itô diffusions* or *Itô processes* $\{x(t), t \geq 0\}$ that are solutions of stochastic differential equations (SDEs) of the form

$$dx(t) = b(t, x(t))dt + \sigma(t, x(t))dw(t), \tag{6.1.1}$$

where $b(t, x)$ and $\sigma(t, x)$ are given functions from $[0, \infty) \times \mathbb{R}^d$ to $\mathbb{R}^d$ and $\mathbb{R}^{d \times n}$, respectively, and $\{w(t), t \geq 0\}$ is a standard n–dimensional Wiener process. To begin, throughout the following

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*, Texts in Applied Mathematics 76,
https://doi.org/10.1007/978-3-031-21139-3_6

we impose conditions on the coefficients b and σ ensuring that (6.1.1) has indeed a well-defined (and well–behaved) solution.

Assumption 6.1. Itô conditions. The functions $b(t,x)$ and $\sigma(t,x)$ are measurable and satisfy:

(a) Linear growth: For every $T>0$, there is a constant $K = K(T)$ such that, for all $0 \le t \le T$ and $x \in \mathbb{R}^d$,

$$|b(t,x)| \le K(1+|x|), \quad |\sigma(t,x)| \le K(1+|x|),$$

(where, for a matrix $d=(d_{ij})$, we define its norm $|d|^2 := Tr(dd^*) = \sum_{ij} d_{ij}^2$, with $d^* :=$ transpose of d, and $Tr(D) :=$ Trace of a matrix D); and

(b) Lipschitz conditions: For every $T>0$ and $r>0$, there is a constant $K' = K'(T,r)$ such that

$$|b(t,x)-b(t,y)| \le K'|x-y|, \quad |\sigma(t,x)-\sigma(t,y)| \le K'|x-y|$$

for all $0 \le t \le T$ and $|x| \le r$, $|y| \le r$.

The conditions (a) and (b) in Assumption 6.1 are implied, for instance, by the following:

(a') The components of $b(t,x)$ and $\sigma(t,x)$ are continuously differentiable in x with bounded derivatives, uniformly in $t \ge 0$;

(b') $|b(t,0)| + |\sigma(t,0)| \le K$ for all $t \ge 0$, for some constant K.

Under Assumption 6.1 the SDE (6.1.1) has a unique continuous solution $x(\cdot)$, which is a Markov process with transition probabilities

$$P(s,x,t,B) = P(x(t) \in B | x(s)=x) = P(x(t;s,x) \in B) \quad (6.1.2)$$

for all $0 \le s \le t$, $x \in \mathbb{R}^d$, $B \in \mathcal{B}(\mathbb{R}^d)$ where $x(t) = x(t;s,x)$ denotes the solution of (6.1.1) for $t \ge s$, with initial condition $x(s)=x$. Moreover, denoting by $E_{s,x}$ the conditional expectation given the initial condition $x(s)=x$, we also have

$$E_{s,x}|x(t)|^k \le (1+|x|^k)e^{C(t-s)} \quad (k=1,2,\ldots) \qquad (6.1.3)$$

for some constant C depending on the integer k and the constant K in Assumption 6.1(a).

The solution process $x(\cdot)$ has other nice properties. For instance, it is continuous and satisfies the "Feller property", which implies that $x(\cdot)$ is in fact a *strong* Markov process. (We do not use these assertions here.)

We will also suppose:

Assumption 6.2. The functions b and σ are continuous in the time variable $t \geq 0$.

Assumptions 6.1 and 6.2 imply that the solution $x(\cdot)$ of (6.1.1) is a Markov diffusion process with *drift coefficient* $b(t,x)$ and *diffusion matrix* $D(t,x) := \sigma(t,x)\sigma(t,x)^*$, where $\sigma(t,x)$ is the diffusion coefficient in (6.1.1). (Recall that σ^* denotes the transpose of σ.) We will next obtain the infinitesimal generator L (see (5.2.3)) of $x(\cdot)$.

Definition 6.3. Let $C^{1,2} \equiv C^{1,2}([0,\infty) \times \mathbb{R}^d)$ be the class of real–valued continuous functions $v(s,x)$ on $[0,\infty) \times \mathbb{R}^d$ such that v is of class C^1 in s and of class C^2 in x, that is, the partial derivates v_s, v_{x_i}, $v_{x_i x_j}$, for $i,j = 1,\ldots,d$, are continuous. If $v \in C^{1,2}$, let

$$\mathcal{L}v(s,x) := v_s(s,x) + v_x(s,x)b(s,x) + \frac{1}{2}Tr[v_{xx}(s,x)D(s,x)], \tag{6.1.4}$$

where $D(x,s)$ is the diffusion matrix, the row vector v_x:=$(v_{x_1},\ldots,v_{x_d})$ is the gradient of v (in the x–variables), and $v_{xx} = (v_{x_i x_j})$ is the Hessian matrix. The last term in (6.1.4) can be expressed more explicitly as

$$\frac{1}{2}Tr[v_{xx}(s,x)D(s,x)] = \frac{1}{2}\sum_{i,j=1}^{d} v_{x_i x_j}(s,x)d_{ij}(s,x),$$

where d_{ij} are the components of $D = \sigma\sigma^*$.

In terms of $\mathcal{L}$ we may write the important *Itô's differential rule* as in (6.1.5), below.

Theorem 6.4. *Let $x(\cdot)$ be the solution of (6.1.1). If $v \in C^{1,2}$, then the process $v(t,x(t))$ satisfies the SDE*

$$dv(t,x(t)) = \mathcal{L}v(t,x(t))dt + v_x(t,x(t))\sigma(t,x(t))dw(t). \tag{6.1.5}$$

In integral form, we may write (6.1.5) for $t \geq s \geq 0$, given $x(s) = x$, as

$$v(t, x(t)) - v(s, x) = \int_s^t \mathcal{L}v(r, x(r))dr + \int_s^t v_x(r, x(r))\sigma(r, x(r))dw(r). \tag{6.1.6}$$

If v and σ are such that, for each $t > s$,

$$E_{s,x} \int_s^t |v_x(r, x(r))\sigma(r, x(r))|^2 dr < \infty, \tag{6.1.7}$$

then the expected value of the last integral in (6.1.6) is zero. Therefore, if in addition to (6.1.7) we have that

$$E_{s,x} \int_s^t |\mathcal{L}v(r, x(r))|dr < \infty, \tag{6.1.8}$$

then taking expectations $E_{s,x}$ in (6.1.6) we obtain

$$E_{s,x}v(t, x(t)) - v(s, x) = E_{s,x} \int_s^t \mathcal{L}v(r, x(r))dr. \tag{6.1.9}$$

Multiplying both sides of (6.1.9) by $(t - s)^{-1}$ and letting $t \downarrow s$ we obtain, from (5.2.3), the infinitesimal generator $Lv = \mathcal{L}v$. More explicitly, we have shown the following.

Theorem 6.5. *If $v \in C^{1,2}$ is such that (6.1.7) and (6.1.8) hold, then v is in the domain $\mathcal{D}(L)$ and Lv is given by*

$$Lv(s, x) = v_s(s, x) + v_x(s, x)b(s, x) + \frac{1}{2}Tr[v_{xx}(s, x)D(s, x)]. \tag{6.1.10}$$

With $\mathcal{L}v = Lv$, (6.1.9) is a particular form of *Dynkin's formula* in Remark 5.10(a).

Remark 6.6. A function $f(s, x)$ is said to satisfy a *polynomial growth condition* if there are constants K and j such that $|f(s, x)| \leq K(1 + |x|^j)$ for every (s, x). Now, using the inequality (6.1.3) and Assumption 6.1, one can see that if $v \in C^{1,2}$ and its partial derivates v_s, v_{x_i}, $v_{x_i x_j}$ satisfy polynomial growth conditions, then (6.1.7) and (6.1.8) are satisfied. ◇

Remark 6.7. **(a)** Let $v \in C^{1,2}$ be as in Theorem 6.5, and let τ be a stopping time for $x(\cdot)$ such that $E_{s,x}(\tau) < \infty$. Then (6.1.9) holds when t is replaced by τ. (See, for instance, Friedman (1975) p. 85.) This fact is analogous to Remark 5.10(b).

(b) Strictly speaking, for every $t \geq s \geq 0$ and every initial condition $x(s) = x$, the SDE (6.1.1) is a compact form of expressing the integral equation

$$x(t) = x + \int_s^t b(r, x(r))dr + \int_s^t \sigma(r, x(r))dw(r),$$

where the second integral on the right–hand side, with respect to the Wiener process $w(\cdot)$, is an *Itô integral*. Developing this approach, however, is out of the scope of these lecture notes. We are thus proceeding as in Chap. 5, in which instead of analyzing the properties of a Markov process $x(\cdot)$ we use directly the corresponding generator. Formally, this is all we need to develop the dynamic programming approach, as in Sect. 5.4.

6.2 Controlled Diffusion Processes

Let A, the *control* (or *action*) *set*, be a closed subset of $\mathbb{R}^m$ and, instead of (6.1.1), consider the controlled SDE

$$dx(t) = b(t, x(t), a(t))dt + \sigma(t, x(t), a(t))dw(t), \tag{6.2.1}$$

with coefficients b and σ, which are functions from $[0, \infty) \times \mathbb{R}^d \times A$ to $\mathbb{R}^d$ and $\mathbb{R}^{d \times n}$, respectively, and $a(t) \in A$, for $t \geq 0$, being the control process. As in Chap. 5, we will only consider Markov control policies, so that $a(\cdot)$ is of the form $a(t) = \pi(t, x(t))$, with $\pi : [0, \infty) \times \mathbb{R}^d \to A$ a measurable function. We also need to restrict the set Π of *admissible* control policies. Thus, in view of the Assumptions 6.1, 6.2 and Remark 6.6, we define Π as follows.

Definition 6.8. A Markov control policy π is said to be *admissible* (and we write $\pi \in \Pi$) if

(a) The functions $b^\pi(t,x) := b(t,x,\pi(t,x))$ and $\sigma^\pi(t,x) := \sigma(t,x,\pi(t,x))$ satisfy the Assumptions 6.1 and 6.2—the corresponding solution $x(\cdot)$ of (6.2.1) is written as $x^\pi(\cdot)$;

(b) The generator L^π of $x(\cdot)$ satisfies that $L^\pi = L^a$ if $\pi(s,x) = a$, where, from (6.1.10),

$$L^a v(s,x) = v_s(s,x) + v_x(s,x)b(s,x,a) + \frac{1}{2}Tr[v_{xx}(s,x)D(s,x,a)] \tag{6.2.2}$$

with $D(s,x,a) = \sigma(s,x,a)\sigma(s,x,a)^*$ (recall that σ^* = transpose of σ, and $Tr(\cdot)$ = Trace).

In some cases it is easy to give conditions for a policy to be admissible. For instance, suppose that the control set A contains the origin $0 \in \mathbb{R}^m$, and also (see (a') and (b') in the paragraph following Assumption 6.1):

(a') The functions $b(t,x,a)$ and $\sigma(t,x,a)$ are continuous, of class C^1 in $x \in \mathbb{R}^d$ and $a \in A$ with bounded derivatives (i.e., $|b_x|$, $|b_a|, |\sigma_x|, |\sigma_a| \le C$ for some constant C) uniformly in $t \ge 0$;

(b') $|b(t,0,0)| + \sigma|(t,0,0)| \le C \quad \forall t \ge 0$ and some constant C.

Then a continuous function $\pi : [0,\infty) \times \mathbb{R}^d \to A$ is an admissible Markov control policy if, for instance, it satisfies:

(c') For every $T > 0$, there is a constant K_T (which may also depend on π) such that $|\pi(t,x)| \le K_T(1+|x|)$ for all $0 \le t \le T$ and $x \in \mathbb{R}^d$;

(d') For every $T > 0$ and $r > 0$, there is a constant $K_{T,r}$ (which may depend on π) such that

$$|\pi(t,x) - \pi(t,y)| \le K_{T,r}|x-y|$$

for all $0 \le t \le T$, and $|x|, |y| \le r$.

The conditions (c') and (d') are of course suggested by (a'), (b') and the Itô conditions in Assumption 6.1.

Finally, we will suppose that Assumptions 5.16(e) and 5.17 hold. Notice in particular that, for instance (in view of (6.1.3)), a sufficient condition for the cost rate $c(s,x,a)$ to satisfy Assumption 5.16(e) is that $c^\pi(s,x)$ satisfies a polynomial growth condition, say

$$|c^{\pi}(s,x)| \leq K(1+|x|^{j}) \quad \forall \pi \in \Pi, \tag{6.2.3}$$

where K and j are positive constants, and $c^{\pi}(s,x) := c(s,x,\pi(s,x))$.

We have thus completed the description of the controlled SDE (6.2.1) in the general MCP setting of Sect. 5.3.

6.3 Examples: Finite Horizon

For a cost functional as in (5.4.1), with $\rho = 0$, and a controlled diffusion process determined by (6.2.1), the Dynamic Programming (DP) Theorem 5.21 is valid provided of course that v satisfies the conditions in Theorem 6.5 (that is, $v \in C^{1,2}$ and (6.1.7), (6.1.8) hold). In this case, the generator L^a in (5.4.3) is given by (6.2.2), so that (5.4.3), with $\rho = 0$, becomes

$$v_s(s,x) + \min_{a \in A} \Big[v_x(s,x)b(s,x,a) + \frac{1}{2} Tr(v_{xx}(s,x)D(s,x,a)) + c(s,x,a) \Big] = 0 \tag{6.3.1}$$

for (s,x) in $X_T := [0,T] \times \mathbb{R}^d$, with the boundary condition

$$v(T,x) = K(T,x), \quad x \in \mathbb{R}^d. \tag{6.3.2}$$

Moreover, using the Remark 6.7 we obtain in fact a slightly more general form of Theorem 5.21. To state it, let Q be a given open subset of X_T, and let ζ be the exit time of $(t,x(t))$ from Q, given the initial condition $(s,x) \in Q$, that is,

$$\zeta := \inf\{t > s | (t,x(t)) \notin Q\}. \tag{6.3.3}$$

(In particular, if $Q := (0,T) \times \mathbb{R}^d$, then $\zeta = T$.) Now let $\partial^* Q$ be a closed subset of the boundary ∂Q of Q such that $(\zeta, x(\zeta)) \in \partial^* Q$ with probability 1 for every initial condition $(s,x) \in Q$ and every admissible policy π. Finally, in the expression (5.4.1) of the cost functional $V(s,x,\pi)$, with $\rho = 0$, replace T with ζ to obtain

$$V(s,x,\pi) = E^{\pi}_{s,x}\left[\int_s^{\zeta} c^{\pi}(t,x(t))dt + K(\zeta,x(\zeta))\right]. \tag{6.3.4}$$

Then Theorem 5.21 is valid if (6.3.1) is restricted to hold for $(s,x) \in Q$, and instead of the boundary condition (5.4.4) we take

$$v(s,x) = K(s,x) \quad \forall (s,x) \in \partial^* Q. \tag{6.3.5}$$

Remark. For the existence of solutions to (6.3.1) with either the boundary condition (6.3.2) or (6.3.5), see Bensoussan (1982), Fleming and Rishel (1975), Hanson (2007) or Krylov (1980), for instance.

Example 6.9. (LQ systems). To simplify the exposition we consider first the scalar case ($d = n = m = 1$ in (6.2.1)). The state $x(\cdot) \in \mathbb{R}$ of the system is supposed to satisfy the linear SDE

$$dx(t) = [\gamma(t)x(t) + \beta(t)a(t)]dt + \sigma(t)dw(t), \tag{6.3.6}$$

with coefficients $\gamma(\cdot)$, $\beta(\cdot)$ and $\sigma(\cdot)$ of class $C^1[0,T]$, and the cost functional

$$V(s,x,\pi) := E^{\pi}_{s,x}\left[\int_s^{T} (q(t)x^2(t) + r(t)a^2(t))dt + q_T x^2(T)\right], \tag{6.3.7}$$

where $q(\cdot) \geq 0$ and $r(\cdot) \geq \epsilon > 0$ are continuous functions; and $q_T \geq 0$. Thus the cost rate and the terminal cost are given, respectively, by

$$c(s,x,a) := q(s)x^2 + r(s)a^2, \quad \text{and} \quad K(s,x) := q_T x^2. \tag{6.3.8}$$

We assume that there are no control constraints, so $A = \mathbb{R}$. Notice, on the other hand, that the coefficients of (6.3.6),

$$b(t,x,a) = \gamma(t)x + \beta(t)a, \quad \text{and} \quad \sigma(t,x,a) = \sigma(t) \tag{6.3.9}$$

satisfy the conditions (a'), (b') in the paragraph after Definition 6.8.

Now, from (6.3.8)–(6.3.9), the DP Eq. (6.3.1)–(6.3.2) becomes

$$v_s + \gamma(s)xv_x + \frac{1}{2}\sigma^2(s)v_{xx} + q(s)x^2 + \min_{a\in\mathbb{R}}[\beta(s)v_x a + r(s)a^2] = 0 \tag{6.3.10}$$

with

$$v(T,x) = q_T x^2, \quad x \in \mathbb{R}. \tag{6.3.11}$$

The minimum in (6.3.10) is reached at $a^* = \pi^*(s,x)$ given by

$$\pi^*(s,x) = -\beta(s)v_x/2r(s), \tag{6.3.12}$$

which inserted in (6.3.10) yields

$$v_s + \gamma(s)xv_x + \frac{1}{2}\sigma^2(s)v_{xx} + q(s)x^2 - (\beta(s)v_x)^2/4r(s) = 0. \tag{6.3.13}$$

The question now is how to obtain a solution of (6.3.13). However, by the form of this equation (or by analogy with the discrete–time case), we may try a solution of the form

$$v(s,x) = k(s)x^2 + g(s), \tag{6.3.14}$$

with $k(\cdot)$ and $g(\cdot)$ of class C^1, and $k(\cdot) \geq 0$. Moreover, for (6.3.11) to hold we require

$$k(T) = q_T, \quad \text{and} \quad g(T) = 0. \tag{6.3.15}$$

With this value of $v(s,x)$, the Eq. (6.3.13) becomes

$$[k'(s) + q(s) + 2\gamma(s)k(s) - \beta^2(s)k^2(s)/r(s)]x^2 + g'(s) + \sigma^2(s)k(s) = 0$$

(where "prime" denotes derivative with respect to s). Hence, for v in (6.3.14) to be a solution of (6.3.13) it suffices that $k(\cdot)$ and $g(\cdot)$ satisfy

$$k'(s) = -q(s) - 2\gamma(s)k(s) + r(s)^{-1}\beta^2(s)k^2(s) \tag{6.3.16}$$

and

$$g'(s) = -\sigma^2(s)k(s)$$

for $s < T$. Thus, combined with the boundary condition (6.3.15), $g(\cdot)$ is given by

$$g(s) = \int_s^T \sigma^2(t)k(t)dt, \quad s \le T,$$

and $k(\cdot)$ is uniquely determined by the (Riccati equation) (6.3.16) with the terminal condition $k(T) = q_T$. Finally, from (6.3.12) and (6.3.14), the optimal policy π^* is

$$\pi^*(s,x) = -r(s)^{-1}\beta(s)k(s)x. \tag{6.3.17}$$

Observe that π^* satisfies the conditions (c') and (d') in Sect. 6.2, and so it is admissible (in the sense of Definition 6.8). Moreover, from Theorem 5.21(b), with $\rho = 0$, the value function of the LQ problem (6.3.6)–(6.3.7) is

$$V^*(s,x) = v(s,x) = k(s)x^2 + \int_s^T \sigma^2(t)k(t)dt. \tag{6.3.18}$$

The vector case. Let us suppose now that in (6.3.6), $x \in \mathbb{R}^d$, $a \in \mathbb{R}^m$, $w \in \mathbb{R}^n$, with $\gamma(\cdot)$, $\beta(\cdot)$ and $\sigma(\cdot)$ matrices of appropriate dimensions, of class $C^1[0,T]$ again. Furthermore, the costs in (6.3.7) (see (6.3.8)) are now the quadratic forms

$$c(t,x,a) := x^*q(t)x + a^*r(t)a, \quad K(T,x) := x^*q_Tx,$$

where $q(\cdot)$, $q_T \in \mathbb{R}^{d\times d}$ are symmetric and nonnegative definite, and $r(\cdot) \in \mathbb{R}^{m\times m}$ is symmetric and positive definite. We also assume that $q(\cdot)$ and $r(\cdot)$ are continuous.

The analysis in the vector case is completely analogous to that presented in (6.3.10)–(6.3.18), and it yields that the optimal Markov policy and the optimal value function are [see (6.3.17), (6.3.18)]

$$\pi^*(s,x) = -r(s)^{-1}\beta(s)^*k(s)x, \tag{6.3.19}$$

and

$$J^*(s,x) = x^*k(s)x + \int_s^T Tr[D(t)k(t)]dt \tag{6.3.20}$$

where $D(t) = \sigma(t)\sigma(t)^*$, and $k(\cdot) \in \mathbb{R}^{d\times d}$ is the solution to the matrix Riccati equation [see (6.3.16)]

$$k'(s) = -k(s)\gamma(s) - \gamma(s)^* k(s) + k(s)\beta(s)r(s)^{-1}\beta(s)^* k(s) - q(s), \tag{6.3.21}$$

for $s \leq T$, with the boundary condition $k(T) = q_T$. ◇

Example 6.10. (Optimal portfolio selection.) Let us now consider an example formulated in the context of a financial market with two assets, or "securities". One of them is a risk–free asset, called a *bond*, and the other one is a risky asset called a *stock*. In a consumption–investment problem, also known as an *optimal portfolio selection* problem, a "small investor", that is, an economic agent whose actions cannot influence the market prices, may choose a *portfolio* (investment strategy) and a *consumption strategy* that determine the evolution of his wealth. The problem is to choose these strategies to maximize some "utility" criterion. In this example, the problem is to maximize the expected discounted total utility from consumption (6.3.25) below; in the following Example 6.11, the investor wishes to maximize the expected utility from terminal wealth (6.3.31).

Let $x(t)$ denote the investor's wealth at time t, and suppose that the price $p_1(t)$ of the risk–free asset (the bond) is given by

$$dp_1(t) = rp_1(t)dt,$$

whereas the price $p_2(t)$ of the risky asset (the stock) changes according to the linear SDE

$$dp_2(t) = p_2[\alpha dt + \sigma dw(t)],$$

where $w(\cdot)$ is a 1–dimensional standard Wiener process. Here, r, α, and σ are constants with $r < \alpha$, and $\sigma > 0$. A consumption–investment policy π is a pair $(a_1(\cdot), a_2(\cdot))$ consisting of a *portfolio process* $a_1(\cdot)$ and a *consumption rate process* $a_2(\cdot)$. That is, $a_1(t)$ (respectively, $1 - a_1(t)$) is the fraction of wealth invested in the stock (respectively, the bond) at time t, and $a_2(t)$ is the consumption rate, satisfying the control constraints

$$0 \leq a_1(t) \leq 1, \quad a_2(t) \geq 0. \tag{6.3.22}$$

Thus, when using a given consumption/investment policy π, the wealth $x(\cdot) \equiv x^\pi(\cdot)$ changes according to the SDE

$$dx(t) = (1 - a_1(t))x(t)rdt + a_1(t)x(t)[\alpha dt + \sigma dw(t)] - a_2(t)dt. \tag{6.3.23}$$

The three terms on the right–hand side of (6.3.23) correspond, respectively, to:

(i) gains from money invested in the bond,

(ii) gains from investment in the stock, and

(iii) the decrease in wealth due to consumption.

Rewritten in the standard form (6.2.1), Eq. (6.3.23) becomes

$$dx(t) = [(r + (\alpha - r)a_1(t))x(t) - a_2(t)]dt + \sigma a_1(t)x(t)dw(t). \tag{6.3.24}$$

Now let U be a *utility function*, that is, U is a nonnegative function on $[0, \infty)$, of class C^2, strictly increasing, strictly concave, and such that $U'(0) = +\infty$. Then the consumption–investment problem we are concerned with is to maximize the expected discounted utility from consumption:

$$V(s, x, \pi) = E^{\pi}_{s,x} \int_s^T e^{-\rho t} U(a_2(t))dt, \tag{6.3.25}$$

with discount rate $\rho > 0$. In this case, the DP Eq. (6.3.1)–(6.3.2) becomes

$$v_s + \max_a \left\{ e^{-\rho s} U(a_2) + [(r + (\alpha - r)a_1)x - a_2]v_x + \frac{1}{2}(\sigma a_1 x)^2 v_{xx} \right\} = 0, \tag{6.3.26}$$

with terminal condition $v(T, x) = 0$, and the maximization is over the set of pairs $a = (a_1, a_2)$ satisfying (6.3.22). Ignoring for the moment the constraints (6.3.22), an elementary calculation shows that the function within brackets in (6.3.26) is maximized by $a^* = (a_1^*, a_2^*)$ such that

$$a_1^* = -(\alpha - r)v_x/\sigma^2 x v_{xx}, \quad U'(a_2^*) = e^{\rho s} v_x \tag{6.3.27}$$

provided that $v_x > 0$ and $v_{xx} < 0$ for $x > 0$. The solution to our problem will depend of course on the particular utility function U used in (6.3.25).

Let us suppose that the utility function is of the form $U(a_2) = a_2^\gamma$, with $0 < \gamma < 1$, and propose a solution to (6.3.26) of the form

$$v(s,x) = h(s)x^\gamma, \quad \text{with} \quad h(T) = 0. \tag{6.3.28}$$

In this case, (6.3.27) yields

$$a_1^* = (\alpha - r)/\sigma^2(1-\gamma), \quad a_2^* = x[e^{\rho s}h(s)]^{1/(\gamma-1)}. \tag{6.3.29}$$

Replacing these values in (6.3.26) we obtain

$$[h'(s) + C\gamma h(s) + (1-\gamma)h(s)(e^{\rho s}h(s))^{1/(\gamma-1)}]x^\gamma = 0, \tag{6.3.30}$$

where $C := r + (\alpha - r)^2/2\sigma^2(1-\gamma)$. Since the latter equation holds for all $x > 0$, the function in brackets must be 0. This yields a differential equation for h, which can be solved (making the substitution $g = (e^{\rho s}h)^{1/(\gamma-1)}$) to obtain

$$h(s) = e^{-\rho s}[\beta - \beta e^{-(T-s)/\beta}]^{1-\gamma},$$

with $\beta := (1-\gamma)/(\rho - C\gamma)$. Thus with this function h, and if $\alpha - r \le \sigma^2(1-\gamma)$, the optimal consumption–investment policy is given by (6.3.29). Notice in particular that the optimal process $a_1^*(\cdot)$ is constant, and the optimal consumption rate $a_2^*(\cdot)$ is a linear function of x. ◇

Example 6.11. In the same context of Example 6.10, let us now suppose that there is no consumption, say $a_2(\cdot) \equiv 0$, so that the wealth Eqs. (6.3.23)–(6.3.24) becomes

$$dx(t) = [r + (\alpha - r)a(t)]x(t)dt + \sigma a(t)x(t)dw(t),$$

where $a(\cdot) = a_1(\cdot)$. Moreover, we wish to maximize the expected utility from terminal wealth

$$V(s,x;\pi) := E_{s,x}^\pi U[x(\zeta)], \tag{6.3.31}$$

where U is a utility function such that $U(0) = 0$, and ζ is the first exit time from the open set $Q = (0,T) \times (0,\infty)$. Observe that the utility criterion (6.3.31) is of the form (6.3.4) with

$$c(s,x,a) \equiv 0 \quad \text{and} \quad K(s,x) = U(x).$$

Then the DP equation is

$$v_s + \max_{0\le a\le 1}\left[(r+(\alpha-r)a)xv_x + \frac{1}{2}\sigma^2 a^2 x^2 v_{xx}\right] = 0, \qquad (6.3.32)$$

with the boundary condition (see (6.3.5))

$$v(s,x) = U(x) \quad \text{for} \quad (s,x)\in \partial^* Q,$$

where $\partial^* Q$ is the union of $[0,T]\times\{0\}$ and $\{T\}\times[0,\infty)$. Again ignoring for the moment the constraint $0\le a\le 1$ in (6.3.22), we find that the maximization in (6.3.32) is obtained with

$$a^* = \pi^*(s,x) = -(\alpha-r)v_x/\sigma^2 x v_{xx} \qquad (6.3.33)$$

if $v_x > 0$ and $v_{xx} < 0$. Substitution of a^* in (6.3.32) yields

$$v_s + rxv_x - (\alpha-r)^2 v_x^2/2\sigma^2 v_{xx} \quad \text{on} \quad Q, \qquad (6.3.34)$$

with

$$v(s,x) = U(x) \quad \text{for} \quad s = T \quad \text{or} \quad x = 0. \qquad (6.3.35)$$

To solve (6.3.34)–(6.3.35) we suppose that the utility function is $U(x) = x^\gamma$, with $0<\gamma<1$, and we try a solution of the form

$$v(s,x) = h(s)x^\gamma, \quad \text{where} \quad h(T) = 1.$$

With this choice of v, Eq. (6.3.34) becomes

$$h'(s) + C\gamma h(s) = 0,$$

with C as in (6.3.30). Hence $h(s) = e^{C\gamma(T-s)}$ for $s\le T$, so that

$$v(s,x) = e^{C\gamma(T-s)}x^\gamma, \quad \text{and} \quad \pi^*(s,x) = (\alpha-r)/\sigma^2(1-\gamma). \qquad (6.3.36)$$

Thus if $(\alpha-r)/\sigma^2(1-\gamma)\le 1$, then we conclude that the functions in (6.3.36) correspond to the optimal value function $J^*(s,x) = v(s,x)$ and the optimal control policy (or portfolio process) π^*, which is a constant. ◇

6.4 Examples: Discounted Costs

We now specialize the infinite–horizon discounted cost problem in (5.4.6) and Theorem 5.23 to

$$dx(t) = b(x(t), a(t))dt + \sigma(x(t))dw(t), \quad t \geq 0. \tag{6.4.1}$$

In contrast to (6.2.1), the coefficients b and σ are time–invariant, and therefore (6.4.1) is an "autonomous" equation. (Observe that σ in (6.4.1) is independent of the controls $a \in A$.) The cost rate $c(s, x, a)$ is also time–invariant, $c(s, x, a) = c(x, a)$, and so the meaning of the notation b^π, σ^π and c^π in Definition 6.8(a) and (6.2.3) is as in (5.3.3):

$$b^\pi(t, x) := b(x, \pi(t, x)), \quad \sigma^\pi(x) := \sigma(x), \quad c^\pi(t, x) := c(x, \pi(t, x)).$$

If π is *stationary*, we use the notation (5.3.4): $b^\pi(x) = b(x, \pi)$, $c^\pi(x) = c(x, \pi)$.

For a function $v(s, x) = v(x)$ of class C^2 in $x \in \mathbb{R}^d$ (assuming that it satisfies the assumptions of Theorem 6.5), the expression (6.2.2) for the generator L^a reduces to

$$L^a v(x) = v_x(x)b(x, a) + \frac{1}{2}Tr[v_{xx}(x)D(x)], \tag{6.4.2}$$

with $D(x) = \sigma(x)\sigma(x)^*$. Let $V(x, \pi)$ be as in (5.4.6) but in the time–homogeneous case:

$$V(x, \pi) := E_x^\pi \int_0^\infty e^{-\rho t} c^\pi(t, x(t))dt$$

where π is an admissible policy in the sense of Definition 6.8. Moreover, let V^* be the value function

$$V^*(x) := \inf_\pi V(x, \pi) \quad \forall\, x \in \mathbb{R}^d.$$

Then, under Assumption 5.22, $V^*(x) < \infty$ for every $x \in X$ and the DP equation is given by (5.4.7), with L^a as in (6.4.2), i.e.,

$$-\rho v(x)+\frac{1}{2}Tr[v_{xx}(x)D(x)]+\min_{a\in A}[c(x,a)+v_x(x)b(x,a)]=0. \tag{6.4.3}$$

Definition 6.12. Let $v\in\mathcal{D}$ be a solution of (6.4.3). We denote by Π_D the class of *stationary* policies $\pi\in\Pi$ for which the following condition

$$\lim_{t\to\infty} e^{-\rho t}E_x^{\pi}v(x(t))=0 \quad \forall\, x\in\mathbb{R}^d \tag{6.4.4}$$

is satisfied.

Note that (6.4.4) is a time–homogeneous version of (5.4.8).

Remark 6.13. By Theorem 5.23(b), if $\pi^*\in\Pi_D$ attains the minimum in (6.4.3), that is

$$c(x,\pi^*)+v_x(x)b(x,\pi^*)=\min_{a\in A}[c(x,a)+v_x(x)b(x,a)] \quad \forall x,$$

then π^* is ρ–discount optimal in the class Π_D. ◇

Example 6.14. (LQ systems). Consider the time–invariant scalar linear system [see (6.3.6)]

$$dx(t)=[\gamma x(t)+\beta a(t)]dt+\sigma dw(t), \quad t\geq 0, \tag{6.4.5}$$

with constant coefficients γ, β, σ $(\beta\neq 0)$, and the ρ–discounted cost functional

$$V(x,\pi):=E_x^{\pi}\int_0^{\infty} e^{-\rho t}(qx^2(t)+ra^2(t))dt, \tag{6.4.6}$$

where $q\geq 0$, $r>0$. Thus, using the notation $v_x=v'$ and $v_{xx}=v''$, the DP Eq. (6.4.3) becomes

$$-\rho v+\frac{1}{2}\sigma^2 v''+\min_a[qx^2+ra^2+(\gamma x+\beta a)v']=0, \tag{6.4.7}$$

where the minimum is over $A=\mathbb{R}$. As in (6.3.10), we will try to solve (6.4.7) with a function of the form

$$v(x)=kx^2+g \quad \forall\, x\in\mathbb{R}; \quad k \text{ and } g \text{ constants.} \tag{6.4.8}$$

For this function v, (6.4.7) becomes

$$-\rho(kx^2+g)+\sigma^2 k+\min_a[qx^2+ra^2+2(\gamma x+\beta a)kx]=0 \tag{6.4.9}$$

and the minimum is attained at $a=\pi^*$ given by

$$\pi^*(x)=-r^{-1}\beta kx, \quad x\in\mathbb{R}. \tag{6.4.10}$$

Inserting this value of $a=\pi^*$ in (6.4.9), we obtain

$$[q+(2\gamma-\rho)k-r^{-1}\beta^2k^2]x^2+(k\sigma^2-\rho g)$$

which is zero for all $x\in\mathbb{R}$ if

$$g=k\sigma^2/\rho,$$

and k satisfies the equation

$$q+(2\gamma-\rho)k-r^{-1}\beta^2k^2=0. \tag{6.4.11}$$

Assuming that $q>0$, the latter equation has a unique positive solution. Thus the function $v(x)$ in (6.4.8) is given by

$$v(x)=kx^2+k\sigma^2/\rho, \quad x\in\mathbb{R}, \tag{6.4.12}$$

where k is the unique positive solution to (6.4.11).

Thus to conclude that π^* in (6.4.10) is optimal in the sense of Remark 6.13, it only remains to verify that π^* satisfies (6.4.4). To do this, in (6.4.5) take $a(t)=\pi^*(x(t))$, to obtain

$$dx(t)=-\alpha x(t)dt+\sigma dw(t), \quad x(0)=x, \tag{6.4.13}$$

with $\alpha=r^{-1}\beta^2k-\gamma$, which is the so–called *Langevin equation*. Thus, as is well known (see, for instance, Arnold (1974) Sect. 8.3, or Øksendal (2003) p. 75) the solution of (6.4.13) is

$$x(t)=xe^{-\alpha t}+\sigma\int_0^t e^{-\alpha(t-s)}dw(s) \quad \text{for} \quad t\geq 0.$$

Therefore, by the properties of stochastic integrals,

$$E_x^{\pi^*}[e^{-\rho t}x^2(t)]=x^2e^{-(\rho+2\alpha)t}+\sigma^2e^{-\rho t}E_x^{\pi^*}\left(\int_0^t e^{-\alpha(t-s)}dw(s)\right)^2$$

$$= (x^2 - \sigma^2/2\alpha)e^{-(\rho+2\alpha)t} + \sigma^2 e^{-\rho t}/2\alpha. \tag{6.4.14}$$

Finally, from (6.4.11) and the definition of α, we have

$$\rho + 2\alpha = [(2\gamma - \rho)^2 + 4r^{-1}\beta^2 q] > 0.$$

Therefore,

$$\lim_{t\to\infty} E_x^{\pi^*}[e^{-\rho t}x^2(t)] = 0, \tag{6.4.15}$$

which in turn, by (6.4.12), implies (6.4.4) for $\pi = \pi^*$. Hence from the Remark 6.13 we conclude that π^* minimizes (6.4.6) within the class of admissible stationary policies π that satisfy

$$\lim_{t\to\infty} E_x^{\pi}[e^{-\rho t}x^2(t)] = 0, \tag{6.4.16}$$

and the value (or minimum) cost function is $v(\cdot)\equiv v_\rho(\cdot)$ in (6.4.12), where k is the unique positive solution of the quadratic equation (6.4.11). ◇

Remark. The same argument leading from (6.4.13) to (6.4.14) shows that (6.4.15) holds for every policy of the form $\pi(x) = -Gx$ if G is a constant such that $\rho + 2(\beta G - \gamma) > 0$, that is, $G > (\gamma - \rho/2)/\beta$.

The vector case. The above results are easily extended to n–dimensional systems (6.4.5) with q and r in (6.4.6) being symmetric matrices, q nonnegative definite, and r positive definite. In this case, (6.4.8)=(6.4.12) and (6.4.10) result

$$v(x) = x^* kx + g, \quad \text{and} \quad \pi^*(x) = -r^{-1}\beta^* kx,$$

where g is a constant, and k is the symmetric and positive definite solution of a matrix equation corresponding to (6.4.11). ◇

Example 6.15. (Maximization of total discounted utility from consumption). The problem is to choose a consumption process $a(t) = \pi(x(t))$ to *maximize* the expected total discounted utility

$$V(x,\pi) = E_x^{\pi}\int_0^{\infty} e^{-\rho t}U[a(t)]dt, \quad \rho > 0, \tag{6.4.17}$$

where U is a utility function, and the wealth $x(\cdot) = x^{\pi}(\cdot)$ satisfies, for $t \geq 0$,

$$dx(t) = x(t)[\alpha dt + \sigma dw(t)] - a(t)dt, \quad x(0) = x; \tag{6.4.18}$$

[see (6.3.23)–(6.3.24) with $a_1(\cdot) \equiv 1$ and $a_2(\cdot) = a(\cdot)$]. In (6.4.18), which can also be written as

$$dx(t) = [\alpha x(t) - a(t)]dt + \sigma x(t)dw(t), \quad x(0) = x > 0; \tag{6.4.19}$$

we assume that $\alpha > 1$, $\sigma^2 > 0$, and the initial wealth x is positive, whereas $a(t) = \pi(x(t))$ is subject to the constraint

$$0 \leq \pi(x) \leq x \quad \forall x. \tag{6.4.20}$$

Then for a general utility function U, the DP equation (6.4.3) becomes (with $v_x = v'$ and $v_{xx} = v''$)

$$-\rho v(x) + \frac{1}{2}\sigma^2 x^2 v''(x) + \alpha x v'(x) + \max_a [U(a) - a v'(x)] = 0. \tag{6.4.21}$$

We will solve (6.4.21) for the particular utility function $U(a) = a^{\gamma}/\gamma$, with $0 < \gamma < 1$, and try a solution of the form [see (6.3.28)]

$$v(x) = hx^{\gamma}, \quad h > 0. \tag{6.4.22}$$

In this case, the function $U(a) - av'(x)$ is maximized when $a = \pi^*(x)$ is given by

$$\pi^*(x) = (\gamma h)^{-1/(1-\gamma)} x \quad \forall x \geq 0. \tag{6.4.23}$$

Inserting this value of $a = \pi^*(x)$ in (6.4.21) gives the following equation for h:

$$-[\rho + \sigma^2 \gamma(1-\gamma)/2 - \alpha\gamma] + (1-\gamma)(\gamma h)^{-1/(1-\gamma)} = 0.$$

Therefore

$$h = \gamma^{-1}\theta^{-(1-\gamma)}, \quad \text{with} \quad \theta := [\rho + \sigma^2\gamma(1-\gamma)/2 - \alpha\gamma]/(1-\gamma), \tag{6.4.24}$$

and, from (6.4.23),

$$\pi^*(x) = \theta x. \tag{6.4.25}$$

This is the optimal consumption process provided that it satisfies (6.4.20), so that we must have

$$0 < \theta \leq 1, \tag{6.4.26}$$

and provided also that the condition (6.4.4) holds.

To verify (6.4.4), let us apply Itô's differential rule (Theorem 6.4) to the process $y(t) = \log x(t)$, with $x(\cdot) = x^{\pi^*}(\cdot)$ in (6.4.19), to obtain

$$dy(t) = \left(\alpha - \theta - \frac{1}{2}\sigma^2\right) dt + \sigma dw(t).$$

This equation, in integral form, yields

$$\log[x(t)/x] = \left(\alpha - \theta - \frac{1}{2}\sigma^2\right) t + \sigma w(t);$$

that is,

$$x(t) = x \exp\left[\left(\alpha - \theta - \frac{1}{2}\sigma^2\right) t + \sigma w(t)\right]. \tag{6.4.27}$$

Therefore, since $w(t)$ is a Gaussian variable $N(0, t)$ (see Example 5.3), from (6.4.22) we obtain

$$\begin{aligned} E_x^{\pi^*}[e^{-\rho t} v(x(t))] &= h x^\gamma \exp\left(-\left[\rho + \frac{1}{2}\sigma^2\gamma(1-\gamma) - \alpha\gamma + \theta\gamma\right] t\right) \\ &= h x^\gamma e^{-\theta t} \quad [\text{from (6.4.24)}] \\ &\to 0 \quad \text{as} \quad t \to \infty, \end{aligned}$$

provided that (6.4.26) holds. Thus from Remark 6.13 we conclude that, assuming (6.4.26), the consumption process in (6.4.23) is optimal within the class of admissible stationary policies π for which

$$E_x^\pi[e^{-\rho t} v(x(t))] = h e^{-\rho t} E_x^\pi[x^\gamma(t)] \to 0 \quad \text{as} \quad t \to \infty.$$

Moreover, when using π^*, the corresponding wealth is the log–normal process in (6.4.27), and the optimal expected discounted utility is given by (6.4.22) and (6.4.24) for any initial wealth $x > 0$. ◇

6.5 Examples: Average Costs

Let us consider again the autonomous SDE (6.4.1) with generator L^a in (6.4.2). As in Sect. 5.5, we wish to minimize the long–run expected average cost (AC)

$$J(x,\pi) = \limsup_{t\to\infty} J_t(x,\pi)/t, \tag{6.5.1}$$

with $J_t(x,\pi)$ as in (5.5.1). As in Assumption 5.24, we suppose that there exists a control policy $\pi \in \Pi$ such that $J(x,\pi) < \infty$ for every $x \in \mathbb{R}^d$, where Π is the set of *admissible* policies in Definition 6.8. Finally, observe that, from (6.4.2), the *AC optimality equation* (ACOE) in Definition 5.27 can be expressed as

$$j^* = \inf_{a\in A}[c(x,a) + h_x(x)b(x,a)] + \frac{1}{2}Tr[h_{xx}(x)D(x)] \tag{6.5.2}$$

in terms of a solution pair $(j^*, h(\cdot))$.

Example 6.16. (LQ problems). This example is related to the *deterministic* LQ problem in Example 4.22. First, we consider a stochastic version and then we show how it reduces to the deterministic case.

(a) **The stochastic case.** Consider the LQ system in Example 6.14, with state equation (6.4.5), with $\beta \neq 0$. The cost J_t in (6.5.1) is

$$J_t(x,\pi) = E_x^\pi \int_0^t c(x(t),a(t))ds,$$

with quadratic instantaneous (or running) cost $c(x,a) := qx^2 + ra^2$, where both q and r are positive numbers. For notational ease, we will write the derivatives h_x and h_{xx} in the ACOE (6.5.2) as h' and h'', respectively. Hence, (6.5.2) becomes

$$j^* = \inf_{a\in\mathbb{R}}[qx^2 + ra^2 + (\gamma x + \beta a)h'(x)] + \frac{1}{2}\sigma^2 h"(x). \tag{6.5.3}$$

Note that this equation is similar to (6.4.7). Therefore, as a first "guess", we may try to solve (6.5.3) with a function of

the form (6.4.8), i.e.,

$$h(x) = kx^2 + g, \ x \in \mathbb{R}, \tag{6.5.4}$$

for some constants k and g. Observe, however, that *these constants are NOT the same as in* (6.4.8), *and* (6.4.12), because the latter equations depend on the discount factor ρ. Thus, to avoid confusions, we will rewrite (6.4.12) as

$$v_\rho(x) = k(\rho)x^2 + k(\rho)\sigma^2/\rho, \tag{6.5.5}$$

where $k(\rho)$ is the unique positive solution of (6.4.11).

Inserting $h(\cdot)$ in (6.5.3), the same calculations used in (6.4.9)–(6.4.11) now show that the minimum in (6.5.3) is attained at $\bar{a} = \bar{\pi}(x)$ given by

$$\bar{\pi}(x) = -r^{-1}\beta k x \quad \forall\, x \in \mathbb{R}. \tag{6.5.6}$$

With this value of $a = \bar{a}$, (6.5.3) becomes

$$j^* = \sigma^2 k + x^2(q + 2\gamma k - r^{-1}\beta^2 k^2) \tag{6.5.7}$$

for all $x \in \mathbb{R}$. Therefore (recalling that $q > 0$), taking k^* as the unique positive solution of the quadratic equation

$$q + 2\gamma k - r^{-1}\beta^2 k^2 = 0, \tag{6.5.8}$$

we obtain

$$\bar{\pi}(x) = -r^{-1}\beta k^* x. \tag{6.5.9}$$

Thus, from (6.5.7) and (6.5.4), we have that the pair $(j^*, h(\cdot))$ consisting of

$$j^* = \sigma^2 k^*, \quad \text{and} \quad h(x) = k^* x^2 + g \quad \forall\, x, \tag{6.5.10}$$

where g is an arbitrary constant, is a *solution to the ACOE* (6.5.3).

We now wish to show, using Theorem 5.28(b), that $\bar{\pi}$ is AC–optimal in the family Π_S^{AC} of stationary policies that satisfy (5.5.9), which in our present situation becomes

$$\lim_{t\to\infty} t^{-1} E_x^{\bar{\pi}} h(x(t)) = 0 \quad \forall\, x.$$

This condition, by the definition of h in (6.5.10), is equivalent to

$$\lim_{t\to\infty} E_x^{\bar{\pi}}[x^2(t)]/t = 0 \quad \forall\, x. \tag{6.5.11}$$

To prove this, note that inserting $a(t) = \bar{\pi}(x(t))$ in the linear equation (6.4.5) we obtain again a *Langevin equation* (6.4.13) but now with coefficient $\alpha := r^{-1}\beta^2 k^* - \gamma$. Hence, from (6.4.14)–(6.4.15) with $\pi^* = \bar{\pi}$ and $\rho = 0$, we have

$$E_x^{\bar{\pi}}[x^2(t)] = (x^2 - \sigma^2/2\alpha)e^{-2\alpha t} + \sigma^2/2\alpha. \tag{6.5.12}$$

This clearly yields (6.5.11) since, from (6.5.8), $\alpha = (\gamma^2 + \beta^2 q/r)^1/2) > 0$. Therefore, from Theorem 5.28(b) we conclude the desired result: $\bar{\pi}$ is AC–optimal in Π_S^{AC}, and the minimum average cost is j^* in (6.5.10).

(b) **The deterministic LQ case.** The AC results for the deterministic LQ system are obtained by taking the coefficient $\sigma = 0$ in (6.4.5) and "everywhere" in part (a) above; in particular, the state Eq. (6.4.5) is now

$$\dot{x}(t) = \gamma x(t) + \beta a(t), t \geq 0, \tag{6.5.13}$$

with constant coefficients γ, β, with $\beta = 0$. Moreover, from (6.5.7)–(6.5.10) with $\sigma = 0$, we conclude that the optimal average cost is $j^* = 0$, and *the AC–optimal policy is again* $\bar{\pi}$ in (6.5.9). ◇

Remark 6.17. (The certainty–equivalence principle). Consider an stochastic optimal control problem (OCP) with state equation as in, say, (6.2.1):

$$dx(t) = b(t, x(t), a(t))dt + \sigma(t, x(t), a(t))dw(t).$$

Consider also an associated *deterministic* OCP in which the corresponding state equation is obtained from (6.2.1) taking $\sigma(\cdot) \equiv 0$. If the optimal control policy in the stochastic case is the same as in the associated deterministic problem, it is then said that the stochastic system satisfies the certainty–equivalence

principle. From the Example 6.16(a), (b) we can see that this principle is satisfied by the average cost problem for LQ systems. This fact is also true in the discounted cost case. *Can you guess why*? ◇

In the following example we wish to *maximize* the long-run average reward (AR) defined as

$$J_{AR}(x, a(\cdot)) := \liminf_{T\to\infty} j_T(x, a(\cdot))/T, \tag{6.5.14}$$

where, given a control policy $a(\cdot)$ and the running (or instantaneous) reward function $r(x, a)$,

$$j_T(x, a(\cdot)) := E_x[\int_0^T r(x(t), a(t))ds],$$

given the one-dimensional Eq. (6.4.1). In this case, the optimality Eq. (6.5.2) becomes the *average reward optimality equation* (AROE)

$$j^* = \sup_{a\in A}[r(x, a) + h'(x)b(x, a) + \frac{1}{2}h'(x)\sigma^2(x)], \tag{6.5.15}$$

where $j^* = \sup_{a(\cdot)} J_{AR}(x, a(\cdot))$, and h', h'' denote the first and second derivatives of h with respect to x.

Example 6.18. (Average welfare in a pollution accumulation model). This example is a particular case of the control problem in Kawaguchi and Morimoto (2007) or Sect. 10.1 in Morimoto (2010), and it is also an extension to controlled diffusion processes of the *deterministic* Example 4.29. Now, instead of (4.4.28)–(4.4.30), the stock of pollution $x(\cdot)$ evolves according to the one-dimensional stochastic differential equation

$$dx(t) = [a(t) - d_0x(t)]dt + \sigma \cdot x(t)dw(t), x(t) = x > 0, \tag{6.5.16}$$

where $a(\cdot)$ denotes the flow of consumption (or pollution), d_0>0 is the constant rate of pollution decay, and $\sigma > 0$ is a given constant.

The long-run average reward (or average welfare) is as in (6.5.14) with instantaneous reward $r(x, a) = U(a) - D(x)$, where

$U(a)$ is the social utility function of the consumption a, and $D(x)$ is the social disutility of the pollution stock x. Kawaguchi and Morimoto (2007) consider general utility and disutility functions. Here, however, as in (4.4.30), we will assume that U and D are of the particular form

$$U(a) = 2a^{1/2} \quad \text{and} \quad D(x) = d_1 x,\ d > 0,$$

in which case the AROE (6.5.15) becomes

$$j^* = \sup_{a \geq 0}[2a^{1/2} - d_1 x + h'(x)(a - d_0 x)] + \frac{1}{2}h''(x)\sigma^2 x^2. \quad (6.5.17)$$

Note that the right-hand side of (6.5.17) is maximized at $a = a^*(x) = 1/(h'(x))^2$. The problem then is how to find a suitable function h.

Observe that the drift coefficient $b(x, a) = a - d_0 x$ in (6.5.16) is the same as the right-hand side of (4.4.29) with $\varphi(x)$ as in (4.4.30). Therefore, in view of the results in Example 4.29, we conjecture that h is of the form $h(x) = -h_1 x + h_2$ for some constants h_1, h_2 that need to be determined. Replacing this function h in (6.5.17) and using that $a^*(x) = 1/(h'(x))^2 = 1/h_1^2$, (6.5.17) reduces to

$$j^* = \frac{1}{h_1} + (h_1 d_0 - d_1)x \quad \forall x \geq 0.$$

Therefore, $h_1 = d_1/d_0$ and $j^* = 1/h_1 = d_0/d_1$. This gives a pair $(j^*, h(\cdot))$ that satisfies (6.5.16). However, to conclude from Theorem 5.28 that j^* has some optimality property we still need to verify (5.5.9) for some family of policies $\pi \in \Pi$, so that

$$\frac{1}{t}E_x^\pi h(x(t)) \to 0 \quad \text{as } t \to \infty. \quad (6.5.18)$$

To this end, in (6.5.16) take $a(\cdot) \equiv a^* := (d_0/d_1)^2$. Then for any initial state $x(0) = x > 0$ and any positive integer k such that

$$2d_0 > (k-1)\sigma^2 \quad (6.5.19)$$

the solution $x(\cdot)$ of (6.5.16) satisfies that

$$\sup_{t\geq 0} E[x(t)^k] \leq x^k + C$$

for some constant $C > 0$. (See Morimoto (2010), Lemma 10.2.1 or Proposition 10.2.2.) The latter inequality obviously yields (6.5.18) provided that the coefficients d_0 and σ satisfy (6.5.19).

To conclude, note that the optimal control $a^*(x) = 1/(h'(x))^2$ in this example is the same as the optimal control a^* in the *deterministic* control problem of Example 4.29. Hence we have another example of the *certainty-equivalence principle* in Remark 6.17. *Can you explain why?* ◇

Notes—Chapter 6

1. The setting in this chapter—imposing Assumptions 6.1, 6.2, and the notion of admissibility in Definition 6.8—is very restrictive, but it suffices to analyze some stochastic control problems by means of concepts from undergraduate calculus, such as continuity, differentiability, and so forth. The setting can be considerably relaxed but at the cost of introducing nontrivial mathematical complications.

To illustrate this, consider the following innocent–looking one–dimensional control problem in which the system Eq. (6.3.6) and the cost functional (6.3.7) are of the form

$$dx(t) = a(t)dt + dw(t), \quad 0 \leq t \leq 1, \tag{6.5.20}$$

$$J(0, x, \pi) = E_x^\pi[x(1)^2],$$

respectively, with control set $A = [-1, 1]$. Then the optimal control is

$$\pi^*(s, x) = -\text{sign}(x)$$

[see Beneš (1974), or Christopeit and Helmes (1982)], which is *not* an admissible Markov policy in the sense of Definition 6.8. In fact, if we write $a(t) = \pi^*(t, x(t))$ in (6.5.20), the resulting equation does not satisfy the Assumption 6.1(b), so in our context we cannot claim that the equation has a "solution". This kind of situations can be dealt with in a number of ways (typically, by "weakening" the notion of solution of a SDE); see Fleming and Soner

(2006), Hanson (2007), Morimoto (2010), Pham (2009), Yong and Zhou (1999),...

2. The examples in Sects. 6.3 and 6.4 are standard, see e.g. Fleming and Rishel (1975) Chap. 6, or Øksendal (2003) Chap. 11. The Example 6.10 is originally due to Merton (1971). For related applications and further references on financial economics see, for instance, Chang (2004), Karatzas (1989), Karatzas and Shreve (1998), Merton (1990), Morimoto (2010), Pham (2009).

3. The control problems in which the coefficients b and σ of (6.2.1) are of the form $b(t,x,a) = a$, $\sigma(t,x,a) = \sigma(t,x)$, and $A = \mathbb{R}^m$ (which is the case in (6.5.20)), are called *stochastic calculus of variations* problems: Fleming (1983), Loewen (1987).

4. For the existence of solutions to the DP Eq. (6.4.3) see e.g. Bensoussan (1982), Morimoto (2010), Pham (2009), Krylov (1980).

5. The Example 6.15 is due to Merton (1971). For other discounted problems in economics and finance see the references in Note 2 above. For other applications see, for instance, Mangel (1985), and Whittle (1982).

Appendix A

Terminology and Notation

For technical reasons, all the sets and functions considered in these lecture notes are assumed to be *Borel measurable*. If the reader is not familiar with this concept, don't worry: we only consider nice sets and functions, for instance, open sets and continuous (or even differentiable) functions. It is important, however, to know at least the following basic terminology.

Let X be a metric space. The *Borel sigma–algebra* of X, denoted by $\mathcal{B}(X)$, is the smallest sigma–algebra that contains all the open subsets of X. The sets in $\mathcal{B}(X)$ are called *Borel sets*.

If X is a complete and separable metric space (also known as a *Polish space*), then a Borel subset of X is called a *Borel space*. The following are examples of Borel spaces:

- Any open or any closed subset of $\mathbb{R}^n$.
- A *discrete space* X, that is, a finite or denumerable set with the discrete topology (the topology consisting of all the subsets of X).
- A compact metric space (which is complete and separable).
- If $X_1, X_2, \ldots$ is a (finite or countable) sequence of Borel spaces, then the product space $Y := X_1 \times X_2 \times \cdots$ is also a Borel space with the (product) Borel sigma–algebra.
- If X is a Borel space, then the space $\mathbb{P}(X)$ of probability measures on X with the topology of weak convergence is also a Borel space.

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*,
Texts in Applied Mathematics 76, https://doi.org/10.1007/978-3-031-21139-3

For further details on measurability and related concepts, see any introductory book on Real Analysis, for instance, Ash (1972), Bartle (1995), Bass (2020), etc.

Lower Semicontinuous Functions

Let X be a metric space and v a function from X to $\mathbb{R} \cup \{+\infty\}$ such that $v(x) < \infty$ for at least one point $x \in X$. This function v is said to be *lower semicontinuous* (l.s.c.) at $x \in X$ if

$$\liminf_{n\to\infty} v(x_n) \geq v(x)$$

for any sequence $\{x_n\}$ in X that converges to x. The function v is called *lower semicontinuous* (l.s.c.) if it is l.s.c. at every point of X.

Proposition A.1. The following statements are equivalent:

(a) v is l.s.c.;

(b) the set $epi(v) := \{(x, \lambda) \in X \times \mathbb{R} \mid v(x) \leq \lambda\}$, called the *epigraph* of v, is closed;

(c) all of the *lower sections* (or *level sets*) $S_\lambda(v)$ are closed, where

$$S_\lambda(v) := \{x \in X \mid v(x) \leq \lambda\}, \quad \lambda \in \mathbb{R}.$$

Let $L(X)$ be the family of l.s.c. functions on X, and $L^+(X)$ the subfamily of nonnegative l.s.c. functions. (In many applications it suffices to assume that $L^+(X)$ consists of l.s.c. functions that are *bounded below*. Clearly, if v is l.s.c. and $v(\cdot) \geq -m$ for some m, then $v(\cdot) + m$ is in $L^+(X)$.)

Proposition A.2. A function v is in $L^+(X)$ if and only if there exists a sequence of continuous and bounded functions v_n on X such that $v_n \uparrow v$.

Proposition A.3. If $v, v_1, \ldots, v_n$ belong to $L^+(X)$, then

(a) the functions αv, with $\alpha \geq 0$, $v_1 + \ldots + v_n$, and $\min_i v_i$, belong to $L^+(X)$;

(b) if X is compact, then v attains its minimum, that is, there exists a point $x^* \in X$ such that $v(x^*) = \inf_x v(x)$.

For proofs of Proposition A.1–A.3 see Ash (1972), Appendix A6 or Bertsekas and Shreve (1978), Sect. 7.5.

On the other hand, v is *upper semicontinuous* (u.s.c.) if and only if $-v$ is l.s.c. Moreover, v is *continuous* if and only if v is both l.s.c. and u.s.c.

Appendix B

Existence of Measurable Minimizers

In this appendix X and A denote Borel spaces. In the previous chapters, they denote the *state space* and the *action space* (or *control set*) of an OCP, respectively.

Let 2^A denote the family of all nonempty subsets of A. A set–valued mapping $\Phi : X \to 2^A$ is called a *multifunction*, also known as a *correspondence*. In this case, a (measurable) function $f : X \to A$ such that $f(x) \in \Phi(x)$ for all $x \in X$ is said to be a *selector* (or *selection*) of Φ. We denote by $\mathbb{F}$ the family of selectors of Φ.

Sometimes we write $\Phi(x)$ as $A(x)$.

Definition B.1. (a) The *lower inverse* of $B \subset A$ with respect to Φ is defined as

$$\Phi^{-1}[B] := \{x \in X : A(x) \cap B \neq \emptyset\}.$$

(b) The *graph* of Φ, which we will denote by $\mathbb{K}$, is given by

$$\mathbb{K} := \{(x, a) \in X \times A : a \in A(x)\}.$$

(c) Φ is said to be *Borel measurable* if the lower inverse $\Phi^{-1}[B]$ is a Borel subset of X for every closed set $B \subset A$.

Theorems B.2 and B.3, below, are obtained in Himmelberg et al. (1976) and in Schäl (1975); they are also reproduced in Appendix D of Hernández-Lerma and Lasserre (1996).

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*,
Texts in Applied Mathematics 76, https://doi.org/10.1007/978-3-031-21139-3

Theorem B.2. *Suppose that Φ is compact–valued, that is, $\Phi(x)$ is compact for every $x \in X$. Then the following statements are equivalent:*

(a) *Φ is Borel measurable.*

(b) *$\Phi^{-1}[B] \subset X$ is a Borel set for every open set $B \subset A$.*

(c) *The graph of Φ is a Borel subset of $X \times A$.*

Theorem B.3. *Suppose that Φ is compact–valued. Let $v : \mathbb{K} \to \mathbb{R}$ be a Borel function such that $v(x, \cdot)$ is lower semicontinuous (l.s.c) on $\Phi(x)$ for every $x \in X$. Then:*

(a) there exists a (Borel measurable) selector $f^* \in \mathbb{F}$ such that

$$v(x, f^*(x)) = v^*(x) := \min_{a \in A(x)} v(x, a) \quad \forall\, x \in X, \tag{B.1}$$

and v^* is measurable.

(b) If the set-valued mapping $x \mapsto A(x)$ is u.s.c. and v is l.s.c. and bounded below, then there exists $f^* \in \mathbb{F}$ that satisfies (B.1) and, furthermore, v^* is l.s.c. and bounded below.

Theorems B.2 and B.3 require Φ to be compact–valued. In contrast, Theorem B.8 below does not require compactness, but we need the following concepts.

Definition B.4. (a) Consider a function $v : \mathbb{K} \to \mathbb{R}$.

(a1) v is called *inf–compact* if, for every $r \in \mathbb{R}$, the set

$$\{(x, a) \in \mathbb{K} | v(x, a) \leq r\}$$

is compact;

(a2) v is called $\mathbb{K}$*–inf–compact* if, for every compact set $X' \subset X$ and every $r \in \mathbb{R}$, the set

$$\{(x, a) \in G(X') | v(x, a) \leq r\}$$

is compact, where $G(X') := \{(x, a) \in X' \times A | a \in A(x)\}$;

(a3) v is called *inf–compact on* $\mathbb{K}$ if, for every $x \in X$, the function $a \mapsto v(x, a)$ is inf–compact on $A(x)$, that is, the set

$$\{a \in A(x) | v(x, a) \leq r\} \subset A$$

is compact for every $r \in \mathbb{R}$.

(b) Consider a compact-valued multifunction $\Phi : X \to 2^A$ as above, with A a separable metric space. Then Φ is said to be:

(b1) *lower semicontinuous* (l.s.c.) at $x \in X$ if it satisfies that: If $x_n \to x$ and a is in $A(x)$, then there exist $a_n \in A(x_n)$ such that $a_n \to a$. Φ is called l.s.c. on X if it is l.s.c. at every $x \in X$.

(b2) *upper semicontinuous* (u.s.c) at $x \in X$ if it satisfies that: If $x_n \to x$ and $a_n \in A(x_n)$ is such that $a_n \to a$, then a is in $A(x)$. Φ is u.s.c. on X if it is u.s.c. at every $x \in X$.

(b3) *continuous* on X if it is both l.s.c and u.s.c. on X.

Remark B.5. (a) For calculations, it is convenient to restate Definition B.4(a2) as follows (see Feinberg et al. 2021). A function $v : \mathbb{K} \to \mathbb{R}$ is $\mathbb{K}$-inf-compact if and only if for every sequence $\{(x_t, a_t)\}$ in $\mathbb{K}$ such that, for some $x \in X$, $x_t \to x$ and $v(x_t, a_t)$ is bounded above, it holds that the sequence $\{a_t\}$ has an accumulation point $a \in A(x)$.

(b) As a simple example of Definition B.4(b), consider the spaces $X = A = \mathbb{R}$. Then the multifunction

$$\Phi_1(x) := \begin{cases} [0,1] & \text{if } x \neq 0 \\ [0,1/2] & \text{if } x = 0 \end{cases}$$

is l.s.c. On the other hand, the multifunction

$$\Phi_2(x) := \begin{cases} [0,1] & \text{if } x \neq 0 \\ [0,2] & \text{if } x = 0 \end{cases}$$

is u.s.c., whereas the "constant" multifunction $\Phi(x) := [0,1]$ for all $x \in X$ is continuous. ◊

Lemma B.6. *For v as in Definition B.4 (a): (a1) ⇒ (a2) ⇒ (a3).*

To verify that a multifunction is l.s.c., the following proposition from Michael (1970) is useful.

Proposition B.7. (Michael, 1970). Consider a multifunction $\Phi : X \to 2^A$. If for every $x \in X$ and $a \in \Phi(x)$ there is a continuous selector f of Φ such that $f(x) = a$, then Φ is l.s.c.

Theorem B.8. *Let us suppose that $\mathbb{K}$ is a Borel subset of $X \times A$, v is l.s.c., bounded below, and inf-compact on $\mathbb{K}$. Then*

(a) *There exists a Borel selector $f^* \in \mathbb{F}$ for which (B.1) holds.*

(b) *If, in addition, the multifunction $x \mapsto A^*(x)$, where*

$$A^*(x) := \{a \in A(x) : v^*(x) = v(x, a)\}$$

is l.s.c., then v^* is l.s.c. If, moreover, v is continuous, then so is v^*.

Proof. For part (a) see Rieder (1978); for (b) see Hernández-Lerma and Runggaldier (1994). □

The conclusions (a) and (b) in Theorem B.8 can be obtained in several ways. For instance, from our Definition B.4(a) above and Feinberg et al. (2013) we obtain the following.

Theorem B.9. *Let v be a real–valued function on $\mathbb{K}$ (with $\mathbb{K}$ as in Definition B.1(b)). If v is $\mathbb{K}$–inf–compact (Definition B.4(a2)), then there exists $f^* \in \mathbb{F}$ that satisfies (B.1). Moreover, v is l.s.c. on $\mathbb{K}$, and v^* is l.s.c. on X.*

We conclude this appendix with a useful result from Schäl (1975), Proposition 12.2, which is reproduced as Proposition D.7 in Hernández-Lerma and Lasserre (1996).

Theorem B.10. *Let X be an arbitrary metric space, A a separable metric space, and Φ a compact-valued multifunction from X to 2^A. Let $\mathbb{F}$ be the family of measurable selectors of Φ. In f_n is a sequence in $\mathbb{F}$, then there exists $f \in \mathbb{F}$ such that, for each $x \in X$, $f(x) \in A(x)$ is an accumulation point of $\{f_n(x)\}$. In other words, for each $x \in X$, there is a sequence $n_i = n_i(x)$ such that $f_{n_i}(x) \to f(x)$ as $i \to \infty$.*

If f_n and f satisfy the conclusion of Theorem B.10, then we say that the sequence $\{f_n\}$ *converges in the sense of Schäl to* f.

Remark B.11. Theorem B.10 is useful, for instance, in optimization problems such as the following. Let X, A and Φ be as in Theorem B.10, and consider a sequence of real-valued functions v_n on $X \times A$ such that $v_n \to v^*$. Suppose that, for each n, there exists $f_n \in \mathbb{F}$ such that

$$v_n(x, f_n(x)) = \inf_{a \in A(x)} v_n(x, a) \quad \forall x \in X.$$

Then, by Remark B.11, the sequence of "minimizers" $\{f_n\}$ converges in the sense of Schäl to some $f^* \in \mathbb{F}$. The obvious question now is, is f^* a "minimizer" for v^*? More explicitly, is $f^*(x) \in A(x)$ such that

$$v^*(x, f^*(x)) = \min_{a \in A(x)} v^*(x, a) \quad \forall x \in X?$$

The answer to this question depends on the underlying assumptions. For examples in which the answer is affirmative, see Sect. 2.4 above, or Sect. 4 in Escobedo-Trujillo et al. (2020). Another example is given in the following proposition, which requires A to be a *locally compact* space (that is, for each $a \in A$, there is an open set containing a and such that its closure is compact). For example, Euclidean spaces $\mathbb{R}^d$ are locally compact. ◊

Proposition B.12. (a) Let v and v_n $(n = 1, 2, \ldots)$ be l.s.c. functions, bounded below, and inf-compact on $\mathbb{K}$. Let, for $x \in X$,

$$v_n^*(x) := \min_{a \in A(x)} v_n(x, a), \quad v^*(x) := \min_{a \in A(x)} v(x, a).$$

For each n, let $f_n \in \mathbb{F}$ be a minimizer of v_n, i.e., $v_n^*(x) = v_n(x, f_n(x))$. If A is locally compact and either $v_n \uparrow v$ or $v_n \downarrow v$, then f_n converges in the sense of Schäl (1975) to some $f \in \mathbb{F}$ that is a minimizer of v, i.e., $v^*(x) = v(x, f(x))$ for all $x \in X$.

(b) The conclusion in part (a) is also valid if the Assumption 2.33 (in Sect. 2.3.3 above) holds, the functions v and v_n belong to $L_w(X)$, and $v_n \to v$ as $n \to \infty$.

For a proof of Proposition B.12(a), see Lemma 4.6.6 in Hernández-Lerma and Lasserre (1996). Part (b) follows from Theorem B.10.

Appendix C

Markov Processes

Let $\{x_n\}$ be a discrete–time stochastic process, that is, a sequence of random variables, with values in a space S. We will assume that S is a Borel space.

The process $\{x_n\}$ is called a **Markov chain** or a *discrete–time Markov process* with *state space* S if, for every $n \geq k \geq 0$ and $B \in \mathcal{B}(S)$,

$$P(x_n \in B \mid x_0, \ldots, x_k) = P(x_n \in B \mid x_k). \tag{C.1}$$

The interpretation of (C.1) is as follows. Let us refer to k as the "present (or current) time", $n \geq k$ as the "future", and $n \leq k-1$ as the "past". Then (C.1) states that the distribution of the sequence at any future time n, given the "history" of the process up to the current time k depends only on the current state x_k.

In fact, the so-called *Markov property* (C.1) holds iff (C.1) holds for $n = k+1$ only; that is, (C.1) is equivalent to the following: For every $k \geq 0$ and $B \in \mathcal{B}(S)$,

$$P(x_{k+1} \in B \mid x_0, \ldots, x_k) = P(x_{k+1} \in B \mid x_k). \tag{C.2}$$

In other words, to verify the Markov property it is not necessary to check (C.1) *for every* $n \geq k$; it suffices to check (C.2) for $n = k+1$.

The right–hand side of (C.2) defines the *one–step transition probabilities*

$$P_k(x, B) := P(x_{k+1} \in B \mid x_k = x)$$

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*,
Texts in Applied Mathematics 76, https://doi.org/10.1007/978-3-031-21139-3

for all $x \in S, B \in \mathcal{B}(S)$, and $k = 0, 1, \ldots$. If the transition probabilities are independent of k, that is,

$$P(x, B) \equiv P(x_{k+1} \in B \mid x_k = x) \quad \forall\, k \geq 0,$$

then $\{x_n\}$ is said to be a *stationary* or *time–homogeneous* Markov chain. Here, unless noted otherwise, we will consider *stationary* Markov chains only.

Remark C.1. We will try to explain the origin of the name *Markov chain*. The Russian mathematician Andrei A. Markov (1856–1922) considered a sequence $\{x_k\}$ of random variables with values in a finite set S, and such that

$$\begin{aligned} &P(x_0 = i_0, x_1 = i_1, \ldots, x_n = i_n) \\ &= P(x_0 = i_0) p(i_0, i_1) p(i_1, i_2) \cdots p(i_{n-1}, i_n) \end{aligned} \tag{C.3}$$

for every $n \geq 1$ and every sequence of states $i_0, \ldots, i_n$ in S, where

$$p(i_k, i_{k+1}) := P(x_{k+1} = i_{k+1} | x_k = i_k)$$

denote the one–step transition probabilities. It can be shown (Exercise 2) that (C.2) and (C.3) are equivalent; that is, the finite–valued sequence $\{x_k\}$ is a Markov chain iff (C.3) holds. Moreover, because of the right–hand side of (C.3), Markov referred to $\{x_k\}$ as a sequence whose probabilities were "chained". These facts appeared in a paper by Markov, in 1906. The name "Markov chain" was used for the first time by Bernstein (1927). For additional details on Markov's contributions see the paper by Basharin et al. (2004). ◊

The following definition generalizes the concept of transition probability.

Definition C.2. Let X and Y be Borel spaces. A *stochastic kernel* on X given Y (also known as a *transition probability* from Y to X) is a real–valued function $Q(\cdot|\cdot)$ such that

(a) $B \to Q(B|y)$ is a probability measure on $\mathcal{B}(X)$ for each fixed $y \in Y$, and

(b) $y \to Q(B|y)$ is a measurable function on Y for each fixed Borel set $B \subset X$.

Note that a (Markov) transition probability $P(x, B)$ as above, written in the form $P(B|x)$, is a stochastic kernel with $X = Y$.

Proposition C.3. Let $\{x_n, n = 0, 1, \dots\}$ and $\{\xi_n, n = 0, 1, \dots\}$ be stochastic processes in Borel spaces X and S, respectively. Suppose that $\xi_0, \xi_1, \dots$ are independent, and also independent of x_0. If there is a measurable function $F : X \times S \to X$ such that

$$x_{n+1} = F(x_n, \xi_n) \quad \forall\, n \geq 0, \tag{C.4}$$

then $\{x_n\}$ is a Markov chain. The converse is also true.

Example C.4. Let $\{\xi_n\}$ be a sequence of independent random variables, and x_0 a given random variable independent of $\{\xi_n\}$. By Proposition C.3, the following are examples of Markov chains.

(a) A Markov chain that evolves as

$$x_{n+1} = F(x_n) + \xi_n \quad \forall\, n = 0, 1, \dots \tag{C.5}$$

is called a **first order autoregressive process**, and $\{\xi_n\}$ is said to be an **additive noise**. (C.5) includes **linear systems**

$$x_{n+1} = Gx_n + \xi_n, \tag{C.6}$$

where G is a constant, which can be a matrix in the vector case. If the ξ_n and the initial state x_0 are Gaussian random variables, then (C.6) is called a Gaussian–Markov system.

(b) Consider an **inventory–production system**

$$x_{n+1} = x_n + f(x_n) - \xi_n,\ n = 0, 1, \dots,$$

where x_n is the stock or inventory level (of a certain product) at time n, $f(x)$ is the *production strategy* given the stock level x, and ξ_n is the *demand* of the product in period n.

(c) Consider a water reservoir with capacity Q. Let x_n be the amount of water in the reservoir at time n, and $f(x_n)$ the amount of water discharged in period n (for instance, for irrigation or to produce electrical energy). Then we can express

x_{n+1} as

$$x_{n+1} = \min\left[x_n - f(x_n) + \xi_n, Q\right],$$

where ξ_n is the amount rainwater deposited in the reservoir in period n.

Remark C.5. In Example C.4(a), the sequence $\{\xi_n\}$ is, in general, a *random noise*, that is, a sequence of arbitrary random variables with no particular interpretation or meaning. In contrast, in (b) and (c) the sequence is a *driving process*, that is, the random variables ξ_n have a physical or economic interpretation. Usually, this is also the case in queueing systems, in which the "random perturbations" ξ_n represent the arrival process.

Continuous–Time Markov Processes

Let $\{x_t, t \geq 0\}$ be a continuous–time stochastic process with values in a Borel space X. We say that $\{x_t\}$ is a **Markov process** if, for every $t \geq s \geq 0$ and $B \in \mathcal{B}(X)$,

$$P(x_t \in B \mid x_r \quad \forall\, r \leq s) = P(x_t \in B \mid x_s) \tag{C.7}$$

The interpretation of (C.7) is analogous, of course, to that of (C.1), where s is the "present time", $t > s$ is the "future", and $r < s$ is the "past".

From the right–hand side of (C.7) we obtain the transition probabilities

$$P(s, x, t, B) := P(x_t \in B \mid x_s = x) \tag{C.8}$$

for every $t \geq s, x \in X$, and $B \in \mathcal{B}(X)$. The transition probabilities are called *stationary* or *time–homogeneous* if they depend only on the time–difference $t - s$. In this case we rewrite (C.8) as

$$P(t, x, B) := P(x_t \in B \mid x_0 = x).$$

Example C.6. The simplest example of a continuous–time Markov process is the solution $\{x_t\}$ of an ordinary differential equation

$$\dot{x}_t = F(x_t) \quad \text{for} \quad t \geq 0,$$

for some given initial condition x_0. Under suitable assumptions on F, there is a unique solution

$$x_t = x_0 + \int_0^t F(x_u)du,$$

which can be expressed, for every $t \geq s \geq 0$, as

$$x_t = x_s + \int_s^t F(x_u)du.$$

This is the deterministic analogue of the Markov property (C.7).

Example C.7. [Brownian motion.] The one–dimensional Brownian motion, also known as *Wiener process*, is a process $\{w(t), t \geq 0\}$ with values $w(\cdot) \in \mathbb{R}$ and:

(a) continuous trajectories $t \mapsto w(t)$, with $w(0) = 0$;

(b) *independent increments*, that is, for any positive integer n and times $0 = t_0 < t_1 < \cdots < t_n$, the increments

$$w(t_1) - w(t_0), w(t_2) - w(t_1), ..., w(t_n) - w(t_{n-1})$$

are independent random variables; and

(c) *stationary Gaussian increments*, that is, for any $t > s \geq 0$, the increment $w(t) - w(s)$ has a normal (or Gaussian) distribution with zero mean, and variance $E[(w(t) - w(s))^2] = t - s$.

The Wiener process has many interesting properties. In particular, from the independence of increments (b), it is a continuous–time Markov process.

A process $w(t), t \geq 0$ with values $w(\cdot) = (w_1(\cdot), \ldots, w_n(\cdot)) \in \mathbb{R}^n$ is a n–dimensional Brownian motion (or Wiener process) if the components $w_1(\cdot), \ldots, w_n(\cdot)$ are independent one–dimensional Brownian motions.

Theorem of C. Ionescu–Tulcea

The following proposition in used in the analysis of Markov chains.

Proposition C.8. (Theorem of C. Ionescu Tulcea.). Let X_0, $X_1, \ldots$ be a sequence of Borel spaces and, for $n = 0, 1, \ldots$, define $Y_n := X_0 \times \cdots \times X_n$ and $Y := \Pi_{n=0}^{\infty} X_n$. Let ν be an arbitrary probability measure on X_0 and, for every $n = 0, 1, \ldots$, let $P_n(dx_{n+1}|y_n)$ be a stochastic kernel on X_{n+1} given Y_n. Then there exists a unique probability measure P_ν on Y such that, for every measurable rectangle $B_0 \times \cdots \times B_n$ in Y_n,

$$\begin{aligned} P_\nu(B_0 \times \cdots \times B_n) &= \int_{B_0} \nu(dx_0) \int_{B_1} P_0(dx_1|x_0) \int_{B_2} P_1(dx_2|x_0, x_1) \\ &\cdots \int_{B_n} P_{n-1}(dx_n|x_0, \ldots, x_{n-1}). \end{aligned} \tag{C.9}$$

Moreover, for any nonnegative measurable function u on Y, the function

$$x \mapsto \int u(y) P_x(dy)$$

is measurable on X_0, where P_x stands for P_ν when ν is the probability concentrated at $x \in X_0$.

Proof. See Ash (1972, p. 109), Bertsekas and Shreve (1978, pp. 140–141), or Neveu (1965, p. 162). □

Remark C.9. Let us (informally) write the measure P_ν in (C.9) as

$$P_\nu(dx_0, dx_1, dx_2, \ldots) = \nu(dx_0) P_0(dx_1|x_0) P_1(dx_2|x_0, x_1) \cdots,$$

and let $\pi = \{\pi_t\}$ be an arbitrary control policy. Then the measure P_ν^π in Chap. 3 [see (3.1.10a)–(3.1.10c)] can be written in the form

$$P_\nu^\pi(dx_0, da_0, dx_1, da_1, \ldots) = \nu(dx_0) \pi_0(da_0|x_0) Q(dx_1|x_0, a_0).$$

$$\cdot \pi_1(da_1|x_0, a_0, x_1) Q(dx_2|x_1, a_1) \cdots.$$

Exercises—Appendix C

1. Prove that (C.1) and (C.2) are equivalent.
2. Prove that (C.2) and (C.3) are equivalent.

Bibliography

Adukov, V.M., Adukova, N.V., Kudryavtsev, K.N.: On a discrete model of optimal advertising. J. Comput. Eng. Math. **2**(3), 13–24 (2015)

Alla, A., Falcone, M., Kalise, D.: An efficient policy iteration algorithm for dynamic programming equations. SIAM J. Sci. Comput. **37**(1), A181–A200 (2015)

Anulova, S.V., Mai, H., Veretennikov, A.Y.: On iteration improvement for averaged expected cost control for one-dimensional ergodic diffusions. SIAM J. Control Optim. **58**(4), 2312–2331 (2020)

Arapostathis, A., Borkar, V.S., Ghosh, M.K.: Ergodic Control of Diffusion Processes. Cambridge University Press, Cambridge (2012)

Arisawa, M.: Ergodic problem for the hamilton-jacobi-bellman equation. i. existence of the ergodic attractor. Annales de l'Institut Henri Poincaré C, Analyse Non Linéaire. **14**(4), 415–438 (1997)

Arnold, L.: Stochastic Differential Equations: Theory and Applications. Wiley-Interscience, New York (1974)

Ash, R.B.: Real Analysis and Probability. Academic Press, New York (1972)

Ash, R.B., Gardner, M.F.: Topics in Stochastic Processes. Academic Press, New York (1975)

Bardi, M., Capuzzo-Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations. Systems & Control: Foundations & Applications with Appendices by Maurizio Falcone and Pierpaolo Soravia. Birkhäuser Boston, Inc., Boston, MA (1997)

Bartle, R.G.: The Elements of Integration and Lebesgue Measure. Wiley, New York (1995)

Basharin, G.P., Langville, A.N., Naumov, V.A.: The life and work of AA Markov. Linear Algebra Appl. **386**, 3–26 (2004)

Bass, R.F.: Real analysis for graduate students. Version 4.2, 2020. Available from: https://bass.math.uconn.edu/real.html

Bäuerle, N., Rieder, U.: Markov Decision Processes with Applications to Finance. Springer, Berlin (2011)

Bellman, R.: Dynamic Programming. Princeton University Press, New York (1957)

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*,
Texts in Applied Mathematics 76, https://doi.org/10.1007/978-3-031-21139-3

Bellman, R.: A Markovian decision process. J. Math. Mech. **6**(5), 679–684 (1957)

Bellman, R.: Adaptive Control Processes. Princeton University Press, Princeton, NJ (1961)

Beneš, V.E.: Girsanov functionals and optimal bang-bang laws for final value stochastic control. Stoch. Process. Appl. **2**(2), 127–140 (1974)

Bensoussan, A.: Stochastic Control by Functional Analysis Methods. North-Holland, Amsterdam (1982)

Bernstein, S.: Sur l'extension du théoréme limite du calcul des probabilités aux sommes de quantités dépendantes. Math. Ann. **97**(1), 1–59 (1927)

Bertsekas, D.P.: Dynamic Programming: Deterministic and Stochastic Models. Prentice-Hall, Englewood Cliffs, N.J. (1987)

Bertsekas, D.P.: Dynamic Programming and Optimal Control, vol. 1, 3rd edn. Athena Scientific, Belmont,Massachuts (2005)

Bertsekas, D.P., Shreve, S.: Stochastic Optimal Control: the Discrete-Time Case. Academic Press, New York (1978)

Bishop, C.J., Feinberg, E.A., Zhang, J.: Examples concerning Abel and Cesàro limits. J. Math. Anal. Appl. **420**(2), 1654–1661 (2014)

Blackwell, D.: Discounted dynamic programming. Ann. Math. Stat. **36**(1), 226–235 (1965)

Borkar, V.S., Gaitsgory, V., Shvartsman, I.: LP formulations of discrete time long-run average optimal control problems: The nonergodic case. SIAM J. Control Optim. **57**(3), 1783–1817 (2019)

Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization. Theory and Examples, vol. 3, 2nd edn, of CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York (2006)

Brock, W.A., Mirman, L.J.: Optimal economic growth and uncertainty: the discounted case. J. Econ. Theory **4**, 479–513 (1972)

Brock, W.A., Mirman, L.J.: Optimal economic growth and uncertainty: the no discounting case. Int. Econ. Rev. 560–573 (1973)

Carlson, D.A., Haurie, A.B., Leizarowitz, A.: Infinite Horizon Optimal Control: Deterministic and Stochastic Systems. Springer, Berlin (1991)

Chang, F.-R.: Stochastic Optimization in Continuous Time. Cambridge University Press, Cambridge, UK (2004)

Chow, G.C.: Dynamic Economics: Optimization by the Lagrange Method. Oxford University Press, New York (1997)

Christopeit, N., Helmes, K.: On Beneš'bang-bang control problem. Appl. Math. Optim. **9**(1), 163–176 (1982)

Costa, O.L.V., Dufour, F.: Average control of markov decision processes with feller transition probabilities and general action spaces. J. Math. Anal. Appl. **396**(1), 58–69 (2012)

Cruz-Suárez, H., Montes-de Oca, R.: An envelope theorem and some applications to discounted Markov decision processes. Math. Methods Oper. Res. **67**(2), 299–321 (2008)

Dasgupta, P., Heal, G.: The optimal depletion of exhaustible resources. Rev. Econ. Stud. **41**, 3–28 (1974)

Davis, M.H.A.: Linear Estimation and Stochastic Control. Chapman and Hall, London (1977)

Dockner, E.J., Jorgensen, S., Van Long, N., Sorger, G.: Differential Games in Economics and Management Science. Cambridge University Press, Cambridge, UK (2000)

Domínguez-Corella, A., Hernández-Lerma, O.: The maximum principle for discrete-time control systems and applications to dynamic games. J. Math. Anal. Appl. **475**(1), 253–277 (2019)

Doshi, B.T.: Continuous time control of Markov processes on an arbitrary state space: average return criterion. Stoch. Process. Appl. **4**(1), 55–77 (1976)

Doshi, B.T.: Continuous time control of Markov processes on an arbitrary state space: discounted rewards. Ann. Statist. **4**(6), 1219–1235 (1976)

Doshi, BT.: Generalized semi-Markov decision processes. J. Appl. Probab. 618–630 (1979)

Down, D., Meyn, S.P., Tweedie, R.L.: Exponential and uniform ergodicity of Markov processes. Ann. Probab. **23**(4), 1671–1691 (1995)

Dynkin, E.B.: Markov Processes. Springer, Berlin (1965)

Dynkin, E.B., Yushkevich, A.A.: Controlled Markov Processes. Springer, Berlin (1979)

Escobedo-Trujillo, B., Hernández-Lerma, O., Alaffita-Hernández, F.: Adaptive control of diffusion processes with a discounted reward criterion. Appl. Math. **47**, 225–253 (2020)

Evans, L.C.: An Introduction to Stochastic Differential Equations. American Mathematical Society, Providence, Rhode Island (2013)

Fabbri, G., Gozzi, F., Swiech, A.: Stochastic Optimal Control in Infinite Dimensions: Dynamic Programming and HJB Equations. Springer, Berlin (2017)

Feinberg, E.A., Kasyanov, P.O., Zadoianchuk, N.V.: Average cost markov decision processes with weakly continuous transition probabilities. Math. Oper. Res. **37**(4), 591–607 (2012)

Feinberg, E.A., Kasyanov, P.O., Zadoianchuk, N.V.: Berge's theorem for noncompact image sets. J. Math. Anal. Appl. **397**(1), 255–259 (2013)

Fleming, W.H.: Stochastic calculus of variations and mechanics. J. Optim. Theory Appl. **41**(1), 55–74 (1983)

Fleming, W.H.: Optimal control of Markov processes. In: Proceedings of the International Congress of Mathematicians, August 16–24 1983, vol. I, pp. 71–84. Polish Scientific Publisher and Elsevier, Warsaw (1984)

Fleming, W.H., Rishel, R.W.: Deterministic and Stochastic Optimal Control. Springer, New York (1975)

Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions, 2nd edn. Springer, New York (2006)

Friedman, A.: Stochastic Differential Equations and Applications, vol. I. Academic Press, New York (1975)

Gamkrelidze, R.V.: Discovery of the maximum principle. J. Dyn. Control Syst. **5**(4), 437–451 (1999)

Gihman, I.I., Skorohod, A.V.: Controlled Stochastic Processes. Springer-Verlag, New York (1979)

Guo, X., Hernández-del Valle, A., Hernández-Lerma, O.: Nonstationary discrete-time deterministic and stochastic control systems with infinite horizon. Int. J. Control **83**(9), 1751–1757 (2010)

Guo, X., Hernández-Lerma, O.: Continuous-Time Markov Decision Processes: Theory and Applications. Springer, Berlin (2009)

Halkin, H.: Necessary conditions for optimal control problems with infinite horizons. Econometrica **42**, 267–272 (1974)

Hanson, F.B.: Applied Stochastic Processes and Control for Jump-Diffusions: Modeling. Analysis and Computation. Society for Industrial and Applied Mathematics, Philadelphia (2007)

Hernández-Lerma, O.: Adaptive Markov Control Processes. Springer, New York (1989)

Hernández-Lerma, O.: Lectures on Continuous-Time Markov Control Processes. Sociedad Matemática Mexicana, Mexico City (1994)

Hernández-Lerma, O., Govindan, T.E.: Nonstationary continuous-time Markov control processes with discounted costs on infinite horizon. Acta Appl. Math. **67**(3), 277–293 (2001)

Hernández-Lerma, O., Lasserre, J.B.: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer, New York (1996)

Hernández-Lerma, O., Lasserre, J.B.: Further Topics on Discrete-Time Markov Control Processes. Springer, New York (1999)

Hernández-Lerma, O., Laura-Guarachi Leonardo R., Mendoza-Palacios, S.: A survey of average cost problems in deterministic discrete-time control systems. J. Math. Anal. Appl. 522(1), (2023)

Hernández-Lerma, O., Runggaldier, W.J.: Monotone approximations for convex stochastic control problems. J. Math. Syst. Estimation Control **4**(4), 99–140 (1994)

Himmelberg, C.J., Parthasarathy, T., VanVleck, F.S.: Optimal plans for dynamic programming problems. Math. Oper. Res. **1**(4), 390–394 (1976)

Hinderer, K., Rieder, U., Stieglitz, M.: Dynamic Optimization: Deterministic and Stochastic models. Springer, Cham (2016)

Howard, R.A.: Dynamic Programming and Markov Processes. John Wiley, New York (1960)

Hung, N.M., Quyen, N.V.: Dynamic Timing Decisions Under Uncertainty. Lecture Notes in Economics and Mathematical Systems, vol. 406. Springer-Verlag, Berlin (1994)

Jacka, S.D., Mijatović, A.: On the policy improvement algorithm in continuous time. Stochastics **89**(1), 348–359 (2017)

Josa-Fombellida, R., Rincón-Zapatero, J.P.: Certainty equivalence principle in stochastic differential games: An inverse problem approach. Optimal Control Appl. Methods **40**(3), 545–557 (2019)

Karatzas, I.: Optimization problems in the theory of continuous trading. SIAM J. Control Optim. **27**(6), 1221–1259 (1989)

Karatzas, I., Shreve, S.E.: Methods of Mathematical Finance. Springer, New York (1998)

Kawaguchi, K.: Optimal control of pollution accumulation with long-run average welfare. Environ. Res. Econ. **26**(3), 457–468 (2003)

Kawaguchi, K., Morimoto, H.: Long-run average welfare in a pollution accumulation model. J. Econ. Dyn. Control **31**(2), 703–720 (2007)

Kendrick, D.A.: Stochastic Control for Economic Models, 2nd edn. McGraw Hill, New York (2002)

Krylov, N.V.: Controlled Diffusion Processes. Springer, New York (1980)

Li, X.J., Yong, J.M.: Optimal Control Theory for Infinite-dimensional Systems. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA (1995)

Loewen, P.D.: Existence theory for a stochastic Bolza problem. IMA J. Math. Control Inf. **4**(4), 301–320 (1987)

Long, N.V., Kemp, M.C.: Essays in the Economics of Exhaustible Resources. Elsevier Science Publishers, Amsterdam (1984)

Luenberger, D.G.: Optimization by Vector Space Methods. John Wiley & Sons, New York (1969)

Lund, R.B., Meyn, S.P., Tweedie, R.L.: Computable exponential convergence rates for stochastically ordered Markov processes. Ann. Appl. Probab. **6**(1), 218–237 (1996)

Mangasarian, O.L.: Sufficient conditions for the optimal control of nonlinear systems. SIAM J. Control Optim. **4**(1), 139–152 (1966)

Mangel, M.: Decision and Control in Uncertain Resource Systems. Mathematics in Science and Engineering, 172. Academic Press, Orlando (1985)

Merton, R.C.: Optimum consumption and portfolio rules in a continuous-time model. J. Econ. Theory **3**(4) (1971)

Merton, R.C.: Continuous-Time Finance. Basil Blackwell, London (1990)

Michael, E.: A survey of continuous selections. In Set-Valued Mappings, Selections and Topological Properties of 2^X (Proc. Conf., SUNY, Buffalo,N.Y., 1969), Lecture Notes in Mathematics, vol. 171, pp. 54–58. Springer, Berlin (1970)

Midler, J.L.: Optimal control of a discrete time stochastic system linear in the state. J. Math. Anal. Appl. **25**(1), 114–120 (1969)

Mikosch, T.: Elementary Stochastic Calculus with Finance in View. World Scientific Publishing, Singapore (1998)

Mitra, T., Wan, H.Y., Jr.: Some theoretical results on the economics of forestry. Rev. Econ. Stud. **52**(2), 263–282 (1985)

Morimoto, H.: Stochastic Control and Mathematical Modeling. Cambridge University Press, Cambridge (2010)

Neveu, J.: Mathematical Foundations of The Calculus of Probability. Calif, Holden-Day, San Francisco (1965)

Øksendal, B.: Stochastic Differential Equations: an Introduction with Applications. Springer-Verlag, Berlin (2003)

Øksendal, B., Sulem, A.: Applied Stochastic Control of Jump Diffusions. Springer, Berlin (2007)

Pham, H.: Continuous-Time Stochastic Control and Optimization with Financial Applications. Springer, Berlin (2009)

Pindyck, R.S.: Uncertainty and exhaustible resource markets. J. Polit. Econ. **88**(6), 1203–1225 (1980)

Piunovskiy, A., Zhang, Y.: Continuous-time Markov Decision Processes-Borel Space Models and General Control Strategies. Springer, Cham (2020)

Prieto-Rumeau, T., Hernández-Lerma, O.: Selected Topics on Continuous-Time Controlled Markov Chains and Markov Games. Imperial College Press, London, London (2012)

Raković, S.V., Levine, W.S.: Handbook of Model Predictive Control. Springer, Cham (2019)

Rieder, U.: Measurable selection theorems for optimization problems. Manuscripta Math. **24**(1), 115–131 (1978)

Rishel, R.: Controlled Continuous Time Markov Processes. In Heyman, D.P., S.J.M. (eds.) Stochastic Models, Handbooks Operational Resource Management Science, vol. 2, pp. 435–467. North-Holland, Amsterdam (1990)

Rudin, W.: Principles of Mathematical Analysis. McGraw-Hill Book Co., New York-Auckland-Düsseldorf (1976)

Schäl, M.: Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **32**(3), 179–196 (1975)

Sennott, L.I.: A new condition for the existence of optimal stationary policies in average cost Markov decision processes. Oper. Res. Lett. **5**(1), 17–23 (1986)

Sethi, S.P.: Optimal control theory: Applications to Management Science and Economics, 4th edn. Springer, Cham (2021)

Shapley, L.S.: Stochastic games. Proceed. Nat. Acad. Sci. **39**(10), 1095–1100 (1953)

Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction MIT Press, 2nd edn. MIT Press, Cambridge, MA (2018)

Sznajder, R., Filar, J.A.: Some comments on a theorem of Hardy and Littlewood. J. Optim. Theory Appl. **75**(1), 201–208 (1992)

Vega-Amaya, Ó.: On the vanishing discount factor approach for markov decision processes with weakly continuous transition probabilities. J. Math. Anal. Appl. **426**(2), 978–985 (2015)

Vega-Amaya, Ó.: Solutions of the average cost optimality equation for markov decision processes with weakly continuous kernel: The fixed-point approach revisited. J. Math. Anal. Appl. **464**(1), 152–163 (2018)

Wei, Q., Liao, Z., Yang, Z., Li, B., Liu, D.: Continuous-time time-varying policy iteration. IEEE Trans. Cybern. **50**(12), 4958–4971 (2020)

Wessels, J.: Markov programming by successive approximations with respect to weighted supremum norms. J. Math. Anal. Appl. **58**(2), 326–335 (1977)

Whittle, P.: Optimization Over Time-Volme II: Dynamic Programming and Stochastic Control. John Wiley & Sons, Chichester (1982)

Widder, D.V.: The Laplace Transform. Princeton Mathematical Series, vol. 6. Princeton University Press, Princeton, N. J.: (Dover, p. 2010. N.Y, Mineola (1941)

Yong, J., Zhou, X.Y.: Stochastic Controls: Hamiltonian Systems and HJB Equations. Springer, New York (1999)

Index

A
Abelian theorem
  continuous–time, 159
  discrete–time, 73
Action, *see* Policy
Average cost optimality equation
  continuous–time, 148, 206
  discrete–time, 63, 116
Average cost optimality inequality
  continuous–time, 162, 206
  discrete–time, 113
Average cost problems
  continuous–time, 147
  discrete–time, 62, 111

B
Banach's fixed point theorem, 38
Bellman equation, *see* Dynamic programming equation
Bellman operator, 31, 102
Bellman's principle of optimality, *see* Principle of optimality (PO)
Blackwell's conditions, 46
Borel space, 11, 245
Bounding function, *see* Weight function
Brock–Mirman model, 27, 43, 44, 65, 68, 75
Brownian motion, 186

C
Canonical pair
  continuous–time, 151, 196
  discrete–time, 63
Canonical triplet
  continuous–time, 151
  discrete–time, 63
Cesàro limits, 73, 159
Chapman–Kolmogorov equation, 189
Consumption–investment problem, 6, 98, 227
Control, *see* Policy
Control model, *see* Optimal control problem
Control policy, *see* Policy
Control problem, *see* Optimal control problem

D
Differential equation
  ordinary, 9, 127, 185
  stochastic, 10, 187
Diffusion coefficient, 187
Diffusion process, 187, 217
Dirac measure, 85, 184, 185
Discounted problems
  continuous-time, 141, 143
  discrete-time, 20, 29, 100
Drift coefficient, 187, 219
Driving process, 5, 84, 258
Dynamic programming algorithm, 90
Dynamic programming equation, 14, 16, 31, 88, 91, 102, 131
  forward form, 23, 88, 92, 106
Dynkin's formula, 191, 192

O. Hernández-Lerma et al., *An Introduction to Optimal Control Theory*,
Texts in Applied Mathematics 76, https://doi.org/10.1007/978-3-031-21139-3

E
Economic growth model, *see* Brock–Mirman model
Envelope Theorems, 136
Ergodic Markov process, 197

F
Feller property, 94, 219

G
Gauge function, *see* Weight function
Generator, *see* Infinitesimal generator
Geometrically ergodic Markov process, 197, 209

H
Hamilton–Jacobi–Bellman (HJB), 132, 144
Hamilton–Jacobi–Bellman (HJB) equation, 131
  classical solution, 134
Hamiltonian function, 24, 131
Hardy-Littlewood Theorem, 73

I
Indicator function, 85, 184
Infinitesimal generator, 132, 188, 190
Invariant probability measure, 197
Ionescu Tulcea Theorem, 260
Itô integral, 187
Itô processes, 217

L
Langevin equation, 233
Law of motion, 198
Long–run expected average cost, 7, 194, 195, 205, 237
LQ control problem
  continuous–time, 140, 141, 146, 151, 155, 224, 232, 237
  discrete–time, 18, 30, 57, 75, 96, 98, 122

M
Majorant, *see* Weight function
Markov control model (MCM), 83, 87, 91, 100
Markov control problem
  continuous–time, 132
Markov control process
  continuous–time, 183, 198
  discrete–time, 84, 87
Markov decision process, *see* Markov control process
Markov process
  continuous–time, 183, 185, 187, 188, 190
Minimal cost, *see* Value function
Minimum principle
  continuous-time, 135
  discrete-time, 38
Minimum steady state
  continuous–time, 153
  discrete–time, 67
Minimumprinciple
  discrete-time, 23
Multifunction, 32, 84, 101, 249
  continuous, 251
  lower semicontinuous (l.s.c.), 251
  upper semicontinuous (u.s.c.), 251

O
Optimal control problem
  finite–horizon, 128
Optimal control problem (OCP), 1, 13, 87
  continuous–time, 9, 127, 183
  discounted case, 141
  discrete–time, 1, 13, 87, 100
  finite-horizon, 87
  infinite–horizon, 28, 100, 143
  long–run average cost (AC), 62, 72
  stationary discounted, 28
Optimality equation, *see* Dynamic programing equation

P
Partially observable systems, 8
Poisson equation, 149, 195, 196, 198, 209
Policy, 86
  history–dependent, 5
  action, *see* Control policy
  closed–loop, *see* Markov
  control, 5, 86
  feedback, *see* Markov policy
  Markov, 5, 10, 199
  open–loop, 5, 10, 127
  optimal, 3, 88
  randomized control, 86
  stationary Markov, 53, 199

Policy iteration (PI) algorithm, 54, 108, 110, 163
Pollution accumulation, 156, 240
Pontryagin's Maximum Principle, *see* Minimum principle
Portfolio selection problem, *see* Consumption–investment problem
Principle of optimality (PO), 15, 129
Production–inventory system, 4

R
Random noise, *see* Driving process

S
Sample path, 184
Selector, 32, 249
Semigroup of operators, 189, 190
Stationary measure, *see* Invariant probability measure
Stochastic differential equations (SDEs), *see* Differential equation
Stochastic kernel, 84, 86, 93, 256, 260
Stochastic process
  continuous–time, 183, 186
Stopping time, 192
Strategy, *see* Policy

T
Tracking problem, 7, 30
Transition probability, 84, 93, 184, 187, 198, 256, 260

U
Uniformly ergodic Markov process, *see* Geometrically ergodic Markov process
Uniformly geometrically ergodic Markov process, *see* Geometrically ergodic Markov process

V
Value function, 3, 30, 62, 88, 100, 129, 201, 205, 236, 237
Value iteration (VI)
  algorithm, 51
  functions, 32, 106
Vanishing discount
  continuous–time, 158, 210
  discrete–time, 72
Verification theorem, 133, 134, 144

W
Weight function, 45
Weighted-norm, 44
Wiener process, *see* Brownian motion