1
Introduction

Contents

1.1. Structure of Dynamic Programming Problems . . . . . p. 2
1.2. Abstract Dynamic Programming Models . . . . . p. 5
1.2.1. Problem Formulation . . . . . p. 5
1.2.2. Monotonicity and Contraction Assumptions . . . . p. 7
1.2.3. Some Examples . . . . . p. 9
1.2.4. Approximation-Related Mappings . . . . . p. 21
1.3. Organization of the Book . . . . . p. 23
1.4. Notes, Sources, and Exercises . . . . . p. 25

1.1 STRUCTURE OF DYNAMIC PROGRAMMING PROBLEMS

Dynamic programming (DP for short) is the principal method for analysis of a large and diverse class of sequential decision problems. Examples are deterministic and stochastic optimal control problems with a continuous state space, Markov and semi-Markov decision problems with a discrete state space, minimax problems, and sequential zero sum games. While the nature of these problems may vary widely, their underlying structures turn out to be very similar. In all cases there is an underlying mapping that depends on an associated controlled dynamic system and corresponding cost per stage. This mapping, the DP operator, provides a "compact signature" of the problem. It defines the cost function of policies and the optimal cost function, and it provides a convenient shorthand notation for algorithmic description and analysis.

More importantly, the structure of the DP operator defines the mathematical character of the associated problem. The purpose of this book is to provide an analysis of this structure, centering on two fundamental properties: *monotonicity* and (weighted sup-norm) *contraction*. It turns out that the nature of the analytical and algorithmic DP theory is determined primarily by the presence or absence of these two properties, and the rest of the problem's structure is largely inconsequential.

A Deterministic Optimal Control Example

To illustrate our viewpoint, let us consider a discrete-time deterministic optimal control problem described by a system equation

$$x_{k+1} = f(x_k, u_k), \qquad k = 0, 1, \ldots. \tag{1.1}$$

Here x_k is the state of the system taking values in a set X (the state space), and u_k is the control taking values in a set U (the control space). At stage k, there is a cost

$$\alpha^k g(x_k, u_k)$$

incurred when u_k is applied at state x_k, where α is a scalar in $(0, 1]$ that has the interpretation of a discount factor when $\alpha < 1$. The controls are chosen as a function of the current state, subject to a constraint that depends on that state. In particular, at state x the control is constrained to take values in a given set $U(x) \subset U$. Thus we are interested in optimization over the set of (nonstationary) policies

$$\Pi = \big\{\{\mu_0, \mu_1, \ldots\} \mid \mu_k \in \mathcal{M},\ k = 0, 1, \ldots\big\},$$

where $\mathcal{M}$ is the set of functions $\mu : X \mapsto U$ defined by

$$\mathcal{M} = \big\{\mu \mid \mu(x) \in U(x),\ \forall\ x \in X\big\}.$$

The total cost of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ over an infinite number of stages and starting at an initial state x_0 is

$$J_\pi(x_0) = \sum_{k=0}^{\infty} \alpha^k g\big(x_k, \mu_k(x_k)\big), \tag{1.2}$$

where the state sequence $\{x_k\}$ is generated by the deterministic system (1.1) under the policy π:

$$x_{k+1} = f\big(x_k, \mu_k(x_k)\big), \quad k = 0, 1, \ldots$$

The optimal cost function is †

$$J^*(x) = \inf_{\pi \in \Pi} J_\pi(x), \qquad x \in X.$$

For any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, consider the policy $\pi_1 = \{\mu_1, \mu_2, \ldots\}$ and write by using Eq. (1.2),

$$J_\pi(x) = g\big(x, \mu_0(x)\big) + \alpha J_{\pi_1}\big(f(x, \mu_0(x)\big).$$

We have for all $x \in X$

$$\begin{aligned} J^*(x) &= \inf_{\pi = \{\mu_0, \pi_1\} \in \Pi} \Big\{ g\big(x, \mu_0(x)\big) + \alpha J_{\pi_1}\big(f(x, \mu_0(x)\big) \Big\} \\ &= \inf_{\mu_0 \in \mathcal{M}} \Big\{ g\big(x, \mu_0(x)\big) + \alpha \inf_{\pi_1 \in \Pi} J_{\pi_1}\big(f(x, \mu_0(x)\big) \Big\} \\ &= \inf_{\mu_0 \in \mathcal{M}} \Big\{ g\big(x, \mu_0(x)\big) + \alpha J^*\big(f(x, \mu_0(x)\big) \Big\}. \end{aligned}$$

The minimization over $\mu_0 \in \mathcal{M}$ can be written as minimization over all $u \in U(x)$, so we can write the preceding equation as

$$J^*(x) = \inf_{u \in U(x)} \Big\{ g(x, u) + \alpha J^*\big(f(x, u)\big) \Big\}, \qquad \forall\ x \in X. \tag{1.3}$$

This equation is an example of *Bellman's equation*, which plays a central role in DP analysis and algorithms. If it can be solved for J^*, an optimal stationary policy $\{\mu^*, \mu^*, \ldots\}$ may typically be obtained by minimization of the right-hand side for each x, i.e.,

$$\mu^*(x) \in \arg\min_{u \in U(x)} \Big\{ g(x, u) + \alpha J^*\big(f(x, u)\big) \Big\}, \qquad \forall\ x \in X. \tag{1.4}$$

† For the informal discussion of this section, we will disregard a few mathematical issues. In particular, we assume that the series defining J_π in Eq. (1.2) is convergent for all allowable π, and that the optimal cost function J^* is real-valued. We will address such issues later.

We now note that both Eqs. (1.3) and (1.4) can be stated in terms of the expression

$$H(x,u,J) = g(x,u) + \alpha J\big(f(x,u)\big), \qquad x \in X,\ u \in U(x).$$

Defining

$$(T_\mu J)(x) = H\big(x,\mu(x),J\big), \qquad x \in X,$$

and

$$(TJ)(x) = \inf_{u\in U(x)} H(x,u,J) = \inf_{\mu\in\mathcal{M}} (T_\mu J)(x), \qquad x \in X,$$

we see that Bellman's equation (1.3) can be written compactly as

$$J^* = TJ^*,$$

i.e., J^* is the fixed point of T, viewed as a mapping from the set of real-valued functions on X into itself. Moreover, it can be similarly seen that J_μ, the cost function of the stationary policy $\{\mu,\mu,\ldots\}$, is a fixed point of T_μ. In addition, the optimality condition (1.4) can be stated compactly as

$$T_{\mu^*} J^* = TJ^*.$$

We will see later that additional properties, as well as a variety of algorithms for finding J^* can be analyzed using the mappings T and T_μ.

One more property that holds in some generality is worth noting. For a given policy $\pi = \{\mu_0,\mu_1,\ldots\}$ and a terminal cost $\alpha^N \bar{J}(x_N)$ for the state x_N at the end of N stages, consider the N-stage cost function

$$J_{\pi,N}(x_0) = \alpha^N \bar{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k,\mu_k(x_k)\big). \tag{1.5}$$

Then it can be verified by induction that for all initial states x_0, we have

$$J_{\pi,N}(x_0) = (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x_0). \tag{1.6}$$

Here $T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}}$ is the composition of the mappings $T_{\mu_0}, T_{\mu_1}, \ldots T_{\mu_{N-1}}$, i.e., for all J,

$$(T_{\mu_0} T_{\mu_1} J)(x) = \big(T_{\mu_0}(T_{\mu_1} J)\big)(x), \qquad x \in X,$$

and more generally

$$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} J)(x) = \big(T_{\mu_0}(T_{\mu_1}(\cdots(T_{\mu_{N-1}} J)))\big)(x), \qquad x \in X,$$

(our notational conventions are summarized in Appendix A). Thus the finite horizon cost functions $J_{\pi,N}$ of π can be defined in terms of the mappings T_μ [cf. Eq. (1.6)], and so can their infinite horizon limit J_π:

$$J_\pi(x) = \lim_{N\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad x \in X, \tag{1.7}$$

where $\bar{J}$ is the zero function, $\bar{J}(x) = 0$ for all $x \in X$ (assuming the limit exists).

Connection with Fixed Point Methodology

The Bellman equation (1.3) and the optimality condition (1.4), stated in terms of the mappings T_μ and T, highlight the central theme of this book, which is that DP theory is intimately connected with the theory of abstract mappings and their fixed points. Analogs of the Bellman equation, $J^* = TJ^*$, optimality conditions, and other results and computational methods hold for a great variety of DP models, and can be stated compactly as described above in terms of the corresponding mappings T_μ and T. The gain from this abstraction is greater generality and mathematical insight, as well as a more unified, economical, and streamlined analysis.

1.2 ABSTRACT DYNAMIC PROGRAMMING MODELS

In this section we formally introduce and illustrate with examples an abstract DP model, which embodies the ideas discussed in the preceding section.

1.2.1 Problem Formulation

Let X and U be two sets, which we loosely refer to as a set of "states" and a set of "controls," respectively. For each $x \in X$, let $U(x) \subset U$ be a nonempty subset of controls that are feasible at state x. We denote by $\mathcal{M}$ the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$.

In analogy with DP, we refer to sequences $\pi = \{\mu_0, \mu_1, \ldots\}$, with $\mu_k \in \mathcal{M}$ for all k, as "nonstationary policies," and we refer to a sequence $\{\mu, \mu, \ldots\}$, with $\mu \in \mathcal{M}$, as a "stationary policy." In our development, stationary policies will play a dominant role, and with slight abuse of terminology, we will also refer to any $\mu \in \mathcal{M}$ as a "policy" when confusion cannot arise.

Let $R(X)$ be the set of real-valued functions $J : X \mapsto \Re$, and let $H : X \times U \times R(X) \mapsto \Re$ be a given mapping.† For each policy $\mu \in \mathcal{M}$, we consider the mapping $T_\mu : R(X) \mapsto R(X)$ defined by

$$(T_\mu J)(x) = H\big(x, \mu(x), J\big), \qquad \forall\ x \in X,\ J \in R(X),$$

and we also consider the mapping T defined by‡

$$(TJ)(x) = \inf_{u \in U(x)} H(x, u, J), \qquad \forall\ x \in X,\ J \in R(X).$$

† Our notation and mathematical conventions are outlined in Appendix A. In particular, we denote by $\Re$ the set of real numbers, and by $\Re^n$ the space of n-dimensional vectors with real components.

‡ We assume that H, $T_\mu J$, and TJ are real-valued for $J \in R(X)$ in the present chapter and in Chapter 2. In Chapters 3-5 we will allow $H(x, u, J)$, and hence also $(T_\mu J)(x)$ and $(TJ)(x)$, to take the values ∞ and $-\infty$.

Similar to the deterministic optimal control problem of the preceding section, the mappings T_μ and T serve to define a multistage optimization problem and a DP-like methodology for its solution. In particular, for some function $\bar{J} \in R(X)$, and nonstationary policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we define for each integer $N \geq 1$ the functions

$$J_{\pi,N}(x) = (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad x \in X,$$

where $T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}}$ denotes the composition of the mappings T_{μ_0}, T_{μ_1}, $\ldots$, $T_{\mu_{N-1}}$, i.e.,

$$T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} J = \big(T_{\mu_0}(T_{\mu_1}(\cdots(T_{\mu_{N-2}}(T_{\mu_{N-1}} J)))\cdots)\big), \quad J \in R(X).$$

We view $J_{\pi,N}$ as the "N-stage cost function" of π [cf. Eq. (1.5)]. Consider also the function

$$J_\pi(x) = \limsup_{N\to\infty} J_{\pi,N}(x) = \limsup_{N\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad x \in X,$$

which we view as the "infinite horizon cost function" of π [cf. Eq. (1.7); we use lim sup for generality, since we are not assured that the limit exists]. We want to minimize J_π over π, i.e., to find

$$J^*(x) = \inf_\pi J_\pi(x), \qquad x \in X,$$

and a policy π^* that attains the infimum, if one exists.

The key connection with fixed point methodology is that J^* "typically" (under mild assumptions) can be shown to satisfy

$$J^*(x) = \inf_{u \in U(x)} H(x, u, J^*), \qquad \forall\ x \in X,$$

i.e., it is a fixed point of T. We refer to this as *Bellman's equation* [cf. Eq. (1.3)]. Another fact is that if an optimal policy π^* exists, it "typically" can be selected to be stationary, $\pi^* = \{\mu^*, \mu^*, \ldots\}$, with $\mu^* \in \mathcal{M}$ satisfying an optimality condition, such as for example

$$T_{\mu^*} J^* = T J^*$$

[cf. Eq. (1.4)]. Several other results of an analytical or algorithmic nature also hold under appropriate conditions, which will be discussed in detail later.

However, Bellman's equation and other related results may not hold without T_μ and T having some special structural properties. Prominent among these are a monotonicity assumption that typically holds in DP problems, and a contraction assumption that holds for some important classes of problems.

1.2.2 Monotonicity and Contraction Assumptions

Let us now formalize the monotonicity and contraction assumptions. We will require that both of these assumptions hold throughout the next chapter, and we will gradually relax the contraction assumption in Chapters 3-5. Recall also our assumption that T_μ and T map $R(X)$ (the space of real-valued functions over X) into $R(X)$. In Chapters 3-5 we will relax this assumption as well.

Assumption 1.2.1: (Monotonicity) If $J, J' \in R(X)$ and $J \leq J'$, then

$$H(x,u,J) \leq H(x,u,J'), \qquad \forall\ x \in X,\ u \in U(x).$$

Note that by taking infimum over $u \in U(x)$, we have

$$J \leq J' \qquad \Rightarrow \qquad \inf_{u \in U(x)} H(x,u,J) \leq \inf_{u \in U(x)} H(x,u,J'), \quad \forall\ x \in X,$$

or equivalently,

$$J \leq J' \qquad \Rightarrow \qquad TJ \leq TJ'.$$

Another way to arrive at this relation, is to note that the monotonicity assumption is equivalent to

$$J \leq J' \quad \Rightarrow \quad T_\mu J \leq T_\mu J', \qquad \forall\ \mu \in \mathcal{M},$$

and to use the simple but important fact

$$\inf_{u \in U(x)} H(x,u,J) = \inf_{\mu \in \mathcal{M}} (T_\mu J)(x), \qquad \forall\ x \in X,\ J \in R(x),$$

i.e., *infimum over u is equivalent to infimum over μ*, which holds in view of the definition $\mathcal{M} = \big\{\mu \mid \mu(x) \in U(x),\ \forall\ x \in X\big\}$. We will be writing this relation as $TJ = \inf_{\mu \in \mathcal{M}} T_\mu J$.

For the contraction assumption, we introduce a function $v : X \mapsto \Re$ with

$$v(x) > 0, \qquad \forall\ x \in X.$$

Let us denote by $B(X)$ the space of real-valued functions J on X such that $J(x)/v(x)$ is bounded as x ranges over X, and consider the weighted sup-norm

$$\|J\| = \sup_{x \in X} \frac{\big|J(x)\big|}{v(x)}$$

on $B(X)$. The properties of $B(X)$ and some of the associated fixed point theory are discussed in Appendix B. In particular, as shown there, $B(X)$

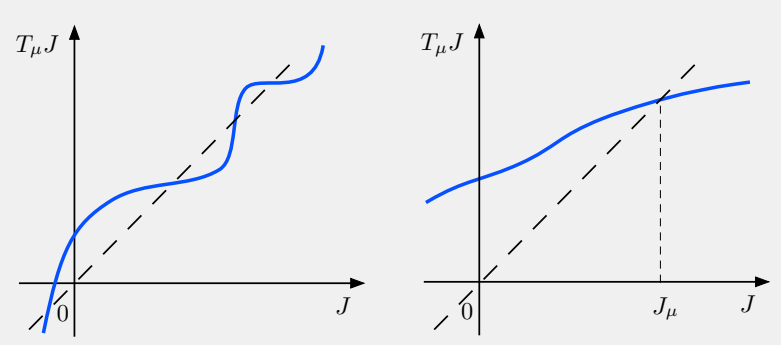

Figure 1.2.1. Illustration of the monotonicity and the contraction assumptions in one dimension. The mapping T_μ on the left is monotone but is not a contraction. The mapping T_μ on the right is both monotone and a contraction. It has a unique fixed point at J_μ.

is a complete normed space, so any mapping from $B(X)$ to $B(X)$ that is a contraction or an m-stage contraction for some integer $m > 1$, with respect to $\|\cdot\|$, has a unique fixed point (cf. Props. B.1 and B.2).

Assumption 1.2.2: (Contraction) For all $J \in B(X)$ and $\mu \in \mathcal{M}$, the functions $T_\mu J$ and TJ belong to $B(X)$. Furthermore, for some $\alpha \in (0,1)$, we have

$$\|T_\mu J - T_\mu J'\| \leq \alpha \|J - J'\|, \qquad \forall\ J, J' \in B(X),\ \mu \in \mathcal{M}. \tag{1.8}$$

Figure 1.2.1 illustrates the monotonicity and the contraction assumptions. It is important to note that the contraction condition (1.8) implies that

$$\|TJ - TJ'\| \leq \alpha \|J - J'\|, \qquad \forall\ J, J' \in B(X), \tag{1.9}$$

so that T is also a contraction with modulus α. To see this we use Eq. (1.8) to write

$$(T_\mu J)(x) \leq (T_\mu J')(x) + \alpha \|J - J'\|\, v(x), \qquad \forall\ x \in X,$$

from which, by taking infimum of both sides over $\mu \in \mathcal{M}$, we have

$$\frac{(TJ)(x) - (TJ')(x)}{v(x)} \leq \alpha \|J - J'\|, \qquad \forall\ x \in X.$$

Reversing the roles of J and J', we also have

$$\frac{(TJ')(x) - (TJ)(x)}{v(x)} \leq \alpha \|J - J'\|, \qquad \forall\ x \in X,$$

and combining the preceding two relations, and taking the supremum of the left side over $x \in X$, we obtain Eq. (1.9).

Nearly all mappings related to DP satisfy the monotonicity assumption, and many important ones satisfy the weighted sup-norm contraction assumption as well. When both assumptions hold, the most powerful analytical and computational results can be obtained, as we will show in Chapter 2. These are:

(a) Bellman's equation has a unique solution, i.e., T and T_μ have unique fixed points, which are the optimal cost function J^* and the cost functions J_μ of the stationary policies $\{\mu, \mu, \ldots\}$, respectively [cf. Eq. (1.3)].

(b) A stationary policy $\{\mu^*, \mu^*, \ldots\}$ is optimal if and only if

$$T_{\mu^*} J^* = T J^*,$$

[cf. Eq. (1.4)].

(c) J^* and J_μ can be computed by the *value iteration* method,

$$J^* = \lim_{k \to \infty} T^k J, \qquad J_\mu = \lim_{k \to \infty} T_\mu^k J,$$

starting with any $J \in B(X)$.

(d) J^* can be computed by the *policy iteration* method, whereby we generate a sequence of stationary policies via

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k},$$

starting from some initial policy μ^0 [here J_{μ^k} is obtained as the fixed point of T_{μ^k} by several possible methods, including value iteration as in (c) above].

These are the most favorable types of results one can hope for in the DP context, and they are supplemented by a host of other results, involving approximate and/or asynchronous implementations of the value and policy iteration methods, and other related methods that combine features of both. As the contraction property is relaxed and is replaced by various weaker assumptions, some of the preceding results may hold in weaker form. For example J^* turns out to be a solution of Bellman's equation in all the models to be discussed, but it may not be the unique solution. The interplay between the monotonicity and contraction-like properties, and the associated results of the form (a)-(d) described above is the recurring analytical theme in this book.

1.2.3 Some Examples

In what follows in this section, we describe a few special cases, which indicate the connections of appropriate forms of the mapping H with the most

popular total cost DP models. In all these models the monotonicity Assumption 1.2.1 (or some closely related version) holds, but the contraction Assumption 1.2.2 may not hold, as we will indicate later. Our descriptions are by necessity brief, and the reader is referred to the relevant textbook literature for more detailed discussion.

Example 1.2.1 (Stochastic Optimal Control - Markovian Decision Problems)

Consider the stationary discrete-time dynamic system

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots, \tag{1.10}$$

where for all k, the state x_k is an element of a space X, the control u_k is an element of a space U, and w_k is a random "disturbance," an element of a space W. We consider problems with infinite state and control spaces, as well as problems with discrete (finite or countable) state space (in which case the underlying system is a Markov chain). However, for technical reasons that relate to measure theoretic issues to be explained later in Chapter 5, we assume that W is a countable set.

The control u_k is constrained to take values in a given nonempty subset $U(x_k)$ of U, which depends on the current state x_k [$u_k \in U(x_k)$, for all $x_k \in X$]. The random disturbances w_k, $k = 0, 1, \ldots$, are characterized by probability distributions $P(\cdot \mid x_k, u_k)$ that are identical for all k, where $P(w_k \mid x_k, u_k)$ is the probability of occurrence of w_k, when the current state and control are x_k and u_k, respectively. Thus the probability of w_k may depend explicitly on x_k and u_k, but not on values of prior disturbances $w_{k-1}, \ldots, w_0$.

Given an initial state x_0, we want to find a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, where $\mu_k : X \mapsto U$, $\mu_k(x_k) \in U(x_k)$, for all $x_k \in X$, $k = 0, 1, \ldots$, that minimizes the cost function

$$J_\pi(x_0) = \limsup_{N \to \infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}, \tag{1.11}$$

subject to the system equation constraint

$$x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big), \qquad k = 0, 1, \ldots.$$

This is a classical problem, which is discussed extensively in various sources, including the author's text [Ber12a]. It is usually referred to as the *stochastic optimal control problem* or the *Markovian Decision Problem* (MDP for short).

Note that the expected value of the N-stage cost of π,

$$\mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\},$$

is defined as a (possibly countably infinite) sum, since the disturbances w_k, $k = 0, 1, \ldots$, take values in a countable set. Indeed, the reader may verify

that all the subsequent mathematical expressions that involve an expected value can be written as summations over a finite or a countable set, so they make sense without resort to measure-theoretic integration concepts. †

In what follows we will often impose appropriate assumptions on the cost per stage g and the scalar α, which guarantee that the infinite horizon cost $J_\pi(x_0)$ is defined as a limit (rather than as a lim sup):

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k\\ k=0,1,\dots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}.$$

In particular, it can be shown that the limit exists if $\alpha < 1$ and g is uniformly bounded, i.e., for some $B > 0$,

$$\big|g(x,u,w)\big| \le B, \qquad \forall\ x \in X,\ u \in U(x),\ w \in W. \tag{1.12}$$

In this case, we obtain the classical discounted infinite horizon DP problem, which generally has the most favorable structure of all infinite horizon stochastic DP models (see [Ber12a], Chapters 1 and 2).

To make the connection with abstract DP, let us define

$$H(x,u,J) = E\Big\{ g(x,u,w) + \alpha J\big(f(x,u,w)\big) \Big\},$$

so that

$$(T_\mu J)(x) = E\Big\{ g\big(x,\mu(x),w\big) + \alpha J\big(f(x,\mu(x),w)\big) \Big\},$$

and

$$(TJ)(x) = \inf_{u\in U(x)} E\Big\{ g(x,u,w) + \alpha J\big(f(x,u,w)\big) \Big\}.$$

Similar to the deterministic optimal control problem of Section 1.1, the N-stages cost of π, can be expressed in terms of T_μ:

$$(T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J})(x_0) = \mathop{E}_{\substack{w_k\\ k=0,1,\dots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\},$$

† As noted in Appendix A, the formula for the expected value of a random variable w defined over a space Ω is

$$E\{w\} = E\{w^+\} + E\{w^-\},$$

where w^+ and w^- are the positive and negative parts of w,

$$w^+(\omega) = \max\big\{0, w(\omega)\big\}, \qquad w^-(\omega) = \min\big\{0, w(\omega)\big\}, \qquad \forall\ \omega \in \Omega.$$

In this way, taking also into account the rule $\infty - \infty = \infty$ (see Appendix A), $E\{w\}$ is well-defined as an extended real number if Ω is finite or countably infinite.

where $\bar{J}$ is the zero function, $\bar{J}(x) = 0$ for all $x \in X$. The same is true for the infinite stages cost [cf. Eq. (1.11)]:

$$J_\pi(x_0) = \limsup_{N\to\infty} (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(x_0).$$

It can be seen that the mappings T_μ and T are monotone, and it is well-known that if $\alpha < 1$ and the boundedness condition (1.12) holds, they are contractive as well (under the unweighted sup-norm); see e.g., [Ber12a], Chapter 1. In this case, the model has the powerful analytical and algorithmic properties (a)-(d) mentioned at the end of the preceding subsection. In particular, the optimal cost function J^* [i.e., $J^*(x) = \inf_\pi J_\pi(x)$ for all $x \in X$] can be shown to be the unique solution of the fixed point equation $J^* = TJ^*$, also known as Bellman's equation, which has the form

$$J^*(x) = \inf_{u\in U(x)} E\Big\{g(x,u,w) + \alpha J^*\big(f(x,u,w)\big)\Big\}, \qquad x \in X,$$

and parallels the one given for deterministic optimal control problems [cf. Eq. (1.3)].

These properties can be expressed and analyzed in an abstract setting by using just the mappings T_μ and T, both when T_μ and T are contractive (see Chapter 2), and when they are only monotone and not contractive (see Chapter 4). Moreover, under some conditions, it is possible to analyze these properties in cases where T_μ is contractive for some but not all μ (see Chapter 3, and Sections 4.4-4.5).

Example 1.2.2 (Finite-State Discounted Markovian Decision Problems)

In the special case of the preceding example where the number of states is finite, the system equation (1.10) may be defined in terms of the transition probabilities

$$p_{xy}(u) = \text{Prob}\big(y = f(x,u,w) \mid x\big), \qquad x, y \in X,\ u \in U(x),$$

so H takes the form

$$H(x,u,J) = \sum_{y\in X} p_{xy}(u)\big(g(x,u,y) + \alpha J(y)\big).$$

When $\alpha < 1$ and the boundedness condition

$$\big|g(x,u,y)\big| \le B, \qquad \forall\ x, y \in X,\ u \in U(x),$$

[cf. Eq. (1.12)] holds (or more simply when U is a finite set), the mappings T_μ and T are contraction mappings with respect to the standard (unweighted) sup-norm. This is a classical problem, referred to as *discounted finite-state MDP*, which has a favorable theory and has found extensive applications (cf. [Ber12a], Chapters 1 and 2). The model is additionally important, because it is often used for computational solution of continuous state space problems via discretization.

Example 1.2.3 (Discounted Semi-Markov Problems)

With x, y, and u as in Example 1.2.2, consider a mapping of the form

$$H(x,u,J) = G(x,u) + \sum_{y \in X} m_{xy}(u) J(y),$$

where G is some function representing expected cost per stage, and $m_{xy}(u)$ are nonnegative scalars with

$$\sum_{y \in X} m_{xy}(u) < 1, \qquad \forall\, x \in X,\, u \in U(x).$$

The equation $J^* = TJ^*$ is Bellman's equation for a finite-state continuous-time semi-Markov decision problem, after it is converted into an equivalent discrete-time problem (cf. [Ber12a], Section 1.4). Again, the mappings T_μ and T are monotone and can be shown to be contraction mappings with respect to the unweighted sup-norm.

Example 1.2.4 (Discounted Zero-Sum Dynamic Games)

Let us consider a zero sum game analog of the finite-state MDP Example 1.2.2. Here there are two players that choose actions at each stage: the first (called the *minimizer*) may choose a move i out of n moves and the second (called the *maximizer*) may choose a move j out of m moves. Then the minimizer gives a specified amount a_{ij} to the maximizer, called a *payoff*. The minimizer wishes to minimize a_{ij}, and the maximizer wishes to maximize a_{ij}.

The players use mixed strategies, whereby the minimizer selects a probability distribution $u = (u_1, \ldots, u_n)$ over his n possible moves and the maximizer selects a probability distribution $v = (v_1, \ldots, v_m)$ over his m possible moves. Since the probability of selecting i and j is $u_i v_j$, the expected payoff for this stage is $\sum_{i,j} a_{ij} u_i v_j$ or $u'Av$, where A is the $n \times m$ matrix with components a_{ij}.

In a single-stage version of the game, the minimizer must minimize $\max_{v \in V} u'Av$ and the maximizer must maximize $\min_{u \in U} u'Av$, where U and V are the sets of probability distributions over $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$, respectively. A fundamental result (which will not be proved here) is that these two values are equal:

$$\min_{u \in U} \max_{v \in V} u'Av = \max_{v \in V} \min_{u \in U} u'Av. \tag{1.13}$$

Let us consider the situation where a separate game of the type just described is played at each stage. The game played at a given stage is represented by a "state" x that takes values in a finite set X. The state evolves according to transition probabilities $q_{xy}(i,j)$ where i and j are the moves selected by the minimizer and the maximizer, respectively (here y represents the next game to be played after moves i and j are chosen at the game represented by x). When the state is x, under $u \in U$ and $v \in V$, the one-stage

expected payoff is $u'A(x)v$, where $A(x)$ is the $n \times m$ payoff matrix, and the state transition probabilities are

$$p_{xy}(u,v) = \sum_{i=1}^{n} \sum_{j=1}^{m} u_i v_j q_{xy}(i,j) = u' Q_{xy} v,$$

where Q_{xy} is the $n \times m$ matrix that has components $q_{xy}(i,j)$. Payoffs are discounted by $\alpha \in (0,1)$, and the objectives of the minimizer and maximizer, roughly speaking, are to minimize and to maximize the total discounted expected payoff. This requires selections of u and v to strike a balance between obtaining favorable current stage payoffs and playing favorable games in future stages.

We now introduce an abstract DP framework related to the sequential move selection process just described. We consider the mapping G given by

$$\begin{aligned} G(x,u,v,J) &= u' A(x) v + \alpha \sum_{y \in X} p_{xy}(u,v) J(y) \\ &= u' \left(A(x) + \alpha \sum_{y \in X} Q_{xy} J(y) \right) v, \end{aligned} \tag{1.14}$$

where $\alpha \in (0,1)$ is discount factor, and the mapping H given by

$$H(x,u,J) = \max_{v \in V} G(x,u,v,J).$$

The corresponding mappings T_μ and T are

$$(T_\mu J)(x) = \max_{v \in V} G\big(x, \mu(x), v, J\big), \qquad x \in X,$$

and

$$(TJ)(x) = \min_{u \in U} \max_{v \in V} G(x,u,v,J).$$

It can be shown that T_μ and T are monotone and (unweighted) sup-norm contractions. Moreover, the unique fixed point J^* of T satisfies

$$J^*(x) = \min_{u \in U} \max_{v \in V} G(x,u,v,J^*), \qquad \forall\ x \in X,$$

(see [Ber12a], Section 1.6.2).

We now note that since

$$A(x) + \alpha \sum_{y \in X} Q_{xy} J(y)$$

[cf. Eq. (1.14)] is a matrix that is independent of u and v, we may view $J^*(x)$ as the value of a static game (which depends on the state x). In particular, from the fundamental minimax equality (1.13), we have

$$\min_{u \in U} \max_{v \in V} G(x,u,v,J^*) = \max_{v \in V} \min_{u \in U} G(x,u,v,J^*), \qquad \forall\ x \in X.$$

This implies that J^* is also the unique fixed point of the mapping

$$(\overline{T}J)(x) = \max_{v \in V} \overline{H}(x, v, J),$$

where

$$\overline{H}(x, v, J) = \min_{u \in U} G(x, u, v, J),$$

i.e., J^* is the fixed point regardless of the order in which minimizer and maximizer select mixed strategies at each stage.

In the preceding development, we have introduced J^* as the unique fixed point of the mappings T and $\overline{T}$. However, J^* also has an interpretation in game theoretic terms. In particular, it can be shown that $J^*(x)$ is the value of a dynamic game, whereby at state x the two opponents choose multistage (possibly nonstationary) policies that consist of functions of the current state, and continue to select moves using these policies over an infinite horizon. For further discussion of this interpretation, we refer to [Ber12a] and to books on dynamic games such as [FiV96]; see also [PaB99] and [Yu11] for an analysis of the undiscounted case ($\alpha = 1$) where there is a termination state, as in the stochastic shortest path problems of the subsequent Example 1.2.6.

Example 1.2.5 (Minimax Problems)

Consider a minimax version of Example 1.2.1, where w is not random but is rather chosen by an antagonistic player from a set $W(x, u)$. Let

$$H(x, u, J) = \sup_{w \in W(x,u)} \Big[g(x, u, w) + \alpha J\big(f(x, u, w)\big) \Big].$$

Then the equation $J^* = TJ^*$ is Bellman's equation for an infinite horizon minimax DP problem. A special case of this mapping arises in zero-sum dynamic games (cf. Example 1.2.4).

Example 1.2.6 (Stochastic Shortest Path Problems)

The stochastic shortest path (SSP for short) problem is the special case of the stochastic optimal control Example 1.2.1 where:

(a) There is no discounting ($\alpha = 1$).

(b) The state space is $X = \{0, 1, \ldots, n\}$ and we are given transition probabilities, denoted by

$$p_{xy}(u) = P(x_{k+1} = y \mid x_k = x, u_k = u), \qquad x, y \in X,\ u \in U(x).$$

(c) The control constraint set $U(x)$ is finite for all $x \in X$.

(d) A cost $g(x, u)$ is incurred when control $u \in U(x)$ is selected at state x.

(e) State 0 is a special termination state, which is absorbing and cost-free, i.e.,

$$p_{00}(u) = 1,$$

and for all $u \in U(0)$, $g(0,u) = 0$.

To simplify the notation, we have assumed that the cost per stage does not depend on the successor state, which amounts to using expected cost per stage in all calculations.

Since the termination state 0 is cost-free and absorbing, the cost starting from 0 is zero for every policy. Accordingly, for all cost functions, we ignore the component that corresponds to 0, and define

$$H(x,u,J) = g(x,u) + \sum_{y=1}^{n} p_{xy}(u)J(y), \quad x = 1,\ldots,n,\ u \in U(x),\ J \in \Re^n.$$

The mappings T_μ and T are defined by

$$(T_\mu J)(x) = g\big(x,\mu(x)\big) + \sum_{y=1}^{n} p_{xy}\big(\mu(x)\big)J(y), \qquad x = 1,\ldots,n,$$

$$(TJ)(x) = \min_{u\in U(x)} \left[g(x,u) + \sum_{y=1}^{n} p_{xy}(u)J(y) \right], \qquad x = 1,\ldots,n.$$

Note that the matrix that has components $p_{xy}(u)$, $x,y = 1,\ldots,n$, is substochastic (some of its row sums may be less than 1) because there may be positive transition probability from a state x to the termination state 0. Consequently T_μ may be a contraction for some μ, but not necessarily for all $\mu \in \mathcal{M}$.

The SSP problem has been discussed in many sources, including the books [Pal67], [Der70], [Whi82], [Ber87], [BeT89], [HeL99], and [Ber12a], where it is sometimes referred to by earlier names such as "first passage problem" and "transient programming problem." In the framework that is most relevant to our purposes, there is a classification of stationary policies for SSP into *proper* and *improper*. We say that $\mu \in \mathcal{M}$ is proper if, when using μ, there is positive probability that termination will be reached after at most n stages, regardless of the initial state; i.e., if

$$\rho_\mu = \max_{x=1,\ldots,n} P\{x_n \neq 0 \mid x_0 = x, \mu\} < 1.$$

Otherwise, we say that μ is improper. It can be seen that μ is proper if and only if in the Markov chain corresponding to μ, each state x is connected to the termination state with a path of positive probability transitions.

For a proper policy μ, it can be shown that T_μ is a weighted sup-norm contraction, as well as an n-stage contraction with respect to the unweighted sup-norm. For an improper policy μ, T_μ is not a contraction with respect to any norm. Moreover, T also need not be a contraction with respect to any norm (think of the case where there is only one policy, which is improper).

However, T is a weighted sup-norm contraction in the important special case where all policies are proper (see [BeT96], Prop. 2.2, or [Ber12a], Prop. 3.3.1).

Nonetheless, even in the case where there are improper policies and T is not a contraction, results comparable to the case of discounted finite-state MDP are available for SSP problems assuming that:

(a) There exists at least one proper policy.

(b) For every improper policy there is an initial state that has infinite cost under this policy.

Under the preceding two assumptions, it was shown in [BeT91] that T has a unique fixed point J^*, the optimal cost function of the SSP problem. Moreover, a policy $\{\mu^*, \mu^*, \ldots\}$ is optimal if and only if

$$T_{\mu^*} J^* = T J^*.$$

In addition, J^* and J_μ can be computed by value iteration,

$$J^* = \lim_{k \to \infty} T^k J, \qquad J_\mu = \lim_{k \to \infty} T_\mu^k J,$$

starting with any $J \in \Re^n$ (see [Ber12a], Chapter 3, for a textbook account). These properties are in analogy with the desirable properties (a)-(c), given at the end of the preceding subsection in connection with contractive models.

Regarding policy iteration, it works in its strongest form when there are no improper policies, in which case the mappings T_μ and T are weighted sup-norm contractions. When there are improper policies, modifications to the policy iteration method are needed; see [YuB11a], [Ber12a], and also Sections 3.3.2, 3.3.3, where these modifications will be discussed in an abstract setting.

Let us also note that there is an alternative line of analysis of SSP problems, whereby favorable results are obtained assuming that there exists an optimal proper policy, and the one-stage cost is nonnegative, $g(x,u) \geq 0$ for all (x,u) (see [Pal67], [Der70], [Whi82], and [Ber87]). This analysis will also be generalized in Chapter 3 and in Section 4.4, and the nonnegativity assumption on g will be relaxed.

Example 1.2.7 (Deterministic Shortest Path Problems)

The special case of the SSP problem where the state transitions are deterministic is the classical shortest path problem. Here, we have a graph of n nodes $x = 1, \ldots, n$, plus a destination 0, and an arc length a_{xy} for each directed arc (x,y). At state/node x, a policy μ chooses an outgoing arc from x. Thus the controls available at x can be identified with the outgoing neighbors of x [the nodes u such that (x,u) is an arc]. The corresponding mapping H is

$$H(x,u,J) = \begin{cases} a_{xu} + J(u) & \text{if } u \neq 0, \\ a_{x0} & \text{if } u = 0, \end{cases} \qquad x = 1, \ldots, n.$$

A stationary policy μ defines a graph whose arcs are $\big(x, \mu(x)\big)$, $x = 1, \ldots, n$. The policy μ is proper if and only if this graph is acyclic (it consists of a tree of directed paths leading from each node to the destination). Thus there

exists a proper policy if and only if each node is connected to the destination with a directed path. Furthermore, an improper policy has finite cost starting from every initial state if and only if all the cycles of the corresponding graph have nonnegative cycle cost. It follows that the favorable analytical and algorithmic results described for SSP in the preceding example hold if the given graph is connected and the costs of all its cycles are positive. We will see later that significant complications result if the cycle costs are allowed to be nonpositive, even though the shortest path problem is still well posed in the sense that shortest paths exist if the given graph is connected (see Section 3.1.2).

Example 1.2.8 (Multiplicative and Risk-Sensitive Models)

With x, y, u, and transition probabilities $p_{xy}(u)$, as in the finite-state MDP of Example 1.2.2, consider the mapping

$$H(x,u,J) = \sum_{y \in X} p_{xy}(u) g(x,u,y) J(y) = E\Big\{ g(x,u,y) J(y) \mid x, u \Big\}, \tag{1.15}$$

where g is a scalar function with $g(x,u,y) \geq 0$ for all x, y, u (this is necessary for H to be monotone). This mapping corresponds to the multiplicative model of minimizing over all $\pi = \{\mu_0, \mu_1, \ldots\}$ the cost

$$\begin{aligned} J_\pi(x_0) = \limsup_{N \to \infty} E\Big\{ & g\big(x_0, \mu_0(x_0), x_1\big) g\big(x_1, \mu_1(x_1), x_2\big) \cdots \\ & g\big(x_{N-1}, \mu_{N-1}(x_{N-1}), x_N\big) \mid x_0 \Big\}, \end{aligned} \tag{1.16}$$

where the state sequence $\{x_0, x_1, \ldots\}$ is generated using the transition probabilities $p_{x_k x_{k+1}}\big(\mu_k(x_k)\big)$.

To see that the mapping H of Eq. (1.15) corresponds to the cost function (1.16), let us consider the unit function

$$\bar{J}(x) \equiv 1, \qquad x \in X,$$

and verify that for all $x_0 \in X$, we have

$$\begin{aligned} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x_0) = E\Big\{ & g\big(x_0, \mu_0(x_0), x_1\big) g\big(x_1, \mu_1(x_1), x_2\big) \cdots \\ & g\big(x_{N-1}, \mu_{N-1}(x_{N-1}), x_N\big) \mid x_0 \Big\}, \end{aligned} \tag{1.17}$$

so that

$$J_\pi(x) = \limsup_{N \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad x \in X.$$

Indeed, taking into account that $\bar{J}(x) \equiv 1$, we have

$$\begin{aligned} (T_{\mu_{N-1}} \bar{J})(x_{N-1}) &= E\Big\{ g\big(x_{N-1}, \mu_{N-1}(x_{N-1}), x_N\big) \bar{J}(x_N) \mid x_{N-1} \Big\} \\ &= E\Big\{ g\big(x_{N-1}, \mu_{N-1}(x_{N-1}), x_N\big) \mid x_{N-1} \Big\}, \end{aligned}$$

$$(T_{\mu_{N-2}}T_{\mu_{N-1}}\bar{J})(x_{N-2}) = \big((T_{\mu_{N-2}}(T_{\mu_{N-1}}\bar{J})\big)(x_{N-2})$$
$$= E\big\{g\big(x_{N-2},\mu_{N-2}(x_{N-2}),x_{N-1}\big)$$
$$\cdot E\big\{g\big(x_{N-1},\mu_{N-1}(x_{N-1}),x_N\big) \mid x_{N-1}\big\} \mid x_{N-2}\big\},$$

and continuing similarly,

$$(T_{\mu_0}T_{\mu_1}\cdots T_{\mu_{N-1}}\bar{J})(x_0) = E\Big\{g\big(x_0,\mu_0(x_0),x_1\big)E\big\{g\big(x_1,\mu_1(x_1),x_2\big)\cdots$$
$$E\big\{g\big(x_{N-1},\mu_{N-1}(x_{N-1}),x_N\big) \mid x_{N-1}\big\} \mid x_{N-2}\big\}\cdots\big\} \mid x_0\Big\},$$

which by using the iterated expectations formula (see e.g., [BeT08]) proves the expression (1.17).

An important special case of a multiplicative model is when g has the form

$$g(x,u,y) = e^{h(x,u,y)}$$

for some one-stage cost function h. We then obtain a finite-state MDP with an exponential cost function,

$$J_\pi(x_0) = \limsup_{N\to\infty} E\Big\{e^{\left(h(x_0,\mu_0(x_0),x_1)+\cdots+h(x_{N-1},\mu_{N-1}(x_{N-1}),x_N)\right)}\Big\},$$

which is often used to introduce risk aversion in the choice of policy through the convexity of the exponential.

There is also a multiplicative version of the infinite state space stochastic optimal control problem of Example 1.2.1. The mapping H takes the form

$$H(x,u,J) = E\big\{g(x,u,w)J\big(f(x,u,w)\big)\big\},$$

where $x_{k+1} = f(x_k,u_k,w_k)$ is the underlying discrete-time dynamic system; cf. Eq. (1.10).

Multiplicative models and related risk-sensitive models are discussed extensively in the literature, mostly for the exponential cost case and under different assumptions than ours; see e.g., [HoM72], [Jac73], [Rot84], [ChS87], [Whi90], [JBE94], [FlM95], [HeM96], [FeM97], [BoM99], [CoM99], [BoM99], [BoM02], [BBB08]. The works of references [DeR79], [Pat01], and [Pat07] relate to the stochastic shortest path problems of Example 1.2.6, and are the closest to the semicontractive models discussed in Chapters 3 and 4.

Issues of risk-sensitivity have also been dealt within frameworks that do not quite conform to the multiplicative model of the preceding example, and are based on the theory of multi-stage risk measures; see e.g., [Rus10], [CaR12], and the references quoted there. Still these formulations involve abstract monotone DP mappings and are covered by our theory.

Example 1.2.9 (Distributed Aggregation)

The abstract DP framework is useful not only in modeling DP problems, but also in modeling algorithms arising in DP and even other contexts. We illustrate this with an example from [BeY10b] that relates to the distributed

solution of large-scale discounted finite-state MDP using cost function approximation based on aggregation.† It involves a partition of the n states into m subsets for the purposes of distributed computation, and yields a corresponding approximation $(V_1, \ldots, V_m)$ to the cost vector J^*.

In particular, we have a discounted n-state MDP (cf. Example 1.2.2), and we introduce aggregate states $S_1, \ldots, S_m$, which are disjoint subsets of the original state space $\{1, \ldots, n\}$. We assume that these sets form a partition, i.e., each $x \in \{1, \ldots, n\}$ belongs to one and only one of the aggregate state/subsets. We envision a network of processors $\ell = 1, \ldots, m$, each assigned to the computation of a local cost function V_ℓ, defined on the corresponding aggregate state/subset S_ℓ:

$$V_\ell = \{V_{\ell y} \mid y \in S_\ell\}.$$

Processor ℓ also maintains a scalar aggregate cost R_ℓ for its aggregate state, which is a weighted average of the detailed cost values $V_{\ell x}$ within S_ℓ:

$$R_\ell = \sum_{x \in S_\ell} d_{\ell x} V_{\ell x},$$

where $d_{\ell x}$ are given probabilities with $d_{\ell x} \geq 0$ and $\sum_{x \in S_\ell} d_{\ell x} = 1$. The aggregate costs R_ℓ are communicated between processors and are used to perform the computation of the local cost functions V_ℓ (we will discuss computation models of this type in Section 2.6).

We denote $J = (V_1, \ldots, V_m, R_1, \ldots, R_m)$, so that J is a vector of dimension $n + m$. We introduce the mapping $H(x, u, J)$ defined for each of the n states x by

$$H(x, u, J) = W_\ell(x, u, V_\ell, R_1, \ldots, R_m), \qquad \text{if } x \in S_\ell.$$

where for $x \in S_\ell$

$$W_\ell(x, u, V_\ell, R_1, \ldots, R_m) = \sum_{y=1}^{n} p_{xy}(u) g(x, u, y) + \alpha \sum_{y \in S_\ell} p_{xy}(u) V_{\ell y} + \alpha \sum_{y \notin S_\ell} p_{xy}(u) R_{s(y)};$$

and for each original system state y, we denote by $s(y)$ the index of the subset to which y belongs [i.e., $y \in S_{s(y)}$].

We may view H as an abstract mapping on the space of J, and aim to find its fixed point $J^* = (V_1^*, \ldots, V_m^*, R_1^*, \ldots, R_m^*)$. Then, for $\ell = 1, \ldots, m$, we

† See [Ber12a], Section 6.5.2, for a more detailed discussion. Other examples of algorithmic mappings that come under our framework arise in asynchronous policy iteration (see Sections 2.6.3, 3.3.2, and [BeY10a], [BeY10b], [YuB11a]), and in constrained forms of policy iteration (see [Ber11c], or [Ber12a], Exercise 2.7).

may view V_ℓ^* as an approximation to the optimal cost vector of the original MDP starting at states $x \in S_\ell$, and we may view R_ℓ^* as a form of aggregate cost for S_ℓ. The advantage of this formulation is that it involves significant decomposition and parallelization of the computations among the processors, when performing various DP algorithms. In particular, the computation of $W_\ell(x, u, V_\ell, R_1, \ldots, R_m)$ depends on just the local vector V_ℓ, whose dimension may be potentially much smaller than n.

1.2.4 Approximation-Related Mappings

Given an abstract DP model described by a mapping H, we may be interested in fixed points of related mappings other than T and T_μ. Such mappings may arise in various contexts; we have seen one that is related to distributed asynchronous aggregation in Example 1.2.9. An important context is subspace approximation, whereby T_μ and T are restricted onto a subspace of functions for the purpose of approximating their fixed points. Much of the theory of approximate DP and reinforcement learning relies on such approximations (see e.g., the books by Bertsekas and Tsitsiklis [BeT96], Sutton and Barto [SuB98], Gosavi [Gos03], Cao [Cao07], Chang, Fu, Hu, and Marcus [CFH07], Meyn [Mey07], Powell [Pow07], Borkar [Bor08], Haykin [Hay08], Busoniu, Babuska, De Schutter, and Ernst [BBD10], Szepesvari [Sze10], Bertsekas [Ber12a], and Vrabie, Vamvoudakis, and Lewis [VVL13]).

For an illustration, consider the approximate evaluation of the cost vector of a discrete-time Markov chain with states $i = 1, \ldots, n$. We assume that state transitions (i, j) occur at time k according to given transition probabilities p_{ij}, and generate a cost $\alpha^k g(i, j)$, where $\alpha \in (0, 1)$ is a discount factor. The cost function over an infinite number of stages can be shown to be the unique fixed point of the Bellman equation mapping $T : \Re^n \mapsto \Re^n$ whose components are given by

$$(TJ)(i) = \sum_{j=1}^{n} p_{ij}(u)\big(g(i, j) + \alpha J(j)\big), \qquad i = 1, \ldots, n, \ J \in \Re^n.$$

This is the same as the mapping T in the discounted finite-state MDP Example 1.2.2, except that we restrict attention to a single policy. Finding the cost function of a fixed policy is the important policy evaluation subproblem which arises prominently within the context of policy iteration.

The approximation of the fixed point of T is often based on the solution of lower-dimensional equations defined on the subspace $\{\Phi R \mid R \in \Re^s\}$ that is spanned by the columns of a given $n \times s$ matrix Φ. Two such approximating equations have been studied extensively (see [Ber12a], Chapter 6, for a detailed account and references; also [BeY07], [BeY09], [YuB10], [Ber11a] for extensions to abstract contexts beyond approximate DP). These are:

(a) The projected equation

$$\Phi R = \Pi_\xi T(\Phi R), \tag{1.18}$$

where Π_ξ denotes projection onto S with respect to a weighted Euclidean norm

$$\|J\|_\xi = \sum_{i=1}^{n} \xi_i J(i) \tag{1.19}$$

with $\xi = (\xi_1, \ldots, \xi_n)$ being a probability distribution with positive components.

(b) The aggregation equation

$$\Phi R = \Phi DT(\Phi R), \tag{1.20}$$

with D being an $s \times n$ matrix whose rows are restricted to be probability distributions; these are known as the disaggregation probabilities. Also, in this approach, the rows of Φ are restricted to be probability distributions; they are known as the aggregation probabilities.

We now see that solution of the projected equation (1.18) and the aggregation equation (1.20) amounts to finding a fixed point of the mappings $\Pi_\xi T$ and ΦDT, respectively. These mappings derive their structure from the DP operator T, so they have some DP-like properties, which can be exploited for analysis and computation.

An important fact is that the aggregation mapping ΦDT preserves the monotonicity and the sup-norm contraction property of T, while the projected equation mapping $\Pi_\xi T$ does not. The reason for preservation of monotonicity is the nonnegativity of the components of the matrices Φ and D. The reason for preservation of sup-norm contraction is that the matrices Φ and D are sup-norm nonexpansive, because their rows are probability distributions. In fact, it can be shown that the solution R of Eq. (1.20) can be viewed as the *exact* DP solution of an "aggregate" DP problem that represents a lower-dimensional approximation of the original (see [Ber12a], Section 6.5).

By contrast, the projected equation mapping $\Pi_\xi T$ need not be monotone, because the components of Π_ξ need not be nonnegative. Moreover while the projection Π_ξ is nonexpansive with respect to the projection norm $\|\cdot\|_\xi$, it need not be nonexpansive with respect to the sup-norm. As a result the projected equation mapping $\Pi_\xi T$ need not be a sup-norm contraction. These facts play a significant role in approximate DP methodology.

Let us also mention that multistep versions of the mapping T have been used widely for approximations, particularly in connection with the projected equation approach. For example, the popular temporal difference methods, such as TD(λ), LSTD(λ), and LSPE(λ) (see the book references on reinforcement learning and approximate DP cited earlier), are based on the mapping $T^{(\lambda)} : \Re^n \mapsto \Re^n$ whose components are given by

$$\big(T^{(\lambda)}J\big)(i) = (1-\lambda)\sum_{k=1}^{\infty} \lambda^{\ell-1}(T^\ell J)(i), \qquad i = 1, \ldots, n,\ J \in \Re^n,$$

for $\lambda \in (0,1]$, where T^ℓ is the ℓ-fold composition of T with itself ℓ times. Here the mapping $T^{(\lambda)}$ is used in place of T in the projected equation (1.18). In the context of the aggregation equation approach, a multistep method based on the mapping $T^{(\lambda)}$ is the λ-aggregation method, given for the case of hard aggregation in [Ber12a], Section 6.5, as well as other forms of aggregation (see [Ber12a], [YuB12]).

A more general form of multistep approach, introduced and studied in [YuB12], uses instead the mapping $T^{(w)} : \Re^n \mapsto \Re^n$, with components

$$\big(T^{(w)}J\big)(i) = \sum_{\ell=1}^{\infty} w_{i\ell}(T^\ell J)(i), \qquad i = 1, \ldots, n, \ J \in \Re^n,$$

where for each i, $(w_{i1}, w_{i2}, \ldots)$ is a probability distribution over the positive integers. Then the multistep analog of the projected Eq. (1.18) is

$$\Phi R = \Pi_\xi T^{(w)}(\Phi R), \tag{1.21}$$

while the multistep analog of the aggregation Eq. (1.20) is

$$\Phi R = \Phi D T^{(w)}(\Phi R). \tag{1.22}$$

The mapping $T^{(\lambda)}$ is obtained for $w_{i\ell} = (1-\lambda)\lambda^{\ell-1}$, independently of the state i. The solution of Eqs. (1.21) and (1.22) by simulation-based methods is discussed in [YuB12] and [Yu12].

In fact, a connection between projected equations of the form (1.21) and aggregation equations of the form (1.22) was established in [YuB12] through the use of a seminorm [this is given by the same expression as the norm $\|\cdot\|_\xi$ of Eq. (1.19), with some of the components of ξ allowed to be 0]. In particular, the most prominent classes of aggregation equations can be viewed as seminorm projected equations because it turns out that ΦD is a seminorm projection (see [Ber12a], p. 639, [YuB12], Section 4). Moreover they can be viewed as projected equations where the projection is oblique (see [Ber12a], Section 7.3.6).

The preceding observations are important for our purposes, as they indicate that much of the theory developed in this book applies to approximation-related mappings based on aggregation. However, this is not true to nearly the same extent for approximation-related mappings based on projection.

1.3 ORGANIZATION OF THE BOOK

The examples of the preceding section have illustrated how the monotonicity assumption is satisfied for many DP models, while the contraction assumption may or may not be satisfied. In particular, the contraction assumption is satisfied for the mapping H in Examples 1.2.1-1.2.5, assuming

that there is discounting and that the cost per stage is bounded, but it need not hold in the SSP Example 1.2.6 and the multiplicative Example 1.2.8.

The main theme of this book is that the presence or absence of monotonicity and contraction is the primary determinant of the analytical and algorithmic theory of a typical total cost DP model. In our development, with some minor exceptions, we will assume that monotonicity holds. Consequently, the rest of the book is organized around the presence or absence of the contraction property. In the next four chapters we will discuss the following four types of models.

(a) **Contractive models:** These models, discussed in Chapter 2, have the richest and strongest algorithmic theory, and are the benchmark against which the theory of other models is compared. Prominent among them are discounted stochastic optimal control problems (cf. Example 1.2.1), finite-state discounted MDP (cf. Example 1.2.2), and some special types of SSP problems (cf. Example 1.2.6).

(b) **Semicontractive models:** In these models T_μ is monotone but it need not be a contraction for all $\mu \in \mathcal{M}$. Instead policies are separated into those that "behave well" with respect to our optimization framework and those that do not. It turns out that the notion of contraction is not sufficiently general for our purposes. We will thus introduce a related notion of "regularity," which is based on the idea that a policy μ should be considered "well-behaved" if the dynamic system defined by T_μ has J_μ as an asymptotically stable equilibrium within some domain. Our models and analysis are patterned to a large extent after the SSP problems of Example 1.2.6 (the regular μ correspond to the proper policies). One of the complications here is that policies that are not regular, may have cost functions that take the value $+\infty$ or $-\infty$. Still under certain conditions, which directly or indirectly guarantee that there exists an optimal regular policy, the complications can be dealt with, and we can prove strong properties for these models, sometimes almost as strong as those of the contractive models.

(c) **Noncontractive models:** These models rely on just the monotonicity property of T_μ, and are more complex than the preceding ones. As in semicontractive models, the various cost functions of the problem may take the values $+\infty$ or $-\infty$, and the mappings T_μ and T must accordingly be allowed to deal with such functions. However, the optimal cost function may take the values ∞ and $-\infty$ as a matter of course (rather than on an exceptional basis, as in semicontractive models). The complications are considerable, and much of the theory of the contractive models generalizes in weaker form, if at all. For example, in general the fixed point equation $J = TJ$ need not

have a unique solution, the value iteration method may work starting with some functions but not with others, and the policy iteration method may not work at all. Of course some of these weaknesses may not appear in the presence of additional structure, and we will discuss noncontractive models that also have some semicontractive structure, and corresponding favorable properties.

(d) **Restricted Policies Models:** These models are variants of some of the preceding ones, where there are restrictions of the set of policies, so that $\mathcal{M}$ may be a strict subset of the set of functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$ for all $x \in X$. Such restrictions may include measurability (needed to establish a mathematically rigorous probabilistic framework) or special structure that enhances the characterization of optimal policies and facilitates their computation.

Examples of DP problems from each of the above model categories, mostly special cases of the specific DP models discussed in Section 1.2, are scattered throughout the book, both to illustrate the theory and its exceptions, and to illustrate the beneficial role of additional special structure. The discussion of algorithms centers on abstract forms of value and policy iteration, and is organized along three characteristics: *exact, approximate, and asynchronous*.

The exact algorithms represent idealized versions, the approximate represent implementations that use approximations of various kinds, and the asynchronous involve irregular computation orders, where the costs and controls at different states are updated at different iterations (for example the cost of a single state being iterated at a time, as in Gauss-Seidel and other methods; see [Ber12a]). Approximate and asynchronous implementations have been the subject of intensive investigations in the last twenty five years, in the context of the solution of large-scale problems. Some of this methodology relies on the use of simulation, which is asynchronous by nature and is prominent in approximate DP and reinforcement learning.

1.4 NOTES, SOURCES, AND EXERCISES

The connection between DP and fixed point theory may be traced to Shapley [Sha53], who exploited contraction mappings in analysis of the two-player dynamic game model of Example 1.2.4. Since that time the underlying contraction properties of discounted DP problems with bounded cost per stage have been explicitly or implicitly used by most authors that have dealt with the subject. Moreover, the value of the abstract viewpoint as the basis for economical and insightful analysis has been widely recognized.

An abstract DP model, based on unweighted sup-norm contraction assumptions, was introduced in the paper by Denardo [Den67]. This model pointed to the fundamental connections between DP and fixed point the-

ory, and provided generality and insight into the principal analytical and algorithmic ideas underlying the discounted DP research up to that time. Abstract DP ideas were also researched earlier, notably in the paper by Mitten (Denardo's Ph.D. thesis advisor) [Mit64]; see also Denardo and Mitten [DeM67]. The properties of monotone contractions were also used in the analysis of sequential games by Zachrisson [Zac64].

Denardo's model motivated a related abstract DP model by the author [Ber77], which relies only on monotonicity properties, and was patterned after the positive DP problem of Blackwell [Bla65] and the negative DP problem of Strauch [Str66]. These two abstract DP models were used extensively in the book by Bertsekas and Shreve [BeS78] for the analysis of both discounted and undiscounted DP problems, ranging over MDP, minimax, multiplicative, Borel space models, and models based on outer integration. Extensions of the analysis of [Ber77] were given by Verdu and Poor [VeP87], which considered additional structure that allows the development of backward and forward value iterations, and in the thesis by Szepesvari [Sze98a], [Sze98b], which introduced non-Markovian policies into the abstract DP framework. The model of [Ber77] was also used by Bertsekas [Ber82], and Bertsekas and Yu [BeY10b], to develop asynchronous value and policy iteration methods for abstract contractive and noncontractive DP models. Another line of related research involving abstract DP mappings that are not necessarily scalar-valued was initiated by Mitten [Mit74], and was followed up by a number of authors, including Sobel [Sob75], Morin [Mor82], and Carraway and Morin [CaM88].

Restricted policies models that aim to address measurability issues in the context of abstract DP were first considered in [BeS98]. Followup research on this highly technical subject has been limited, and some issues have not been fully worked out beyond the classical discounted, positive, and negative stochastic optimal control problems; see Chapter 5.

Generally, noncontractive total cost DP models with some special structure beyond monotonicity, fall in three major categories: monotone increasing models principally represented by negative DP, monotone decreasing models principally represented by positive DP, and transient models, exemplified by the SSP model of Example 1.2.6, where the decision process terminates after a period that is random and subject to control. Abstract DP models patterned after the first two categories have been known since [Ber77] and are further discussed in Section 4.3. The semicontractive models of Chapters 3 and 4, are patterned after the third category, and their analysis is based on the idea of separating policies into those that are well-behaved (have contraction-like properties) and those that are not (but their detrimental effects can be effectively limited thanks to the problem's structure). As far as the author knows, this idea is new in the context of abstract DP. One of the aims of the present monograph is to develop this idea and to show that it leads to an important and insightful paradigm for conceptualization and solution of major classes of practical DP problems.

E X E R C I S E S

1.1 (Multistep Contraction Mappings)

This exercise shows how starting with an abstract mapping, we can obtain multistep mappings with the same fixed points and a stronger contraction modulus. Consider a set of mappings $T_\mu : B(X) \mapsto B(X)$, $\mu \in \mathcal{M}$, satisfying the contraction Assumption 1.2.2, let m be a positive integer, and let $\mathcal{M}_m$ be the set of m-tuples $\nu = (\mu_0, \ldots, \mu_{m-1})$, where $\mu_k \in \mathcal{M}$, $k = 1, \ldots, m-1$. For each $\nu = (\mu_0, \ldots, \mu_{m-1}) \in \mathcal{M}_m$, define the mapping $\overline{T}_\nu$, by

$$\overline{T}_\nu J = T_{\mu_0} \cdots T_{\mu_{m-1}} J, \qquad \forall\ J \in B(X).$$

Show that we have the contraction properties

$$\|\overline{T}_\nu J - \overline{T}_\nu J'\| \leq \alpha^m \|J - J'\|, \qquad \forall\ J, J' \in B(X), \tag{1.23}$$

and

$$\|\overline{T} J - \overline{T} J'\| \leq \alpha^m \|J - J'\|, \qquad \forall\ J, J' \in B(X), \tag{1.24}$$

where $\overline{T}$ is defined by

$$(\overline{T}J)(x) = \inf_{(\mu_0, \ldots, \mu_{m-1}) \in \mathcal{M}_m} (T_{\mu_0} \cdots T_{\mu_{m-1}} J)(x), \qquad \forall\ J \in B(X),\ x \in X.$$

1.2 (State-Dependent Weighted Multistep Mappings [YuB12])

Consider a set of mappings $T_\mu : B(X) \mapsto B(X)$, $\mu \in \mathcal{M}$, satisfying the contraction Assumption 1.2.2. Consider also the mappings $T_\mu^{(w)} : B(X) \mapsto B(X)$ defined by

$$(T_\mu^{(w)} J)(x) = \sum_{\ell=1}^{\infty} w_\ell(x) \big(T_\mu^\ell J\big)(x), \qquad x \in X,\ J \in B(X),$$

where $w_\ell(x)$ are nonnegative scalars such that for all $x \in X$,

$$\sum_{\ell=1}^{\infty} w_\ell(x) = 1.$$

Show that

$$\frac{\big|(T_\mu^{(w)} J)(x) - (T_\mu^{(w)} J')(x)\big|}{v(x)} \leq \sum_{\ell=1}^{\infty} w_\ell(x) \alpha^\ell \|J - J'\|, \qquad \forall\ x \in X,$$

where α is the contraction modulus of T_μ, so that $T_\mu^{(w)}$ is a contraction with modulus

$$\bar{\alpha} = \sup_{x \in X} \sum_{\ell=1}^{\infty} w_\ell(x)\, \alpha^\ell \leq \alpha.$$

Show also that T_μ and $T_\mu^{(w)}$ have a common fixed point for all $\mu \in \mathcal{M}$.