1

Introduction

Contents

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 "mathematical 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 one or both 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

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

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

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

† Our discussion of this section is somewhat informal, without strict adherence to mathematical notation and rigor. We will introduce a rigorous mathematical framework later.

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

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

The total cost of a policy $\pi = \{\mu_0, \mu_1, \cdots\}$ over an infinite number of stages (an infinite horizon) and starting at an initial state x_0 is the limit superior of the N-step costs

$$J_\pi(x_0) = \limsup_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k)\big) \tag{1.2}$$

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

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

(We use limit superior rather than limit to cover the case where the limit does not exist.) The optimal cost function is

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

For any policy $\pi = \{\mu_0, \mu_1, \cdots\}$, consider the policy $\pi_1 = \{\mu_1, \mu_2, \cdots\}$ 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 $\{\mu^*, \mu^*, \cdots\}$ may typically be obtained by minimization of the right-hand side for each x, i.e.,

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

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 functions on X into itself. Moreover, it can be similarly seen that J_μ, the cost function of the stationary policy $\{\mu,\mu,\cdots\}$, 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 stated and analyzed using the mappings T and T_μ.

The mappings T_μ can also be used in the context of DP problems with a finite number of stages (a finite horizon). In particular, for a given policy $\pi = \{\mu_0,\mu_1,\cdots\}$ 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 $T_{\mu_0}, T_{\mu_1}, \cdots, T_{\mu_{N-1}}$, 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 the infinite horizon cost function J_π:

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

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

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 a 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 just discussed.

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 $\pi = \{\mu_0, \mu_1, \cdots\}$, with $\mu_k \in \mathcal{M}$ for all k, as "nonstationary policies," and we refer to a sequence $\{\mu, \mu, \cdots\}$, 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 $\mathcal{R}(X)$ be the set of real-valued functions $J : X \mapsto \Re$, and let $H : X \times U \times \mathcal{R}(X) \mapsto \Re$ be a given mapping.† For each policy $\mu \in \mathcal{M}$, we consider the mapping $T_\mu : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ defined by

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

and we also consider the mapping T defined by‡

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

---

† 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 $J \in \mathcal{R}(X)$ in the present chapter and in Chapter 2. In Chapters 3 and 4 we will allow $H(x, u, J)$, and hence also $(T_\mu J)(x)$ and $(TJ)(x)$, to take the values ∞ and $-\infty$.

We will generally refer to T and T_μ as the (abstract) *DP mappings* or *DP operators* or *Bellman operators* (the latter name is common in the artificial intelligence and reinforcement learning literature).

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 $\bar{J} \in \mathcal{R}(X)$, and nonstationary policy $\pi = \{\mu_0, \mu_1, \cdots\}$, we define for each integer $N \geqslant 1$ 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}, $\cdots$, $T_{\mu_{N-1}}$, i.e.,

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

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, $\pi^* = \{\mu^*, \mu^*, \cdots\}$, with $\mu^* \in \mathcal{M}$ satisfying an optimality condition, such as for example

$$(T_{\mu^*} J^*)(x) = (T J^*)(x), \qquad x \in X$$

[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. We describe these assumptions next.

1.2.2 Monotonicity and Contraction Properties

Let us now formalize the monotonicity and contraction assumptions. We will require that both of these assumptions hold for most of the next chapter, and we will gradually relax the contraction assumption in Chapters 3 and 4. Recall also our assumption that T_μ and T map $\mathcal{R}(X)$ (the space of real-valued functions over X) into $\mathcal{R}(X)$. In Chapters 3 and 4 we will relax this assumption as well.

Assumption 1.2.1: (Monotonicity) If $J, J' \in \mathcal{R}(X)$ and $J \leqslant J'$, then

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

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

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

or equivalently, †

$$J \leqslant J' \qquad \Rightarrow \qquad TJ \leqslant TJ'$$

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

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

and to use the simple but important fact

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

i.e., for a fixed $x \in X$, *infimum over u is equivalent to infimum over μ*, which holds because of the definition $\mathcal{M} = \big\{\mu \mid \mu(x) \in U(x),\ \forall\ x \in X\big\}$, so that $\mathcal{M}$ can be viewed as the Cartesian product $\Pi_{x \in X} U(x)$. 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 $\mathcal{B}(X)$ 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

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

† Unless otherwise stated, in this book, inequalities involving functions, minima and infima of a collection of functions, and limits of function sequences are meant to be pointwise; see Appendix A for our notational conventions.

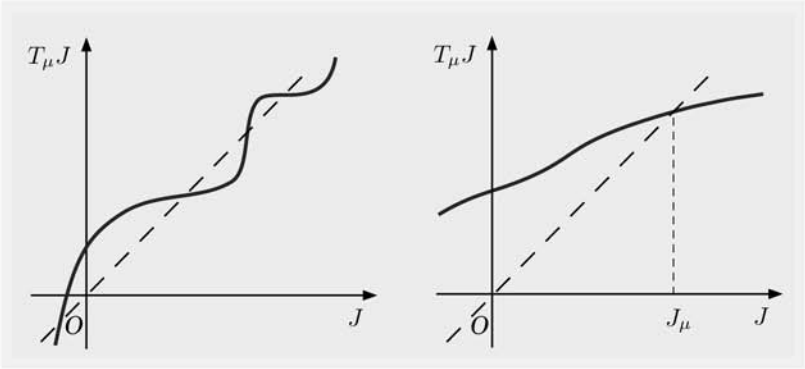

Figure 1.2.1. Illustration of the monotonicity and the contraction assumptions in one dimension. The mapping T_μ on the left is monotone but is not a contraction. The mapping T_μ on the right is both monotone and a contraction. It has a unique fixed point at J_μ.

on $\mathcal{B}(X)$. The properties of $\mathcal{B}(X)$ and some of the associated fixed point theory are discussed in Appendix B. In particular, as shown there, $\mathcal{B}(X)$ is a complete normed space, so any mapping from $\mathcal{B}(X)$ to $\mathcal{B}(X)$ 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 $J \in \mathcal{B}(X)$ and $\mu \in \mathcal{M}$, the functions $T_\mu J$ and TJ belong to $\mathcal{B}(X)$. Furthermore, for some $\alpha \in (0,1)$, we have

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

Figure 1.2.1 illustrates the monotonicity and the contraction assumptions. It can be shown that the contraction condition (1.8) implies that

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

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

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

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

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

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

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

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 $\{\mu, \mu, \cdots\}$, respectively [cf. Eq. (1.3)].

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

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

[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 $J \in \mathcal{B}(X)$.

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

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

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 most of 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 a 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

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

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, *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, $k = 0, 1, \cdots$, 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 $w_{k-1}, \cdots, w_0$.

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

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

subject to the system equation constraint

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

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,

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

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):

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

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

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

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-stage cost of π, can be expressed in terms of T_μ:

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

---

† 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-stage 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 while either $g \geqslant 0$ or $g \leqslant 0$ (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 Section 4.4).

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

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

[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 model, 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 $u = (u_1, \cdots, u_n)$ over his n possible moves and the maximizer selects a probability distribution $v = (v_1, \cdots, v_m)$ over his m possible moves. Thus the probability of selecting i and j is $u_i v_j$, and 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 $\{1, \cdots, n\}$ and $\{1, \cdots, m\}$, 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 $X = \{t, 1, \cdots, n\}$ 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 t is a special termination state, which is cost-free and absorbing, i.e., for all $u \in U(t)$,

$$g(t,u) = 0, \qquad p_{tt}(u) = 1$$

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

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

The mappings T_μ and T are defined by

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

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

Note that the matrix that has components $p_{xy}(u)$, $x,y = 1,\cdots,n$, is sub-stochastic (some of its row sums may be less than 1) because there may be a positive transition probability from a state x to the termination state t. 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], [Ber12a], and [Ber17a], 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

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

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], Chapter 3).

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, referred to as the *strong SSP conditions* in Section 3.5.1, it was shown in [BeT91] that T has a unique fixed point J^*, the optimal cost function of the SSP problem. Moreover, a policy $\{\mu^*, \mu^*, \cdots\}$ is optimal if and only if

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

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 [Ber12a], [YuB13a], and also Section 3.6.2, where these modifications will be discussed in an abstract setting.

In Section 3.5.1 we will also consider SSP problems where the strong SSP conditions (a) and (b) above are not satisfied. Then we will see that unusual phenomena can occur, including that J^* may not be a solution of Bellman's equation. Still our line of analysis of Chapter 3 will apply to such problems.

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 $x = 1, \cdots, n$, plus a destination t, 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

$$H(x, u, J) = \begin{cases} a_{xu} + J(u) & \text{if } u \neq t, \\ a_{xt} & \text{if } u = t, \end{cases} \qquad x = 1, \cdots, n$$

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 zero, 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).

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 satisfying $g(x,u,y) \geqslant 0$ for all x, y, u (this is necessary for H to be monotone). This mapping corresponds to the multiplicative model of minimizing over all $\pi = \{\mu_0, \mu_1, \cdots\}$ the cost

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

where the state sequence $\{x_0, x_1, \cdots\}$ 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

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

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], [BoM02], [BBB08], [Ber16a]. 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, based on the author's paper [Ber16a]; see the next example and Section 3.5.2.

Example 1.2.9 (Affine Monotonic Models)

Consider a finite state space $X = \{1,\cdots,n\}$ and a (possibly infinite) control constraint set $U(x)$ for each state x. For each policy μ, let the mapping T_μ be given by

$$T_\mu J = b_\mu + A_\mu J \tag{1.18}$$

where b_μ is a vector of $\Re^n$ with components $b\big(x,\mu(x)\big)$, $x = 1,\cdots,n$, and A_μ is an $n\times n$ matrix with components $A_{xy}\big(\mu(x)\big)$, $x,y = 1,\cdots,n$. We assume that $b(x,u)$ and $A_{xy}(u)$ are nonnegative,

$$b(x,u) \geqslant 0, \qquad A_{xy}(u) \geqslant 0, \qquad \forall\ x,y = 1,\cdots,n,\ u\in U(x)$$

Thus T_μ and T map nonnegative functions to nonnegative functions $J : X \mapsto [0, \infty]$.

This model was introduced in the first edition of this book, and was elaborated on in the author's paper [Ber16a]. Special cases of the model include the finite-state Markov and semi-Markov problems of Examples 1.2.1-1.2.3, and the stochastic shortest path problem of Example 1.2.6, with A_μ being the transition probability matrix of μ (perhaps appropriately discounted), and b_μ being the cost per stage vector of μ, which is assumed nonnegative. An interesting affine monotonic model of a different type is the multiplicative cost model of the preceding example, where the initial function is $\bar{J}(x) \equiv 1$ and the cost accumulates multiplicatively up to reaching a termination state t. In the exponential case of this model, the cost of a generated path starting from some initial state accumulates additively as in the SSP case, up to reaching t. However, the cost of the model is the expected value of the *exponentiated* cost of the path up to reaching t. It can be shown then that the mapping T_μ has the form

$$(T_\mu J)(x) = p_{xt}\big(\mu(x)\big)\exp\big(g(x,\mu(x),t)\big) + \sum_{y=1}^{n} p_{xy}(\mu(x))\exp\big(g(x,\mu(x),y)\big)J(y) \qquad x \in X,$$

where $p_{xy}(u)$ is the probability of transition from x to y under u, and $g(x,u,y)$ is the cost of the transition; see Section 3.5.2 for a detailed derivation. Clearly T_μ has the affine monotonic form (1.18).

Example 1.2.10 (Aggregation)

Aggregation is an approximation approach that replaces a large DP problem with a simpler problem obtained by "combining" many of its states together into *aggregate states*. This results in an "aggregate" problem with fewer states, which may be solvable by exact DP methods. The optimal cost-to-go function of this problem is then used to approximate the optimal cost function of the original problem.

Consider an n-state Markovian decision problem with transition probabilities $p_{ij}(u)$. To construct an aggregation framework, we introduce a finite set $\mathcal{A}$ of aggregate states. We generically denote the aggregate states by letters such as x and y, and the original system states by letters such as i and j. The approximation framework is specified by combining in various ways the aggregate states and the original system states to form a larger system (see Fig. 1.2.2). To specify the probabilistic structure of this system, we introduce two (somewhat arbitrary) choices of probability distributions, which relate the original system states with the aggregate states:

(1) For each aggregate state x and original system state i, we specify the *disaggregation probability* d_{xi}. We assume that $d_{xi} \geqslant 0$ and

$$\sum_{i=1}^{n} d_{xi} = 1, \qquad \forall\, x \in \mathcal{A}$$