# 3

# *Semicontractive Models*

**Contents**

We will now consider abstract DP models that are intermediate between the contractive models of Chapter 2, where all stationary policies involve a contraction mapping, and noncontractive models to be discussed in Chapter 4, where there are no contraction-like assumptions (although there are some compensating conditions, including monotonicity).

A representative instance of such an intermediate model is the deterministic shortest path problem of Example 1.2.7, where we can distinguish between two types of stationary policies: those that terminate at the destination from every starting node, and those that do not. A more general instance is the stochastic shortest path (SSP for short) problem of Example 1.2.6. In this problem, the analysis revolves around two types of stationary policies $\mu$: those with a mapping $T_\mu$ that is a contraction with respect to some norm, and those with a mapping $T_\mu$ that is not a contraction with respect to any norm (it can be shown that the former are the ones that terminate with probability 1 starting from any state).

In the models of this chapter, like in SSP problems, we divide policies into two groups, one of which has favorable characteristics. We loosely refer to such models as *semicontractive* to indicate that these favorable characteristics include contraction-like properties of the mapping $T_\mu$. To develop a more broadly applicable theory, we replace the notion of contractiveness of $T_\mu$ with a notion of *$S$-regularity of $\mu$*, where $S$ is an appropriate set of functions of the state (roughly, this is a form of "local stability" of $T_\mu$, which ensures that the cost function $J_\mu$ is the unique fixed point of $T_\mu$ within $S$, and that $T_\mu^k J$ converges to $J_\mu$ regardless of the choice of $J$ from within $S$). We allow that some policies are $S$-regular while others are not.

Note that the term "semicontractive" is not used in a precise mathematical sense here. Rather it refers qualitatively to a collection of models where some policies have a regularity/contraction-like property but others do not. Moreover, regularity is a relative property: the division of policies into "regular" and "irregular" depends on the choice of the set $S$. On the other hand, typically in practical applications an appropriate choice of $S$ is fairly evident.

Our analysis will involve two types of assumptions:

(a) *Favorable assumptions*, under which we obtain results that are nearly as strong as those available for the contractive models of Chapter 2. In particular, we show that $J^*$ is a fixed point of $T$, that the Bellman equation $J = TJ$ has a unique solution, at least within a suitable class of functions, and that variants of the VI and PI algorithms are valid. Some of the VI and PI approaches are suitable for distributed asynchronous computation, similar to their Chapter 2 counterparts for contractive models.

(b) *Less favorable assumptions*, under which serious difficulties may occur: $J^*$ may not be a fixed point of $T$, and even when it is, it may not be found using the VI and PI algorithms. These anomalies may appear in simple problems, such as deterministic and stochastic shortest path problems with some zero length cycles. To address the difficulties, we will consider a restricted problem, where the only admissible policies are the ones that are $S$-regular. Under reasonable conditions we show that this problem is better-behaved. In particular, $J_S^*$, the optimal cost function over the $S$-regular policies only, is the unique solution of Bellman's equation among functions $J \in S$ with $J \geqslant J_S^*$, while VI converges to $J_S^*$ starting from any $J \in S$ with $J \geqslant J_S^*$. We will also derive a variety of PI approaches for finding $J_S^*$ and an $S$-regular policy that is optimal within the class of $S$-regular policies.

We will illustrate our analysis in Section 3.5, both under favorable and unfavorable assumptions, by means of four classes of practical problems. Some of these problems relate to finding a path to a destination in a graph under stochastic or set membership uncertainty, while others relate to the control of a continuous-state system to a terminal state. In particular, we will consider SSP problems, affine monotonic problems, including problems with multiplicative or risk-sensitive exponential cost function, minimax-type shortest path problems, and continuous-state deterministic problems with nonnegative cost, such as linear-quadratic problems.

The chapter is organized as follows. In Section 3.1, we illustrate the pathologies regarding solutions of Bellman's equation, and the VI and PI algorithms. To this end, we use four simple examples, ranging from finite-state shortest path problems, to continuous-state linear-quadratic problems. These examples provide orientation and motivation for $S$-regular policies later. In Section 3.2, we formally introduce our abstract DP model, and the notion of an $S$-regular policy. We then develop some of the basic associated results relating to Bellman's equation, and the convergence of VI and PI, based primarily on the ideas underlying the PI algorithm. In Section 3.3 we refine the results of Section 3.2 under favorable conditions, obtaining results and algorithms that are almost as powerful as the ones for contractive models. In Section 3.4 we develop a complementary analytical approach, which is based on the use of perturbations and applies under less favorable assumptions. In Section 3.5, we discuss in detail the application and refinement of the results of Sections 3.2-3.4 in some important shortest path-type practical contexts. In Section 3.6, we focus on variants of VI and PI-type algorithms for semicontractive DP models, including some that are suitable for asynchronous distributed computation.

## 3.1 Pathologies of Noncontractive DP Models

In this section we provide a general overview of the analytical and computational difficulties in noncontractive DP models, using for the most part shortest path-type problems. For illustration we will first use two of the simplest and most widely encountered finite-state DP problems: deterministic and SSP problems, whereby we are aiming to reach a destination state at minimum cost.† We will also discuss an example of continuous-state shortest path problem that involves a linear system and a quadratic cost function.

We will adopt the general abstract DP model of Section 1.2. We give a brief description that is adequate for the purposes of this section, and defer a more formal definition to Section 3.2. In particular, we introduce a set of states $X$, and for each $x \in X$, the nonempty control constraint set $U(x)$. For each policy $\mu$, the mapping $T_\mu$ is given by

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

where $H$ is a suitable function of $(x, u, J)$. The mapping $T$ is given by

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

The cost function of a policy $\pi = \{\mu_0, \mu_1, \cdots\}$ is

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

where $\bar{J}$ is some function.‡ We want to minimize $J_\pi$ over $\pi$, i.e., to find

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

and a policy that attains the infimum.

For orientation purposes, we recall from Chapter 1 (Examples 1.2.1 and 1.2.2) that for a stochastic optimal control problem involving a finite-state Markov chain with state space $X = \{1, \cdots, n\}$, transition probabilities $p_{xy}(u)$, and expected one-stage cost function $g$, the mapping $H$ is given by

$$H(x, u, J) = g(x, u) + \sum_{y=1}^{n} p_{xy}(u) J(y), \qquad x \in X$$

and $\bar{J}(x) \equiv 0$. The SSP problem arises when there is an additional termination state that is cost-free, and corresponding transition probabilities $p_{xt}(u)$, $x \in X$.

---

† These problems are naturally undiscounted, and cannot be readily addressed by introducing a discount factor close to 1, because then the optimal policies may exhibit undesirable behavior. In particular, in the presence of discounting, they may involve moving initially along a small-length cycle in order to postpone the use of an optimal but unavoidably costly path until later, when the discount factor will reduce substantially the cost of that path.

‡ In the contractive models of Chapter 2, the choice of $\bar{J}$ is immaterial, as we discussed in Section 2.1. Here, however, the choice of $\bar{J}$ is important, and affects important characteristics of the model, as we will see later.

A more general undiscounted stochastic optimal control problem involves a stationary discrete-time dynamic system where the state is an element of a space $X$, and the control is an element of a space $U$. The control $u_k$ is constrained to take values in a given set $U(x_k) \subset U$, which depends on the current state $x_k$ [$u_k \in U(x_k)$, for all $x_k \in X$]. For a policy $\pi = \{\mu_0, \mu_1, \cdots\}$, the state evolves according to a system equation

$$x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big), \qquad k = 0, 1, \cdots \tag{3.1}$$

where $w_k$ is a random disturbance that takes values from a space $W$. We assume that $w_k$, $k = 0, 1, \cdots$, 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. Here, we allow 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.†

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

$$J_\pi(x_0) = \limsup_{k\to\infty} E\left\{\sum_{t=0}^{k} g\big(x_t, \mu_t(x_t), w_t\big)\right\} \tag{3.2}$$

subject to the system equation constraint (3.1), where $g$ is the one-stage cost function. The corresponding mapping of the abstract DP problem is

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

and $\bar{J}(x) \equiv 0$. Again here, $(T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x)$ is the expected cost of the first $k + 1$ periods using $\pi$ starting from $x$, and with terminal cost 0.

A discounted version of the problem is defined by the mapping

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

where $\alpha \in (0, 1)$ is the discount factor. It corresponds to minimization of

$$J_\pi(x_0) = \limsup_{k\to\infty} E\left\{\sum_{t=0}^{k} \alpha^t g\big(x_t, \mu_t(x_t), w_t\big)\right\}$$

If the cost per stage $g$ is bounded, then a problem that fits the contractive framework of Chapter 2 is obtained, and can be analyzed using the methods of that chapter. However, there are interesting infinite-state discounted optimal control problems where $g$ is not bounded.

---

† Measure-theoretic issues are not addressed at all in this second edition of the book. The first edition addressed some of these issues within an abstract DP context in its Chapter 5 and Appendix C (this material is posted at the book's web site); see also the monograph by Bertsekas and Shreve [BeS78], and the paper by Yu and Bertsekas [YuB15].

### A Summary of Pathologies

The four examples to be discussed in Sections 3.1.1-3.1.4 are special cases of deterministic and stochastic optimal control problems of the type just described. In each of these examples, we will introduce a subclass of "well-behaved" policies and a restricted optimization problem, which is to minimize the cost over the "well-behaved" subclass (in Section 3.2 the property of being "well-behaved" will be formalized through the notion of $S$-regularity). The optimal cost function over just the "well-behaved" policies is denoted $\hat{J}$ (we will also use the notation $J_S^*$ later). Here is a summary of the examples and the pathologies that they reveal:

(a) *A finite-state, finite-control deterministic shortest path problem (Section 3.1.1).* Here the mapping $T$ can have infinitely many fixed points, including $J^*$ and $\hat{J}$. There exist policies that attain the optimal costs $J^*$ and $\hat{J}$. Depending on the starting point, the VI algorithm may converge to $J^*$ or to $\hat{J}$ or to a third fixed point of $T$ (for cases where $J^* \neq \hat{J}$, VI converges to $\hat{J}$ starting from any $J \geqslant \hat{J}$). The PI algorithm can oscillate between two policies that attain $J^*$ and $\hat{J}$, respectively.

(b) *A finite-state, finite-control stochastic shortest path problem (Section 3.1.2).* The salient feature of this example is that $J^*$ is not a fixed point of the mapping $T$. By contrast $\hat{J}$ is a fixed point of $T$. The VI algorithm converges to $\hat{J}$ starting from any $J \geqslant \hat{J}$, while it does not converge otherwise.

(c) *A finite-state, infinite-control stochastic shortest path problem (Section 3.1.3).* We give three variants of this example. In the first variant (a classical problem known as the "blackmailer's dilemma"), all the policies are "well-behaved," so $J^* = \hat{J}$, and VI converges to $J^*$ starting from any real-valued initial condition, while PI also succeeds in finding $J^*$ as the limit of the generated sequence $\{J_{\mu^k}\}$. However, PI cannot find an optimal policy, because there is no optimal stationary policy. In a second variant of this example, PI generates a sequence of "well-behaved" policies $\{\mu^k\}$ such that $J_{\mu^k} \downarrow \hat{J}$, but $\{\mu^k\}$ converges to a policy that is either infeasible or is strictly suboptimal. In the third variant of this example, the problem data can strongly affect the multiplicity of the fixed points of $T$, and the behavior of the VI and PI algorithms.

(d) *A continuous-state, continuous-control deterministic linear-quadratic problem (Section 3.1.4).* Here the mapping $T$ has exactly two fixed points, $J^*$ and $\hat{J}$, within the class of positive semidefinite quadratic functions. The VI algorithm converges to $\hat{J}$ starting from all positive initial conditions, and to $J^*$ starting from all other initial conditions. Moreover, starting with a "well-behaved" policy, the PI algorithm converges to $\hat{J}$ and to an optimal policy within the class of "well-behaved" policies.

It can be seen that the examples exhibit wide-ranging pathological behavior. In Section 3.2, we will aim to construct a theoretical framework that explains this behavior. Moreover, in Section 3.3, we will derive conditions guaranteeing that much of this type of behavior does not occur. These conditions are natural and broadly applicable. They are used to exclude from optimality the policies that are not "well-behaved," and to obtain results that are nearly as powerful as their counterparts for the contractive models of Chapter 2.

### 3.1.1 Deterministic Shortest Path Problems

Let us consider the classical deterministic shortest path problem, discussed in Example 1.2.7. 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 $(x, y)$. The objective is to find for each $x$ a directed path that starts at $x$, ends at $t$, and has minimum length (the length of a path is defined as the sum of the lengths of its arcs). A standard assumption, which we will adopt here, is that every node $x$ is connected to the destination, i.e., there exists a path from every $x$ to $t$.

To formulate this shortest path problem as a DP problem, we embed it within a "larger" problem, whereby we view all paths as admissible, including those that do not terminate at $t$. We also view $t$ as a cost-free and absorbing node. Of course, we need to deal with the presence of policies that do not terminate, and the most common way to do this is to assume that all cycles have strictly positive length, in which case policies that do not terminate cannot be optimal. However, it is not uncommon to encounter shortest path problems with zero length cycles, and even negative length cycles. Thus we will not impose any assumption on the sign of the cycle lengths, particularly since we aim to use the shortest path problem to illustrate behavior that arises in a broader undiscounted/noncontractive DP setting.

As noted in Section 1.2, we can formulate the problem in terms of an abstract DP model where the states are the nodes $x = 1, \cdots, n$, and 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 mapping $H$ that defines the corresponding abstract DP problem 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 $\mu$ defines the subgraph whose arcs are $\big(x, \mu(x)\big)$, $x = 1, \ldots, n$. We say that $\mu$ is *proper* if this graph is acyclic, i.e., it consists of a tree of paths leading from each node to the destination. If $\mu$ is not

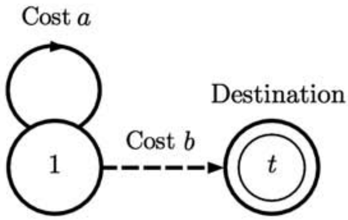


**Figure 3.1.1.** A deterministic shortest path problem with a single node 1 and a termination node $t$. At 1 there are two choices; a self-transition, which costs $a$, and a transition to $t$, which costs $b$.

proper, it is called *improper*. Thus there exists a proper policy if and only if each node is connected to $t$ with a path. Furthermore, an improper policy has cost greater than $-\infty$ starting from every initial state if and only if all the cycles of the corresponding subgraph have nonnegative cycle cost.

Let us now get a sense of what may happen by considering the simple one-node example shown in Fig. 3.1.1. Here there is a single state 1 in addition to the termination state $t$. At state 1 there are two choices: a self-transition, which costs $a$, and a transition to $t$, which costs $b$. The mapping $H$, abbreviating $J(1)$ with just the scalar $J$, is

$$H(1,u,J) = \begin{cases} a+J & \text{if } u\text{: self transition} \\ b & \text{if } u\text{: transition to } t \end{cases} \qquad J \in \Re$$

There are two policies here: the policy $\mu$ that transitions from 1 to $t$, which is proper, and the policy $\mu'$ that self-transitions at state 1, which is improper. We have

$$T_\mu J = b, \quad T_{\mu'} J = a + J, \qquad J \in \Re$$

and

$$TJ = \min\{b,\, a+J\}, \qquad J \in \Re$$

Note that for the proper policy $\mu$, the mapping $T_\mu : \Re \mapsto \Re$ is a contraction. For the improper policy $\mu'$, the mapping $T_{\mu'} : \Re \mapsto \Re$ is not a contraction, and it has a fixed point within $\Re$ only if $a = 0$, in which case every $J \in \Re$ is a fixed point.

We now consider the optimal cost $J^*$, the fixed points of $T$ within $\Re$, and the behavior of the VI and PI methods for different combinations of values of $a$ and $b$.

(a) If $a > 0$, the optimal cost, $J^* = b$, is the unique fixed point of $T$, and the proper policy is optimal.

(b) If $a = 0$, the set of fixed points of $T$ (within $\Re$) is the interval $(-\infty, b]$. Here the improper policy is optimal if $b \geqslant 0$, and the proper policy is optimal if $b \leqslant 0$ (both policies are optimal if $b = 0$).

(c) If $a = 0$ and $b > 0$, the proper policy is strictly suboptimal, yet its cost at state 1 (which is $b$) is a fixed point of $T$. The optimal cost, $J^* = 0$, lies in the interior of the set of fixed points of $T$, which is $(-\infty, b]$. Thus the VI method that generates $\{T^k J\}$ starting with $J \neq J^*$ cannot find $J^*$. In particular if $J$ is a fixed point of $T$, VI stops at $J$, while if $J$ is not a fixed point of $T$ (i.e., $J > b$), VI terminates in two iterations at $b \neq J^*$. Moreover, the standard PI method is unreliable in the sense that starting with the suboptimal proper policy $\mu$, it may stop with that policy because $T_\mu J_\mu = b = \min\{b, J_\mu\} = T J_\mu$ (the improper/optimal policy $\mu'$ also satisfies $T_{\mu'} J_\mu = T J_\mu$, so a rule for breaking the tie in favor of $\mu$ is needed but such a rule may not be obvious in general).

(d) If $a = 0$ and $b < 0$, the improper policy is strictly suboptimal, and we have $J^* = b$. Here it can be seen that the VI sequence $\{T^k J\}$ converges to $J^*$ for all $J \geqslant b$, but stops at $J$ for all $J < b$, since the set of fixed points of $T$ is $(-\infty, b]$. Moreover, starting with either the proper policy or the improper policy, the standard form of PI may oscillate, since $T_\mu J_{\mu'} = T J_{\mu'}$ and $T_{\mu'} J_\mu = T J_\mu$, as can be easily verified [the optimal policy $\mu$ also satisfies $T_\mu J_\mu = T J_\mu$ but it is not clear how to break the tie; compare also with case (c) above].

(e) If $a < 0$, the improper policy is optimal and we have $J^* = -\infty$. There are no fixed points of $T$ within $\Re$, but $J^*$ is the unique fixed point of $T$ within the set $[-\infty, \infty]$. The VI method will converge to $J^*$ starting from any $J \in [-\infty, \infty]$. The PI method will also converge to the optimal policy starting from either policy.

### 3.1.2 Stochastic Shortest Path Problems

We consider the SSP problem, which was described in Example 1.2.6 and will be revisited in Section 3.5.1. Here a policy is associated with a stationary Markov chain whose states are $1, \cdots, n$, plus the cost-free termination state $t$. The cost of a policy starting at a state $x$ is the sum of the expected cost of its transitions up to reaching $t$. A policy is said to be *proper*, if in its Markov chain, every state is connected with $t$ with a path of positive probability transitions, and otherwise it is called *improper*. Equivalently, a policy is proper if its Markov chain has $t$ as its unique ergodic state, with all other states being transient.

In deterministic shortest path problems, it turns out that $J_\mu$ is always a fixed point of $T_\mu$, and $J^*$ is always a fixed point of $T$. This is a generic feature of deterministic problems, which was illustrated in Section 1.1 (see Exercise 3.1 for a rigorous proof). However, in SSP problems where the

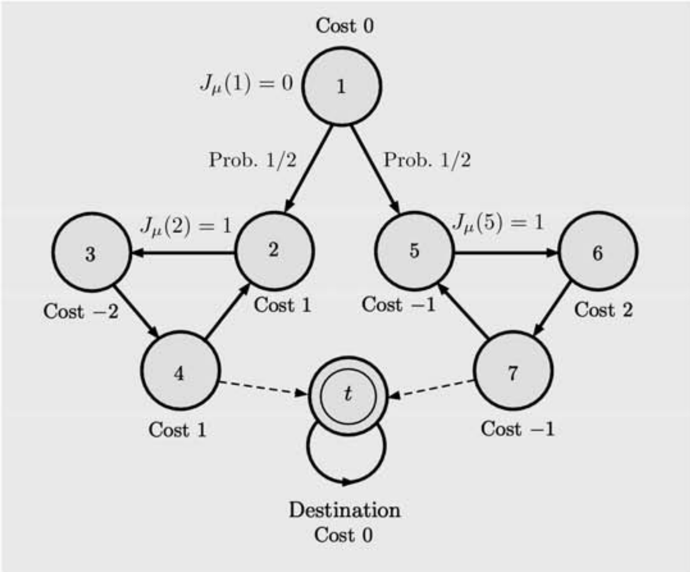


**Figure 3.1.2.** An example of an improper policy $\mu$, where $J_\mu$ is not a fixed point of $T_\mu$. All transitions under $\mu$ are shown with solid lines. These transitions are deterministic, except at state 1 where the next state is 2 or 5 with equal probability 1/2. There are additional high cost transitions from nodes 1, 4, and 7 to the destination (shown with broken lines), which create a suboptimal proper policy. We have $J^* = J_\mu$ and $J^*$ is not a fixed point of $T$.

cost per stage can take both positive and negative values this need not be so, as we will now show with an example due to [BeY16].

Let us consider the problem of Fig. 3.1.2. It involves an improper policy $\mu$, whose transitions are shown with solid lines in the figure, and form the two zero length cycles shown. All the transitions under $\mu$ are deterministic, except at state 1 where the successor state is 2 or 5 with equal probability 1/2. The problem has been deliberately constructed so that corresponding costs at the nodes of the two cycles are negatives of each other. As a result, the expected cost at each time period starting from state 1 is 0, implying that the total cost over any number or even infinite number of periods is 0.

Indeed, to verify that $J_\mu(1) = 0$, let $c_k$ denote the cost incurred at time $k$, starting at state 1, and let $s_N(1) = \sum_{k=0}^{N-1} c_k$ denote the $N$-step accumulation of $c_k$ starting from state 1. We have

$$s_N(1) = 0 \qquad \text{if } N = 1 \text{ or } N = 4 + 3t,\ t = 0, 1, \cdots$$

$$s_N(1) = 1 \text{ or } s_N(1) = -1 \text{ with probability } 1/2 \text{ each}$$
$$\text{if } N = 2 + 3t \text{ or } N = 3 + 3t,\ t = 0, 1, \cdots$$