Chapter 4
Infinite Horizon Dynamic Programming

SELECTED SECTIONS
4

Infinite Horizon Dynamic Programming

<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1. An Overview of Infinite Horizon Problems</td>
<td>174</td>
</tr>
<tr>
<td>4.2. Stochastic Shortest Path Problems</td>
<td>177</td>
</tr>
<tr>
<td>4.3. Discounted Problems</td>
<td>187</td>
</tr>
<tr>
<td>4.4. Semi-Markov Discounted Problems</td>
<td>192</td>
</tr>
<tr>
<td>4.5. Asynchronous Distributed Value Iteration</td>
<td>197</td>
</tr>
<tr>
<td>4.6. Policy Iteration</td>
<td>200</td>
</tr>
<tr>
<td>4.6.1. Exact Policy Iteration</td>
<td>200</td>
</tr>
<tr>
<td>4.6.2. Optimistic and Multistep Lookahead Policy Iteriation</td>
<td>205</td>
</tr>
<tr>
<td>4.6.3. Policy Iteration for Q-factors</td>
<td>208</td>
</tr>
<tr>
<td>4.7. Notes and Sources</td>
<td>209</td>
</tr>
<tr>
<td>4.8. Appendix: Mathematical Analysis</td>
<td>211</td>
</tr>
<tr>
<td>4.8.1. Proofs for Stochastic Shortest Path Problems</td>
<td>212</td>
</tr>
<tr>
<td>4.8.2. Proofs for Discounted Problems</td>
<td>217</td>
</tr>
<tr>
<td>4.8.3. Convergence of Exact and Optimistic Policy Iteration</td>
<td>218</td>
</tr>
</tbody>
</table>
In this chapter, we provide an introduction to the theory of infinite horizon problems. We focus on exact solution by DP methods, leaving the discussion of approximations to subsequent chapters. Infinite horizon problems differ from their finite horizon counterparts in two main respects:

(a) The number of stages is infinite.
(b) The system is stationary, i.e., the system equation, the cost per stage, and the disturbance probability distributions do not change from one stage to the next.

The assumption of an infinite number of stages is never satisfied in practice, but is a reasonable approximation for problems involving a finite but very large number of stages. The assumption of stationarity is often satisfied in practice, and in other cases it approximates well a situation where the system parameters vary relatively slowly with time.

Infinite horizon problems give rise to elegant and insightful analysis, and their optimal policies are often simpler than their finite horizon counterparts. For example, optimal policies are typically stationary, i.e., the optimal rule for choosing controls does not change from one stage to the next.

On the other hand, infinite horizon problems generally require a more sophisticated mathematical treatment. Our discussion will be limited to relatively simple finite-state problems (the theory of problems with an infinite state space is considerably more intricate; see the books [BeS78], [Ber12], [Ber18a]). Still some theoretical results are needed in this chapter. They will be explained intuitively to the extent possible, and their mathematical proofs will be provided in the end-of-chapter appendix.

### 4.1 AN OVERVIEW OF INFINITE HORIZON PROBLEMS

We will focus on two types of infinite horizon problems, where we aim to minimize the total cost over an infinite number of stages, given by

\[
J_\pi(x_0) = \lim_{N \to \infty} E_k \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\};
\]

see Fig. 4.1.1. Here, \(J_\pi(x_0)\) denotes the cost associated with an initial state \(x_0\) and a policy \(\pi = \{\mu_0, \mu_1, \ldots\}\), and \(\alpha\) is a scalar in the interval \([0, 1]\). When \(\alpha\) is strictly less that 1, it has the meaning of a discount factor, and its effect is that future costs matter to us less than the same costs incurred at the present time.

Thus the infinite horizon cost of a policy is the limit of its finite horizon cost as the horizon tends to infinity. (We assume that the limit exists and is finite for the moment, and address the issue later.) The two types of problems, considered in Sections 4.2 and 4.3, respectively, are:
Sec. 4.1 An Overview of Infinite Horizon Problems

Figure 4.1.1 Illustration of an infinite horizon problem. The system and cost per stage are stationary, except for the use of a discount factor $\alpha$. If $\alpha = 1$, there is typically a special cost-free termination state that we aim to reach.

(a) Stochastic shortest path problems (SSP for short). Here, $\alpha = 1$ but there is a special cost-free termination state; once the system reaches that state it remains there at no further cost. We will assume a problem structure such that termination is inevitable. Thus the horizon is in effect finite, but its length is random and may be affected by the policy being used.

(b) Discounted problems. Here, $\alpha < 1$ and there need not be a termination state. However, we will see that a discounted problem can be readily converted to an SSP problem. This can be done by introducing an artificial termination state to which the system moves with probability $1 - \alpha$ at every stage, thus making termination inevitable. As a result, our algorithms and analysis for SSP problems can be easily adapted to discounted problems.

A Preview of Infinite Horizon Theory

There are several analytical and computational issues regarding our infinite horizon problems. Many of them revolve around the relation between the optimal cost function $J^*$ of the infinite horizon problem and the optimal cost functions of the corresponding $N$-stage problems.

In particular, consider the SSP case and let $J_N(x)$ denote the optimal cost of the problem involving $N$ stages, initial state $x$, cost per stage $g(x, u, w)$, and zero terminal cost. This cost is generated after $N$ iterations of the DP algorithm

$$ J_{k+1}(x) = \min_{u \in U(x)} E \left\{ g(x, u, w) + J_k(f(x, u, w)) \right\}, \quad k = 0, 1, \ldots, \quad (4.1) $$

starting from the initial condition $J_0(x) = 0$ for all $x$.† The algorithm (4.1) is known as the value iteration algorithm (VI for short). Since the infinite horizon DP algorithm of Chapter 1. However, we have reversed the time indexing to suit our purposes. Thus the index of the cost functions produced by the algorithm is incremented with each iteration, and not decremented as in the case of finite horizon.

† This is just the finite horizon DP algorithm of Chapter 1. However, we have reversed the time indexing to suit our purposes. Thus the index of the cost functions produced by the algorithm is incremented with each iteration, and not decremented as in the case of finite horizon.
horizon cost of a given policy is, by definition, the limit of the corresponding
$N$-stage costs as $N \to \infty$, it is natural to speculate that:

1. The optimal infinite horizon cost is the limit of the corresponding
$N$-stage optimal costs as $N \to \infty$; i.e.,

$$J^*(x) = \lim_{N \to \infty} J_N(x) \quad (4.2)$$

for all states $x$.

2. The following equation should hold for all states $x$,

$$J^*(x) = \min_{u \in U(x)} \mathbb{E} \left\{ g(x, u, w) + J^*(f(x, u, w)) \right\} \quad (4.3)$$

This is obtained by taking the limit as $N \to \infty$ in the VI algorithm
(4.1) using Eq. (4.2). Equation (4.3) is really a system of equations
(one equation per state $x$), which has as solution the optimal costs-
to-go of all the states. It can also be viewed as a functional equation
for the optimal cost function $J^*$, and it is called Bellman’s equation.

3. If $\mu(x)$ attains the minimum in the right-hand side of the Bellman
equation (4.3) for each $x$, then the policy $\{\mu, \mu, \ldots\}$ should be optimal.
This type of policy is called stationary. Intuitively, optimal policies
can be found within this class of policies, since the future optimization
problem when starting at a given state looks the same regardless of
the time when we start.

All three of the preceding results hold for SSP problems under our
assumptions, as we will state later in Section 4.2 and prove in the appendix
to this chapter. They also hold for discounted problems in suitably modified
form that incorporates the discount factor. In fact the algorithms and
analysis of this chapter are quite similar for discounted problems and SSP
problems (under our assumptions), to the point where we may discuss a
particular method for one of the two problems with the understanding that
its application to the other problem can be straightforwardly adapted.

Transition Probability Notation for Finite-State Infinite Horizon
Problems

Throughout this chapter we assume a finite-state discrete-time dynamic
system, and we will use a special transition probability notation that is
suitable for such a system. We generally denote states by the symbol $i$ and
successor states by the symbol $j$. We will assume that there are $n$ states
(in addition to the termination state for SSP problems). These states are
denoted $1, \ldots, n$, and the termination state is denoted $t$. The control $u$
is constrained to take values in a given finite constraint set $U(i)$, which may
depend on the current state $i$. The use of a control $u$ at state $i$ specifies the transition probability $p_{ij}(u)$ to the next state $j$, at a cost $g(i, u, j)$.†

Given an admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$ (one with $\mu_k(i) \in U(i)$ for all $i$ and $k$) and an initial state $i_0$, the system becomes a Markov chain whose generated trajectory under $\pi$, denoted $\{i_0, i_1, \ldots\}$, has a well-defined probability distribution. The total expected cost associated with an initial state $i$ is

$$J_\pi(i) = \lim_{N \to \infty} E \left\{ \sum_{k=0}^{N-1} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) \mid i_0 = i, \pi \right\},$$

where $\alpha$ is either 1 (for SSP problems) or less than 1 for discounted problems. The expected value is taken with respect to the joint distribution of the states $i_1, i_2, \ldots$, conditioned on $i_0 = i$ and the use of $\pi$. The optimal cost from state $i$, i.e., the minimum of $J_\pi(i)$ over all policies $\pi$, is denoted by $J^*(i)$.‡

The cost function of a stationary policy $\pi = \{\mu, \mu, \ldots\}$ is denoted by $J_\mu(i)$. For brevity, we refer to $\pi$ as the stationary policy $\mu$. We say that $\mu$ is optimal if

$$J_\mu(i) = J^*(i) = \min_\pi J_\pi(i), \quad \text{for all states } i.$$

As noted earlier, under our assumptions, we will show that there will always exist an optimal policy, which is stationary.

† To convert from the transition probability format to the system equation format used in the preceding chapters, we can simply use the system equation $x_{k+1} = w_k$, where $w_k$ is the disturbance that takes values according to the transition probabilities $p_{xk, w_k}(u_k)$.

‡ Because of the underlying Markov chain, stochastic DP problems are called Markovian decision problems by many authors, including by Bellman. We will use instead the term “DP problem” to refer to both deterministic and stochastic problems.