CHAPTER 8

Average Reward and Related Criteria

When decisions are made frequently, so that the discount rate is very close to 1, or when performance criterion cannot easily be described in economic terms, the decision maker may prefer to compare policies on the basis of their average expected reward instead of their expected total discounted reward. Consequently, the average reward criterion occupies a cornerstone of queueing control theory especially when applied to controlling computer systems and communications networks. In such systems, the controller makes frequent decisions and usually assesses system performance on the basis of throughput rate or the average time a job or packet remains in the system. This optimality criterion may also be appropriate for inventory systems with frequent restocking decisions.

In this chapter, we focus on models with the expected average reward optimality criterion. We consider models with discrete-state spaces and focus primarily on finite-state models. Because the average reward criterion depends on the limiting behavior of the underlying stochastic processes, we distinguish models on the basis of this limiting behavior. Consequently we classify models on the basis of the chain structure of the class of stationary policies, and analyze these different model classes separately.

To illustrate key points and avoid many subtleties, this chapter analyzes finite- and countable-state models in which *all* stationary policies generate Markov chains with a single irreducible class. In this case a single optimality equation suffices to characterize optimal policies. We follow this with an analysis of finite-state models with more general chain structure.

In this chapter we assume

Assumption 8.0.1. *Stationary rewards and transition probabilities*; $r(s, a)$ and $p(j|s, a)$ do not depend on the stage;

Assumption 8.0.2. *Bounded rewards*; $|r(s, a)| \leq M < \infty$ for all $a \in A_s$ and $s \in S$ (except in Sec. 8.10),

Assumption 8.0.3. *Finite-state spaces* (except in Secs. 8.10 and 8.11).

Often we restate these assumptions for emphasis.

In contrast to discounted models, approaches for analyzing MDPs with average reward criterion vary with the class structure of Markov chains generated by stationary policies; we review this aspect of Markov chain theory in Appendix A, Secs. A.1-A.4. Sections A.5–A.6 of Appendix A serve as the basis for Sec. 8.2. Section 5.4 also provides some relevant introductory material.

8.1 OPTIMALITY CRITERIA

Referring to our analyses of models with discounted and total-reward criteria suggests that we define an average reward function $g^{\pi}(s)$ for each $\pi \in \Pi^{\text{HR}}$, seek a method for computing

$$g^*(s) = \sup_{\pi \in \Pi^{\text{HR}}} g^{\pi}(s) \tag{8.1.1}$$

and finding a policy $\pi^* \in \Pi^{\text{HR}}$ for which

$$g^{\pi^*}(s) = g^*(s)$$

for all $s \in S$. Unfortunately this approach has some limitations as we show below.

8.1.1 The Average Reward of a Fixed Policy

Recall that for $\pi \in \Pi^{\text{HR}}$

$$v_{N+1}^{\pi}(s) = E_s^{\pi}\left\{ \sum_{t=1}^{N} r(X_t, Y_t) \right\} \tag{8.1.2}$$

denotes the total reward up to decision epoch $N + 1$ or, equivalently, the total reward in an $N + 1$ period problem with terminal reward zero. Define the *average expected reward* of policy π by

$$g^{\pi}(s) = \lim_{N \to \infty} \frac{1}{N} v_{N+1}^{\pi}(s) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} P_{\pi}^{n-1} r_{d_n}(s). \tag{8.1.3}$$

As the following simple example illustrates, even in a finite-state example, this limit need not exist for some policies. Because of this, we require refined concepts of average optimality.

Example 8.1.1. Let $S = \{s_1, s_2\}$, $A_{s_1} = \{a_{1,1}, a_{1,2}\}$, $A_{s_2} = \{a_{2,1}, a_{2,2}\}$, $r(s_1, a_{1,1}) = 2$, $r(s_1, a_{1,2}) = 2$, $r(s_2, a_{2,1}) = -2$, $r(s_2, a_{2,2}) = -2$, $p(s_1|s_1, a_{1,1}) = 1$, $p(s_2|s_1, a_{1,2}) = 1$, $p(s_1|s_2, a_{2,1}) = 1$ and $p(s_2|s_2, a_{2,2}) = 1$ (Fig. 8.1.1).

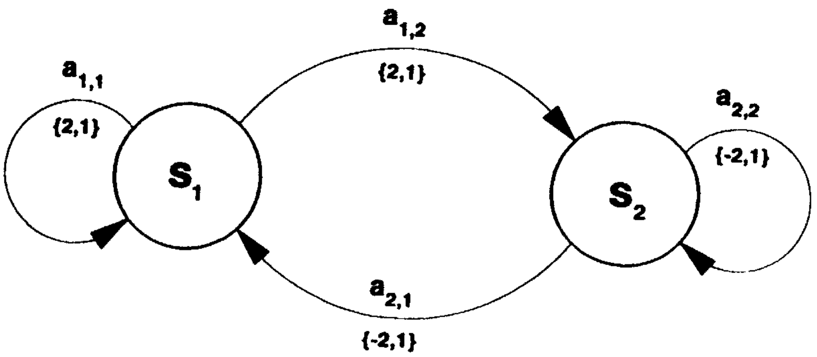

Figure 8.1.1 Symbolic representation of Example 8.1.1

Consider the history-dependent policy π which, on starting in s_1, remains in s_1 for one period, proceeds to s_2 and remains there for three periods, returns to s_1 and remains there for $3^2 = 9$ periods, proceeds to s_2 and remains there for $3^3 = 27$ periods, etc. Then direct computation shows that

$$\liminf_{N\to\infty} \frac{1}{N} v_{N+1}^{\pi}(s_1) = -1 \quad \text{and} \quad \limsup_{N\to\infty} \frac{1}{N} v_{N+1}^{\pi}(s_1) = 1,$$

so that the limit in (8.1.3) does not exist.

Note, however, for each of the four stationary policies in this model, the above lim sup and lim inf are identical and definition (8.1.3) is valid.

The example above suggests that the above approach is valid when we restrict attention to stationary policies. The result below confirms this. Note that we call a non-negative matrix *stochastic* if all of its row sums equal 1, i.e., P is stochastic if $Pe = e$.

Proposition 8.1.1.

a. Let S be countable. Let $d^\infty \in \Pi^{\text{SR}}$ and suppose that the limiting matrix of P_d, P_d^* is stochastic. Then the limit in (8.1.3) exists and

$$g^{d^\infty}(s) = \lim_{N\to\infty} \frac{1}{N} v_{N+1}^{d^\infty}(s) = P_d^* r_d(s). \tag{8.1.4}$$

b. If S is finite, (8.1.4) holds.

Proof. Since

$$v_{N+1}^{d^\infty}(s) = \sum_{n=1}^{N} P_d^{n-1} r_d(s)$$

part (a) follows from (A.3) in Appendix A.4. Part (b) follows by noting that, in a finite-state Markov model, P_d^* is stochastic. □

In Chapters 8 and 9, we will be concerned with finding policies which perform best in terms of limiting behavior of $N^{-1}v_{N+1}^{\pi}$. Since the limit in (8.1.3) need not exist, we introduce two related quantities. For $\pi \in \Pi^{\text{HR}}$, let $g_+^{\pi}(s)$ denote the *lim sup average reward* defined by

$$g_+^{\pi}(s) = \limsup_{N \to \infty} \frac{1}{N} v_{N+1}^{\pi}(s) \tag{8.1.5}$$

and let $g_-^{\pi}(s)$ denote the *liminf average reward* is defined by

$$g_-^{\pi}(s) = \liminf_{N \to \infty} \frac{1}{N} v_{N+1}^{\pi}(s). \tag{8.1.6}$$

Note that $g_+^{\pi}(s) \geq g_-^{\pi}(s)$ and that $g^{\pi}(s)$ exists if and only if $g_+^{\pi}(s) = g_-^{\pi}(s)$. We often refer to g^{π} as the *gain* of policy π. This expression has its origins in control engineering, in which it refers to the ratio of the output of a system to its input. We provide further justification for this designation in Sec. 8.2.

8.1.2 Average Optimality Criteria

We express three optimality criteria in terms of g_+^{π} and g_-^{π}. Note that they are *equivalent in finite-state models*. We say that a policy π^* is *average optimal* if

$$g_-^{\pi^*}(s) \geq g_+^{\pi}(s) \tag{8.1.7}$$

for all $s \in S$ and $\pi \in \Pi^{\text{HR}}$; is *lim sup average optimal* if

$$g_+^{\pi^*}(s) \geq g_+^{\pi}(s) \tag{8.1.8}$$

for all $s \in S$ and $\pi \in \Pi^{\text{HR}}$; and is *lim inf average optimal* if

$$g_-^{\pi^*}(s) \geq g_-^{\pi}(s) \tag{8.1.9}$$

for all $s \in S$ and $\pi \in \Pi^{\text{HR}}$.

We define the following additional quantities:

$$g_+^*(s) = \sup_{\pi \in \Pi^{\text{HR}}} g_+^{\pi}(s) \tag{8.1.10}$$

$$g_-^*(s) = \sup_{\pi \in \Pi^{\text{HR}}} g_-^{\pi}(s). \tag{8.1.11}$$

Lim inf average optimality corresponds to comparing policies in terms of worst case limiting performance, while lim sup average optimality corresponds to comparison in terms of best case performance. Average optimality is the strongest of these three criteria because it requires that

$$\liminf_{N \to \infty} \frac{1}{N} v_{N+1}^{\pi^*}(s) \geq \limsup_{N \to \infty} \frac{1}{N} v_{N+1}^{\pi}(s)$$

for all $\pi \in \Pi^{\mathrm{HR}}$. This means that the lowest possible limiting average reward under policy π^* is as least as great as the best possible limiting average reward under any other policy. Clearly, if π^* is average optimal, it is lim sup and lim inf average optimal. Note that the lim sup and lim inf average optimality criteria are distinct; neither implies the other. As a consequence of Proposition 8.1.1, a *stationary* lim sup average optimal policy with stochastic limiting matrix is average optimal and hence lim inf average optimal.

Example 8.1.1 (ctd.). Observe that in this model the stationary policy $d^*(s_1) = a_{1,1}$, $d^*(s_2) = a_{2,2}$ is *average* optimal, with $g^*(s_1) = g^*(s_2) = 2$.

The following example illustrates the points in the preceding paragraph regarding the relationship of these optimality criteria.

Example 8.1.2. Let $S = \{-1, 0, 1, 2, \dots\}$; $A_0 = \{a_{0,1}, a_{0,2}\}$ and $A_s = \{a_{s,1}\}$ for $s \neq 0$; and $r(0, a_{0,1}) = -1$, $r(0, a_{0,2}) = 0$, and $r(-1, a_{-1,1}) = 0$,

$$r(s, a_{s,1}) = 1, \qquad \sum_{k=0}^{n} 3^k \leq s \leq \sum_{k=0}^{n+1} 3^k - 1, \qquad \text{and } n = 0, 2, 4, \dots$$

and

$$r(s, a_{s,1}) = -1, \qquad \sum_{k=0}^{n} 3^k \leq s \leq \sum_{k=0}^{n+1} 3^k - 1, \qquad \text{and } n = 1, 3, 5, \dots.$$

$p(1|0, a_{0,1}) = 1$, $p(-1|0, a_{0,2}) = 1$, $p(-1|-1, a_{-1,1}) = 1$, and $p(s+1|s, a_{s,1}) = 1$ for $s \geq 1$ (Fig. 8.1.2).

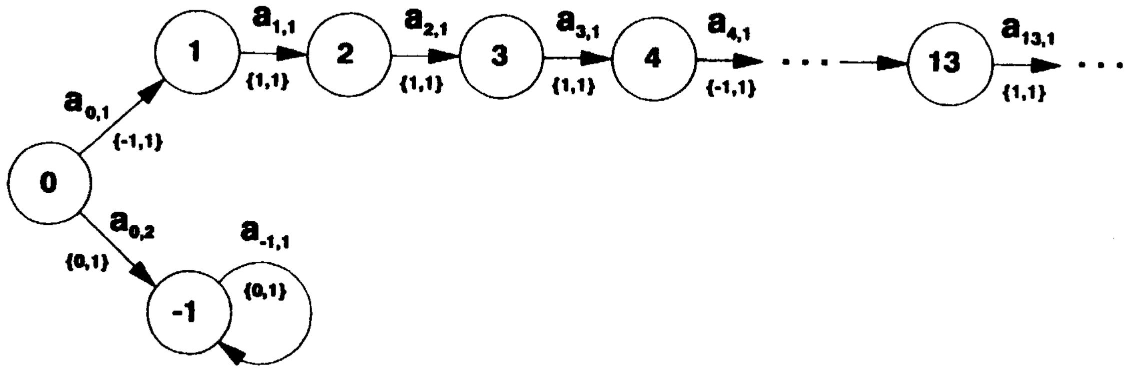

Figure 8.1.2 Symbolic representation of Example 8.1.2.

Action choice in state 0 determines the evolution of the system. Let $\delta(0) = a_{0,1}$ and $\gamma(0) = a_{0,2}$. Then

$$g_+^{\delta^\infty}(0) = \tfrac{1}{2}, \qquad g_-^{\delta^\infty}(0) = -\tfrac{1}{2},$$

and

$$g_+^{\gamma^\infty}(0) = g_-^{\gamma^\infty}(0) = g^{\gamma^\infty}(0) = 0.$$

Observe that

1. δ^∞ is lim sup average optimal,
2. γ^∞ is lim inf average optimal, and
3. no average optimal policy exists.

Note further that (8.1.4) does not hold for δ^∞ because P_δ^* is a matrix with all components equal to 0 and is not stochastic.

In this chapter, we will be concerned with establishing that average optimal policies exist and characterizing their limiting average reward. The following important result, which restates Theorem 5.5.3(b), allows us to restrict attention to Markov policies in the above definitions.

Theorem 8.1.2. For each $\pi \in \Pi^{\text{HR}}$ and $s \in S$, there exists a $\pi' \in \Pi^{\text{MR}}$ for which

a. $g_+^{\pi'}(s) = g_+^{\pi}(s)$,
b. $g_-^{\pi'}(s) = g_-^{\pi}(s)$, and
c. $g^{\pi'}(s) = g^{\pi}(s)$ whenever $g_-^{\pi}(s) = g_+^{\pi}(s)$.

As a consequence of this result, we need only search over the set of Markov randomized policies when determining whether a particular policy is average, lim sup average, or lim inf average optimal. Consequently,

$$g^*(s) = \sup_{\pi \in \Pi^{\text{MR}}} g^{\pi}(s) \tag{8.1.12}$$

with analogous results for g_+^* and g_-^*.

8.2 MARKOV REWARD PROCESSES AND EVALUATION EQUATIONS

Let S denote a finite set. Let P denote the transition probability matrix of a Markov chain $\{X_t\colon t = 1, 2, \dots\}$ and $r(s)$ a reward function. We refer to the bivariate stochastic process $\{(X_t, r(X_t))\colon t = 1, 2, \dots\}$ as a *Markov reward process* (MRP). In Markov decision processes, each stationary policy d^∞ generates a MRP with transition matrix P_d and reward r_d. As a preliminary to deriving result for MDPs, we characterize the average reward and related quantities for a MRP and provide systems of equations which characterize them.

Our analysis in this section includes models with both aperiodic and periodic Markov chains. When chains are periodic, most limiting quantities do not exist. To account for this, we use Cesaro limits (Appendix A.4) instead. Recall that for a sequence (y_n) the Cesaro limit (denoted C-lim) is given by

$$C\text{-}\lim_{n\to\infty} = \lim_{n\to\infty} \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

Note that the Cesaro limit equals the ordinary limit whenever it exists; however, it provides a reasonable limiting quantity whenever $\liminf_{n\to\infty} y_n$ and $\limsup_{n\to\infty} y_n$ are not equal. Including the Cesaro limit makes some of the formulas a bit more complex, but hopefully does not detract from understanding. When reading this chapter for the first time, you might wish to assume that all chains are aperiodic and all limits are ordinary limits. Feller (1950, p. 356) notes in his analysis of Markov chains that

> "The modifications required for periodic chains are rather trite, but the formulations required become unpleasantly involved."

This point applies equally well to analysis of MDPs with average reward criterion, especially with respect to convergence of value iteration.

8.2.1 The Gain and Bias

Results in Section A.4 of Appendix A show that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} P^{t-1} = P^*$$

exists when S is finite or countable. Therefore, whenever r is bounded and P^* is stochastic, the gain of the MRP satisfies

$$g(s) \equiv \lim_{N\to\infty} \frac{1}{N} E_s\left\{ \sum_{t=1}^{N} r(X_t) \right\} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} P^{t-1} r(s) = P^* r(s) \quad (8.2.1)$$

for each $s \in S$. Example 5.1.2 showed that interchanging the limit and the expectation in (8.2.1) may not be justified in models with countable S and unbounded r. In that example, $P^* = 0$, so that $P^* r(s) = 0$ for $s \in S$ but

$$\lim_{N\to\infty} \frac{1}{N} E_s\left\{ \sum_{t=1}^{N} r(X_t) \right\}$$

does not exist.

Suppose P^* is stochastic. Since it has identical rows for states in the same closed irreducible recurrent class (Section A.4 of Appendix A), we have the following important result.

Proposition 8.2.1. Suppose P^* is stochastic. Then if j and k are in the same closed irreducible class, $g(j) = g(k)$. Further, if the chain is irreducible, or has a single recurrent class and possibly some transient states, $g(s)$ is a constant function.

When $g(s)$ is constant, we write ge to denote an $|S|$-vector with identical components g.

Assume now that S is finite and define the *bias*, h of the MRP by

$$h \equiv H_P r \tag{8.2.2}$$

where the fundamental matrix $H_P \equiv (I - P + P^*)^{-1}(I - P^*)$ is defined in (A.14) in Sec. A.5 of Appendix A. From (A.4), $PP^* = P^*$, so (8.2.1) implies that $Pg = PP^*r = P^*r = g$. Therefore, in an *aperiodic* Markov chain, it follows that

$$h = \sum_{t=0}^{\infty} (P^t - P^*)r = \sum_{t=0}^{\infty} P^t(r - g). \tag{8.2.3}$$

We may express this in component notation as

$$h(s) = E_s\left\{\sum_{t=1}^{\infty} [r(X_t) - g(X_t)]\right\}. \tag{8.2.4}$$

From (8.2.1) and the definition of P^*, it follows that the gain represents the average reward per period for a system in steady state. Sometimes we refer to this quantity as the *stationary reward*. Therefore (8.2.4) allows interpretation of the bias as the expected total difference between the reward and the stationary reward. Alternatively, the first expression in (8.2.3) shows that the bias represents the difference between the total reward for a system that starts in state s and one in which the sth row of P^* determines the initial state. Since an aperiodic Markov chain approaches its steady-state exponentially fast, most of the difference in (8.2.4) will be earned during the first few transitions, so that we may regard the bias as a "transient" reward.

In *periodic* chains, (8.2.3) and (8.2.4) hold in the Cesaro limit sense, so that (8.2.4) may be interpreted as

$$h(s) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} E_s\left\{\sum_{t=1}^{k} [r(X_t) - g(X_t)]\right\}.$$

An alternative representation for h provides a different interpretation for the bias for *aperiodic* MRPs. Recall that v_{N+1} denotes the total expected reward over $N + 1$ periods in a system in which the terminal reward equals 0. That is,

$$v_{N+1} = \sum_{t=1}^{N} P^{t-1}r.$$

From (8.2.3),

$$h = \sum_{t=1}^{N} P^{t-1}r - Ng + \sum_{t=N+1}^{\infty} (P^{t-1} - P^*)r.$$

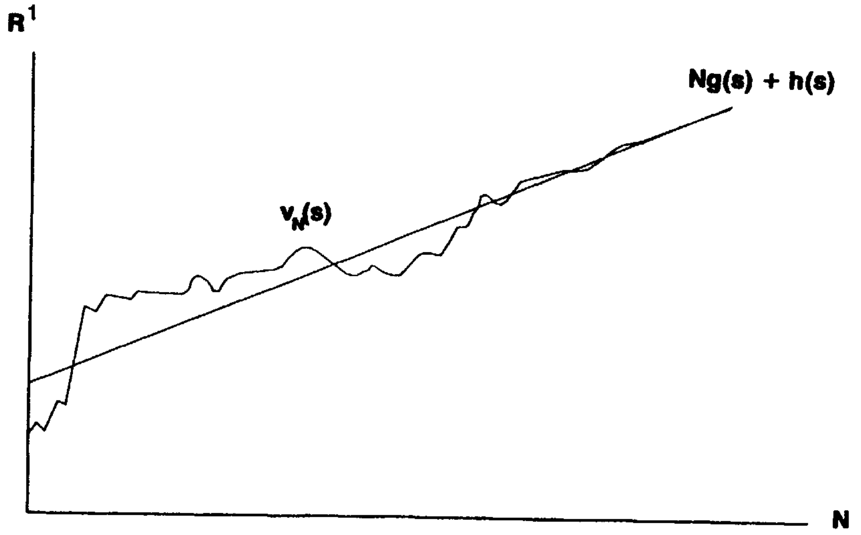

Figure 8.2.1 Graphical representation of (8.2.5). Vertical axis denotes the real line.

Theorem A.7(c) implies that the third term above converges to zero as $N \to \infty$, so that we may write

$$v_{N+1} = Ng + h + o(1) \tag{8.2.5}$$

where $o(1)$ denotes a vector with components which approach 0 pointwise as $N \to \infty$. Thus, as N becomes large, for each $s \in S$, $v_{N+1}(s)$ approaches a line with slope $g(s)$ and intercept $h(s)$ (Fig. 8.2.1).

For j and k in the same closed irreducible class, Proposition 8.1.1 implies that $g(j) = g(k)$, so that (8.2.5) implies that

$$h(j) - h(k) = \lim_{N \to \infty} \left[v_N(j) - v_N(k) \right].$$

Hence h equals the asymptotic relative difference in total reward that results from starting the process in state j instead of in state k. For this reason, we sometimes refer to h as the *relative value* vector.

For *periodic* chains, a similar line of reasoning shows that, for j and k in the same closed irreducible class,

$$h(j) - h(k) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[v_n(j) - v_n(k) \right]$$

so that $h(j) - h(k)$ gives the average relative difference in total reward. Another consequence of representation (8.2.5) is that $v_N - v_{N-1}$ eventually increases at rate g, providing further justification for referring to g as the gain of the process.

The result below follows immediately from (8.2.4).

Proposition 8.2.2. Suppose $g(s) = 0$ for all $s \in S$.

a. Then

$$h(s) = C\text{-}\lim_{N\to\infty} E_s\left\{\sum_{t=1}^{N} r(X_t)\right\} = \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} v_k(s)$$

and

b.

$$h(s) = \lim_{N\to\infty} E_s\left\{\sum_{t=1}^{N} r(X_t)\right\} =_{,} \lim_{N\to\infty} v_{N+1}(s) \tag{8.2.6}$$

whenever the limit exists.

We will use this result in Section 10.4 to further investigate behavior of policy iteration in models with total reward criterion. We showed in Chap. 7 that, in finite-state models, whenever $\lim_{N\to\infty} v_N^{d^\infty}(s)$ exists, $r(s) = 0$ at all states which are recurrent under P_d. Therefore in such models, *the bias equals the expected total reward*. Part (a) of this proposition refers to models which may cycle on recurrent states and accumulate positive and negative rewards which average out to 0. In that case the expected total-reward criterion may be inappropriate for comparing policies.

We demonstrate some of the above concepts by expanding on Example A.1 of Appendix A.

Example 8.2.1. Let $S = \{s_1, s_2\}$, and suppose a Markov chain has transition probability matrix

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

In Appendix A, we show that

$$P^* = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \qquad H_P = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix}.$$

Suppose $r' = (1, -1)$. Then

$$g = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad h = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}.$$

We now interpret these quantities. Starting in s_1, the stream of rewards for this system is $(1, -1, 1, -1, \dots)$ so that the sequence of partial sums of rewards, $\{v_n(s_1)\}$ equals $(1, 0, 1, 0, \dots)$. Therefore the sequence of average rewards starting from s_1

equals $(1, 0, \frac{1}{3}, 0, \frac{1}{5}, 0, \dots)$. Hence the long-run average equals 0 as given by $g(s_1)$. The partial sum sequence $\{v_n(s_1)\}$ converges in the Cesaro sense, but not in the ordinary sense. The sequence of *average* partial sums equals $(1, \frac{1}{2}, \frac{2}{3}, \frac{1}{2}, \frac{3}{5}, \dots)$, so that its Cesaro limit equals $\frac{1}{2}$ as given by $h(s_1)$. Since the sequence of $v_n(s_2)$ equals $(-1, 0, -1, 0, \dots)$, the sequence $v_n(s_1) - v_n(s_2)$ equals $(2, 0, 2, 0, \dots)$, so that the average difference in total rewards between these two states equals 1 as given by $h(s_1) - h(s_2)$.

8.2.2 The Laurent Series Expansion

Laurent series expansions relate the gain and bias to the expected total discounted reward and provide a key tool for analyzing finite-state undiscounted models. Following Sec. A.6 of Appendix A, we parameterize the discounted reward in terms of the interest rate ρ instead of in terms of the discount rate λ. These quantities are related by $\lambda = (1 + \rho)^{-1}$ or $\rho = (1 - \lambda)\lambda^{-1}$; $0 \le \lambda < 1$ implies $\rho > 0$. The quantity $1 + \rho$ represents the amount received at the start of the next period, when one unit is invested at the start of the present period and the interest rate equals ρ. Letting v_λ represent the total expected discounted reward of the MRP, it follows from (6.1.10) that

$$v_\lambda = (I - \lambda P)^{-1} r = (1 + \rho)(\rho I + [P - I])^{-1} r. \tag{8.2.7}$$

We refer to $(\rho I + [I - P])^{-1}$ as the *resolvent* of $I - P$ (at ρ). Multiplying the Laurent series expansion for the resolvent in Theorem A.8 by r, we obtain the following Laurent series expansion of v_λ.

Theorem 8.2.3. Assume finite S. Let ν denote the nonzero eigenvalue of $I - P$ with the smallest modulus. Then, for $0 < \rho < |\nu|$,

$$v_\lambda = (1 + \rho)\left[\rho^{-1} y_{-1} + \sum_{n=0}^{\infty} \rho^n y_n\right] \tag{8.2.8}$$

where $y_{-1} = P^* r = g$, $y_0 = H_P r = h$, and $y_n = (-1)^n H_P^{n+1} r$ for $n = 1, 2, \dots$.

Often, instead of writing $0 < \rho < |\nu|$, we say "for ρ sufficiently small." When analyzing MDPs with average reward criterion, it will suffice to use the following *truncated Laurent expansion* of v_λ.

Corollary 8.2.4. Let g and h represent the gain and bias of a MRP with finite S. Then if Assumption 8.0.2 holds

$$v_\lambda = (1 - \lambda)^{-1} g + h + f(\lambda), \tag{8.2.9}$$

where $f(\lambda)$ denotes a vector which converges to zero as $\lambda \uparrow 1$.

Proof. Expressing (8.2.8) on the λ scale and adding and subtracting h, we obtain

$$v_\lambda = \frac{1}{1-\lambda}g + h + \frac{1-\lambda}{\lambda}h + \frac{1}{\lambda}\sum_{n=0}^{\infty}(-1)^n\left[\frac{1-\lambda}{\lambda}\right]^n y_n.$$

Since the series in (8.2.8) converges for ρ sufficiently small, the last two terms above converge to 0 as $\lambda \uparrow 1$. □

The following corollary provides a powerful tool for extending structure and existence results from the discounted case to the average reward case. It follows immediately by multiplying both sides of (8.2.9) by $1 - \lambda$ and passing to the limit.

Corollary 8.2.5. Let g and v_λ represent the gain and expected discounted reward of a MRP. Then

$$g = \lim_{\lambda \uparrow 1}(1-\lambda)v_\lambda. \tag{8.2.10}$$

The result in Corollary 8.2.5 has the following interesting probabilistic interpretation. Consider a *terminating* Markov reward process in which a random variable τ, with distribution independent of that of the Markov chain, determines the termination time and a reward $r(X_\tau)$ is received only at termination. Suppose τ has a geometric distribution parametrized as

$$P(\tau = n) = (1-\lambda)\lambda^{n-1}, \qquad n = 1, 2, \ldots.$$

Then

$$E_s\{r(X_\tau)\} = \sum_{n=1}^{\infty}(1-\lambda)\lambda^{n-1}E_s\{r(X_n)\} = (1-\lambda)v_\lambda(s).$$

Since the geometric distribution corresponds to the time to first "failure" in a sequence of independent Bernoulli trials with failure probability $1 - \lambda$, as $\lambda \uparrow 1$ or $(1-\lambda) \downarrow 0$, $P(\tau > M)$ approaches 1, for any M, so that

$$\lim_{\lambda \uparrow 1}(1-\lambda)v_\lambda(s) = E_s\{r(X_\infty)\} = g(s).$$

This argument can be made formal to provide a probabilistic derivation of Corollary 8.2.5.

Example 8.2.1 (ctd.). The eigenvalues of $I - P$ are 0 and 2, so that, for $0 < \rho < 2$, the Laurent series expansion for v_λ satisfies

$$v_\lambda = (1+\rho)\left(\frac{1}{\rho}\begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}\frac{1}{2}\\\frac{1}{2}\end{bmatrix} + \rho\begin{bmatrix}-\frac{1}{4}\\\frac{1}{4}\end{bmatrix} + \rho^2\begin{bmatrix}\frac{1}{8}\\-\frac{1}{8}\end{bmatrix} + \cdots\right).$$

By direct calculation through inverting $(I - \lambda P)$ or summing the above series, we observe that

$$v_\lambda = \begin{bmatrix} \dfrac{1+\rho}{2+\rho} \\ -\dfrac{1+\rho}{2+\rho} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{1+\lambda} \\ -\dfrac{1}{1+\lambda} \end{bmatrix}.$$

Note as $\rho \downarrow 0$ or $\lambda \uparrow 1$, that $(1-\lambda)v_\lambda$ converges to $g = 0$, as indicated by Corollary 8.2.5 and that v_λ converges to h, as shown in Corollary 8.2.4.

8.2.3 Evaluation Equations

In all but the simplest examples, computation of g and h through direct evaluation of P^* and H_P may be inefficient. In this section we provide systems of equations which enable computation of these quantities as well as the higher order terms in the Laurent series expansion of v_λ. These equations serve as the basis for the optimality equations for models with average and sensitive optimality criteria.

We provide two approaches to deriving these equations; the first uses identities (A.4), (A.17) and (A.20) of Appendix A. The second approach uses the discounted reward evaluation equation and the Laurent series expansion for the discounted reward.

Theorem 8.2.6. Let S be finite and let g and h denote the gain and bias of a MRP with transition matrix P and reward r.

a. Then

$$(I-P)g = 0 \tag{8.2.11}$$

and

$$g + (I-P)h = r \tag{8.2.12}$$

b. Suppose g and h satisfy (8.2.11) and (8.2.12), then $g = P^*r$ and $h = H_P r + u$ where $(I-P)u = 0$.

c. Suppose g and h satisfy (8.2.11), (8.2.12) and $P^*h = 0$, then $h = H_P r$.

Proof. From (A.4) $(I-P)P^* = 0$, so multiplying r by $(I-P)P^*$ and noting that $g = P^*r$ establishes (8.2.11). From (A.17), $P^* + (I-P)H_P = I$ so applying both sides of this identity to r establishes (8.2.12).

To establish (b), apply P^* to (8.2.12) and add it to (8.2.11) to obtain

$$(I - P + P^*)g = P^*r.$$

Noting the non-singularity of $(I - P + P^*)$, (A.20) shows that

$$g = (I - P + P^*)^{-1}P^*r = Z_P P^*r = P^*r.$$

We have previously shown that $h = H_P r$ satisfies (8.2.12). Suppose h_1 also satisfies (8.2.12). Then $(I - P)(H_P r - h_1) = 0$, so h is unique up to an element of $\{u \in R^{|S|}: (I - P)u = 0\}$.

We now prove part (c). Suppose $P^*h = 0$, then adding this identity to (8.2.12) shows that

$$(I - P + P^*)h = r - g.$$

Therefore,

$$h = (I - P + P^*)^{-1}(I - P^*)r = H_P r. \qquad \square$$

The equations (8.2.11) and (8.2.12) uniquely characterize g and determine h up to an element of the null space of $I - P$. The condition $P^*h = 0$ doesn't provide much practical assistance in finding h because it requires prior determination of P^* (Appendix A.4). We discuss a more efficient approach for computing h below.

First, we show that we need not solve $(I - P)g = 0$ directly. If g satisfies $(I - P)g = 0$, then g is an element of the subspace of $R^{|S|}$ spanned by the eigenvectors of P corresponding to eigenvalue 1. Suppose there are m closed irreducible recurrent classes denoted by $C_1, C_2, \ldots, C_m$, then a basis for this subspace consists of vectors $u_1, u_2, \ldots, u_m$ for which $u_k(s) = 1$ if $s \in C_k$ and 0 otherwise. Thus, for s transient,

$$g(s) - \Sigma_{k \in T} p(k|s)g(k) = \Sigma_{k \in C_1 \cup \cdots \cup C_m} p(k|s)g(k). \qquad (8.2.13)$$

As a consequence of Proposition A.3, once we have found g on recurrent states, (8.2.13) uniquely determines g on transient states.

Therefore, we can find g using the following approach:

1. Classify the states of the Markov chain using the Fox-Landi algorithm from Section A.3,
2. Compute g on each recurrent class of P by solving (8.2.12) on each class separately with g constrained to be constant there, and
3. Compute g on transient states by solving (8.2.13).

When P is irreducible or unichain, g is constant and (8.2.11) becomes superfluous. In this case any solution of (8.2.11) is a scalar multiple of the unit vector so that we need not solve (8.2.11) and can find g uniquely by solving (8.2.12) alone. We state this as follows.

Corollary 8.2.7. Suppose P is unichain or irreducible. Then the average reward $P^*r = ge$ and it is uniquely determined by solving

$$ge + (I - P)h = r. \qquad (8.2.14)$$

Suppose g and h satisfy (8.2.14), then $g = P^*r$ and $h = H_P r + ke$ for arbitrary scalar k. Furthermore, if g and h satisfy (8.2.14) and $P^*h = 0$, then $h = H_P r$.

Another consequence of Theorem 8.2.6 is that (8.2.11) and (8.2.12) uniquely determine h up to an element of the null space of $I - P$. Since this space has dimension m, where m denotes the number of closed irreducible classes of P, this system of equations determines h up to m constants (one on each closed class and possibly all m on transient states). Therefore, any specification, such as $P^*h = 0$, which determines these m constants provides a unique representation for h, but it does not assure that $h = H_P r$. Consequently, for unichain P, we can find *relative values* $h(j) - h(k)$ by setting any component of h equal to zero and solving (8.2.14).

We illustrate the above concepts with the following multichain example.

Example 8.2.2. Suppose $S = \{s_1, \dots, s_5\}$,

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.4 & 0.6 & 0 & 0 \\ 0 & 0.7 & 0.3 & 0 & 0 \\ 0.5 & 0.2 & 0 & 0 & 0.3 \\ 0.2 & 0.3 & 0.3 & 0.2 & 0 \end{bmatrix} \quad \text{and} \quad r = \begin{bmatrix} 2 \\ 5 \\ 4 \\ 1 \\ 3 \end{bmatrix}.$$

Observe that P is in canonical form and that P has two closed irreducible recurrent classes, $C_1 = \{s_1\}, C_2 = \{s_2, s_3\}$ and transient states $T = \{s_4, s_5\}$. Solving $(I - P)g = 0$, does not determine $g(s_1)$, shows that $g(s_2) = g(s_3)$ and that

$$g(s_4) - 0.3g(s_5) = 0.5g(s_1) + 0.2g(s_2)$$
$$-0.2g(s_4) + g(s_5) = 0.2g(s_1) + 0.3g(s_2) + 0.2g(s_3).$$

We may express (8.2.12) in component notation as

$$r(s) - g(s) + \sum_{j \in S} p(j|s)h(j) - h(s) = 0. \tag{8.2.15}$$

Substituting P and r into this equation and noting the above relationships for g shows that $g(s_1) = r(s_1) = 2$,

$$\begin{aligned} 5 - g(s_2) - 0.6h(s_2) + 0.6h(s_3) &= 0, \\ 4 - g(s_2) + 0.7h(s_2) - 0.7h(s_3) &= 0, \end{aligned} \tag{8.2.16}$$

so that $g(s_2) = g(s_3) = 4.538$, and

$$g(s_4) - 0.3g(s_5) = 0.5 \times 2 + 0.2 \times 4.538 = 1.908,$$
$$-0.2g(s_4) + g(s_5) = 0.2 \times 2 + 0.5 \times 4.538 = 2.669,$$

so that $g(s_4) = 2.882$ and $g(s_5) = 3.245$.

To determine h, (8.2.16) implies that

$$\begin{aligned} 0.462 - 0.6h(s_2) + 0.6h(s_3) &= 0, \\ -0.538 + 0.7h(s_2) - 0.7h(s_3) &= 0, \end{aligned} \tag{8.2.17}$$

so that one of these equations is redundant. On the transient states, (8.2.15) yields

$$\begin{aligned} -1.882 + 0.5h(s_1) + 0.2h(s_2) - h(s_4) + 0.3h(s_5) &= 0 \\ -0.245 + 0.2h(s_1) + 0.3h(s_2) + 0.3h(s_3) + 0.2h(s_4) - h(s_5) &= 0 \end{aligned} \tag{8.2.18}$$

As a consequence of Theorem 8.2.6, $P^*h = 0$ uniquely determines h. Using methods in Sec. A.4 shows that

$$P^* = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.538 & 0.462 & 0 & 0 \\ 0 & 0.538 & 0.462 & 0 & 0 \\ 0.595 & 0.218 & 0.187 & 0 & 0 \\ 0.319 & 0.367 & 0.314 & 0 & 0 \end{bmatrix}$$

so that $h(s_1) = 0$, and

$$0.462 - 0.600h(s_2) + 0.600h(s_3) = 0,$$

$$0.538h(s_2) + 0.462h(s_3) = 0$$

Hence $h(s_2) = 0.354$ and $h(s_3) = -0.416$ and from (8.2.18) we obtain $h(s_4) = 2.011$ and $h(s_5) = 0.666$. Note also that $g = P^*r$.

We now provide a system of equations which uniquely characterize g, h, and the higher-order terms of the Laurent series expansions of v_λ. Note that (8.2.11) characterizes g as an element of the null space of $(I - P)$. By adding the additional equation (8.2.12), we uniquely determine g. This suggests that by adding another equation to the above system we can determine h uniquely. We formalize this observation.

Theorem 8.2.8. In a Markov reward process with transition matrix P and reward r, let y_n, $n = -1, 0, \ldots$ denote the coefficients of the Laurent series expansion of v_λ.

a. Then

$$(I - P)y_{-1} = 0, \tag{8.2.19}$$

$$y_{-1} + (I - P)y_0 = r, \tag{8.2.20}$$

and for $n = 1, 2, \ldots$

$$y_{n-1} + (I - P)y_n = 0. \tag{8.2.21}$$

b. Suppose for some $M \geq 0$, that $w_{-1}, w_0, \ldots, w_M$ satisfy (8.2.19), (8.2.20), and if $M \geq 1$, (8.2.21) for $n = 1, 2, \ldots, M$. Then $w_{-1} = y_{-1}$, $w_0 = y_0$, $\ldots$, $w_{M-1} = y_{M-1}$ and $w_M = y_M + u$ where $(I - P)u = 0$.

c. Suppose the hypotheses of part (b) hold and in addition $P^*w_M = 0$. Then $w_M = y_M$.

Proof. Since $(I - \lambda P)v_\lambda = r$ and $\lambda = (1 + \rho)^{-1}$ it follows from (8.2.8) that

$$(1 + \rho)\big(I - (1 + \rho)^{-1}P\big)\left[\rho^{-1}y_{-1} + \sum_{n=0}^{\infty}(-\rho)^n y_n\right] = r$$

Rearranging terms shows that

$$\rho^{-1}(I - P)y_{-1} + \sum_{n=0}^{\infty}(-\rho)^n[y_{n-1} + (I - P)y_n] = r$$

for $0 < \rho < |\nu|$ where ν was defined in Theorem 8.2.3. Equating terms with like powers of ρ shows that $y_{-1}, y_0, \ldots$ satisfy (8.2.19)–(8.2.21).

The proofs of parts (b) and (c) follow by inductively applying the argument in part (b) of Theorem 8.2.6. and noting from (A.18) and (A.19) that $Z_P(H_P)^n = (H_P)^{n+1}$. □

When we choose $M = 1$ in the above theorem we obtain the following system of equations for computing the bias h.

Corollary 8.2.9. Suppose u, v, and w satisfy

$$(I - P)u = 0, \quad u + (I - P)v = r \quad \text{and} \quad v + (I - P)w = 0 \qquad (8.2.22)$$

then $u = g$, $v = h$, and $w = y_1 + z$ where $(I - P)z = 0$.

Example 8.2.2 (ctd.). We now compute the bias h using the approach of Corollary 8.2.9. Note that once we determine $h(s_1)$, $h(s_2)$, and $h(s_3)$, we may find $h(s_4)$ and $h(s_5)$ using (8.2.18). We express $h + (I - P)w = 0$ in component notation as

$$h(s) + w(s) - \sum_{j \in S} p(j|s)w(j) = 0 \qquad (8.2.23)$$

so that on substitution of the values for $p(j|s)$, we obtain

$$\begin{aligned} h(s_1) &= 0 \\ h(s_2) + 0.6w(s_2) - 0.6w(s_3) &= 0 \\ h(s_3) - 0.7w(s_2) + 0.7w(s_3) &= 0 \end{aligned}$$

Hence $h(s_2)$ and $h(s_3)$ satisfy

$$0.7h(s_2) + 0.6h(s_3) = 0$$

Noting from (8.2.17) that

$$0.462 - 0.6h(s_2) + 0.6h(s_3) = 0$$

we obtain $h(s_2) = 0.355$ and $h(s_3) = -0.414$ as above. We leave it as an exercise to show that (8.2.21) determines the higher-order terms of the Laurent series expansion and that they agree with those computed directly.

8.3 CLASSIFICATION OF MARKOV DECISION PROCESSES

With the exception of a few results in Chap. 7, we have ignored the chain structure (Appendix A) of the transition matrices of Markov chains generated by stationary policies. In average reward models we can no longer take this liberty, since results and analyses depend on state accessibility patterns of chains corresponding to stationary policies. In this section we describe the different classes of models, discuss the relationship between them, describe a simple algorithm for classifying a model, and show how the model class effects the form of the optimal average reward.

8.3.1 Classification Schemes

We classify MDPs in two ways:

1. On the basis of the chain structure of the set of Markov chains induced by all stationary policies.
2. On the basis of patterns of states which are accessible from each other under *some* stationary policy.

We refer to any MDP as *general* and distinguish the following classes of models. We say that a MDP is

a. *Recurrent* or *ergodic* if the transition matrix corresponding to *every* deterministic stationary policy consists of a single recurrent class;

b. *Unichain* if the transition matrix corresponding to *every* deterministic stationary policy is unichain, that is, it consists of a single recurrent class plus a possibly empty set of transient states;

c. *Communicating* if, for every pair of states s and j in S, there exists a deterministic stationary policy d^∞ under which j is accessible from s, that is, $p_d^n(j|s) > 0$ for some $n \geq 1$;

d. *Weakly communicating* if there exists a *closed* set of states, with each state in that set accessible from every other state in that set under some deterministic stationary policy, plus a possibly empty set of states which is transient under every policy; and

e. *Multichain* if the transition matrix corresponding to *at least one* stationary policy contains two or more closed irreducible recurrent classes.

Many authors use the expressions *multichain* and *general* interchangably. We distinguish them and use the more restrictive notion of a multichain model above. In our terminology, unichain and recurrent models are special cases of general models but are distinct from multichain models. Weakly communicating models may be viewed as communicating models with "extra" transient states. Figure 8.3.1 represents the relationship between these classifications. Classifications appear in order of generality; the most general classification appears at the top. Connecting lines indicate that the lower class is included in the class above it.

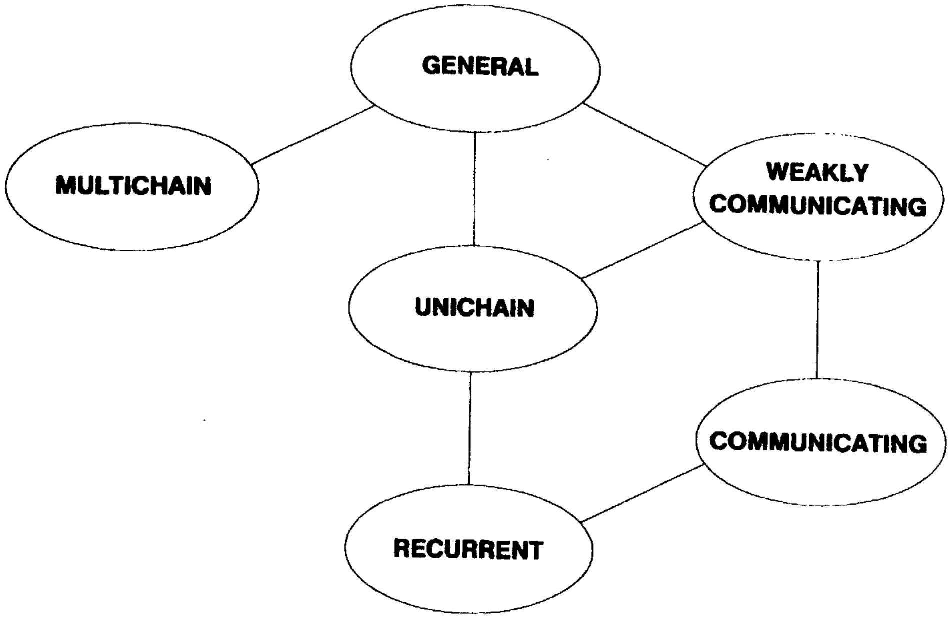

Figure 8.3.1 Hierarchy of MDP classes.

Unichain models generalize recurrent models while weakly communicating models generalize communicating models. Recurrent MDPs are communicating, and unichain MDPs are weakly communicating. Multichain MDPs may or may not be communicating, as the following inventory model illustrates. Refer to Sec. 3.2 for a formulation of the inventory model.

Example 8.3.1. Demands $\{D_t\}$ are independent and identically distributed with $P\{D_t = 0\} = p$ and $P\{D_t = 1\} = 1 - p$, $0 < p < 1$; the warehouse has a capacity of three units, and unfilled demand is lost. Following Sec. 3.2, $S = \{0, 1, 2, 3\}$ and $A_s = \{0, \ldots, 3 - s\}$. Under stationary policy d^∞ with $d(0) = 1$, $d(1) = 0$, $d(2) = 1$, and $d(3) = 0$, the Markov chain (Fig. 8.3.2) has two closed irreducible classes $\{0, 1\}$ and $\{2, 3\}$ so that the model is multichain. Alternatively, consider the stationary deterministic policy δ^∞ which orders three units when the stock level equals 0 and does not order otherwise; i.e., $\delta(0) = 3$, $\delta(1) = 0$, $\delta(2) = 0$, and $\delta(3) = 0$. Under this policy each state is accessible from each other state so that the model is communicat-

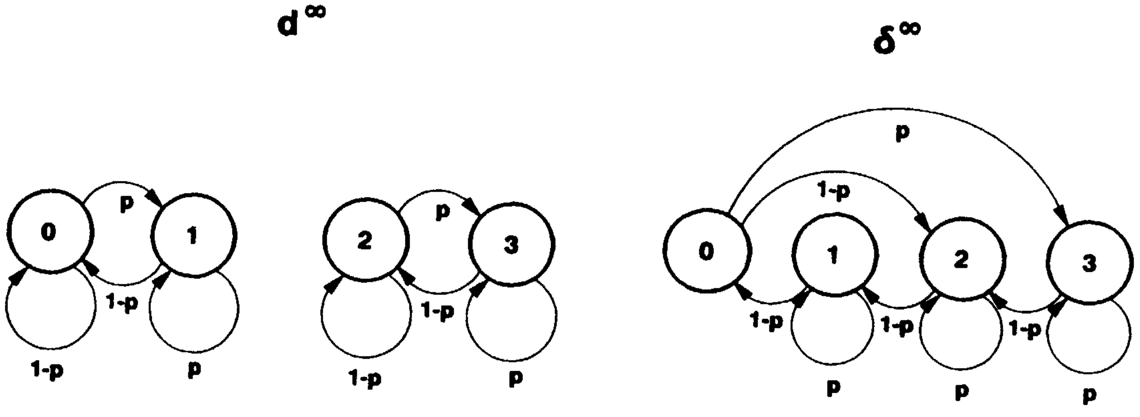

Figure 8.3.2 Markov chains corresponding to policies d^∞ and δ^∞ in Example 8.3.1.

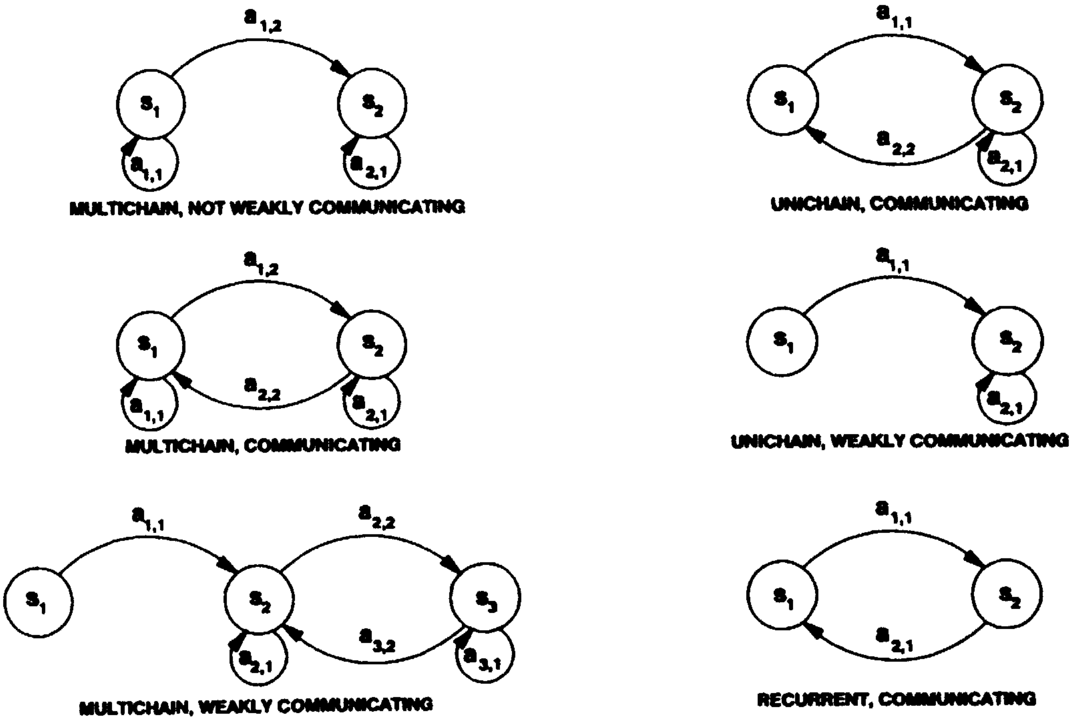

Figure 8.3.3 Examples of MDP classification. All transitions occur with probability 1.

ing (Fig. 8.3.2).

Figure 8.3.3 provides examples which illustrate each distinct model class.

8.3.2 Classifying a Markov Decision Process

In some applications, we might be able to determine whether all stationary policies are unichain or recurrent by inspection. If not, classification of a model as unichain or recurrent requires applying the Fox-Landi Algorithm of Appendix A, Sec. A.3 to the Markov chain corresponding to *each* stationary policy. Since there are $\Pi_{s\in S}|A_s| \le N^{|S|}$ stationary policies where $N = \max_{s\in S}|A_s|$, in the worse case, classification using this approach would be an exponential task requiring $O(N^{|S|}|S|^2)$ comparisons. In practical applications in which the model structure is not obvious, classification by this approach would be prohibitive. In this case, we solve these MDPs using the more complex algorithms of Chap. 9 which do not require prior determination of chain structure.

Classifying models on the basis of communication properties is considerably easier and we now describe an approach. It relies on the following observation.

Proposition 8.3.1

a. An MDP is communicating if and only if there exists a randomized stationary policy which induces a recurrent Markov chain.

b. An MDP is weakly communicating if and only if there exists a randomized stationary policy which induces a Markov chain with a single closed irreducible class and a set of states which is transient under all stationary policies.

Noting this, we can classify a model as communicating, weakly communicating or general in the following way.

Model Classification Algorithm

1. Define the matrix Q as follows: set $q(j|s) = 1$ if $p(j|s, a) > 0$ for some $a \in A_s$; otherwise set $q(j|s) = 0$.
2. Apply the Fox-Landi Algorithm of Section A.4 to Q.
3. If the matrix Q consists of a single closed class and no transient states, classify the MDP as communicating; if it consists of a single closed class plus transient states, go to 4; otherwise classify the MDP as general.
4. Set $c(s) = 1$ if s is in the closed class; set $c(s) = 0$ otherwise. Repeat the following until $c(s)$ no longer changes for any $s \in S$: For each $s \in S$ for which $c(s) = 0$, set $c(s) = 1$ if $\Sigma_{j \in S} p(j|s, a)c(j) > 0$ for *all* $a \in A_s$.

   If $c(s) = 1$ for all $s \in S$, classify the MDP as weakly communicating; otherwise classify it as general.

Matrix Q has the same pattern of positive entries and zeros as the transition probability matrix of a randomized stationary policy which in each state assigns positive probability to all actions. The algorithm finds the closed classes of that policy. Proposition 8.3.1 justifies the classification in step 3. Note that, after forming Q, this algorithm requires $O(|S|^2)$ comparisons to determine if the MDP is communicating.

8.3.3 Model Classification and the Average Reward Criterion

We now show how the structure of the optimal average reward (gain) relates to model type. Proposition 8.2.1 implies that the optimal gain is constant in recurrent and unichain models. We show that, in communicating and weakly communicating models, whenever a stationary policy has nonconstant gain, we may construct a stationary policy with constant gain which dominates the nonconstant gain policy. This need not be the case when the model is multichain but not weakly communicating.

We summarize observations in the following theorem. We provide a rather formal proof below. It can best be understood by referring to an example such as the continuation of Example 8.3.1 below or Problem 8.7. The basic idea is that, in a weakly communicating model, any closed class can be reached from every other state.

Theorem 8.3.2. Assume a weakly communicating model and let $d \in D^{\text{MD}}$.

a. Let C denote a closed, irreducible, and recurrent set of states in the Markov chain generated by the stationary policy d^∞. Then there exists a $\delta \in D^{\text{MD}}$ with $\delta(s) = d(s)$ for all $s \in C$, and for which the Markov chain generated by δ has C as its only closed, irreducible, recurrent class.

b. Suppose stationary policy d^∞ has $g^{d^\infty}(s) < g^{d^\infty}(s')$ for some s and s' in S. Then there exists a stationary policy δ^∞ for which $g^{\delta^\infty}(s) = g^{\delta^\infty}(s') \geq g^{d^\infty}(s')$.

c. Given any $d \in D^{\text{MD}}$, there exists a $\delta \in D^{\text{MD}}$ for which g^{δ^∞} is constant and $g^{\delta^\infty} \geq g^{d^\infty}$.

d. If there exists a stationary optimal policy, there exists a stationary optimal policy with constant gain.

Table 8.3.1 Relationship between model class and gain structure

Model Class	Optimal Gain	Gain of a Stationary Policy
Recurrent	Constant	Constant
Unichain	Constant	Constant
Communicating	Constant	Possibly nonconstant
Weakly Communicating	Constant	Possibly nonconstant
Multichain	Possibly nonconstant	Possibly nonconstant

Proof. Let T denote the set of states that is transient under every policy. From the definition of a weakly communicating model, there exists an $s \in S/(T \cup C)$ and an $a' \in A_s$ for which $\Sigma_{j \in C} p(j|s, a') > 0$. Set $\delta(s) = a'$. Augment C with s' and repeat this procedure until $\delta(s)$ is defined for all $s \in S/T$. By the definition of T, for each $s' \in T$, there exists an $a_{s'} \in A_{s'}$ for which $\Sigma_{j \in S/T} p(j|s', a_{s'}) > 0$. Set $\delta(s') = a_{s'}$ for each $s' \in T$. Then δ achieves the conclusions of part (a).

We now prove part (b). Let C be closed, irreducible, and recurrent under d^∞. Then, if $s' \in C$, it follows from (a) and Proposition 8.2.1 that there exists a $\delta \in D^{\mathrm{MD}}$ for which g^{δ^∞} is constant and $g^{\delta^\infty}(s') = g^{d^\infty}(s')$ so the result follows with equality. If s' is transient under d^∞, then there exists an s'' which is recurrent under d^∞ with $g^{d^\infty}(s'') \geq g^{d^\infty}(s')$. The result now follows from (a) and Proposition 8.2.1.

Note that (c) follows easily from (b) and that (d) follows immediately from (c). □

In subsequent sections of Chaps. 8 and 9, we establish the optimality of stationary policies in finite-state models with average reward criterion. Therefore, in light of Theorem 8.3.2, we may summarize the relationship between model classification and optimal gain as in Table 8.3.1.

Note that we include the adjective "possibly" to account for unusual models in which the optimal gain is constant even though the optimal policy is multichain (Example 8.4.2 below with $r(s_2, a_{2,1})$ set equal to 3). We now return to Example 8.3.1 and illustrate the conclusions of Theorem 8.3.2.

Example 8.3.1 (ctd.). We add assumptions regarding costs and revenues. Let $K = 2$, $c(u) = 2u$, $h(u) = u$, and $f(u) = 8u$. Similar calculations to those in Sect. 3.2 show that $r(s, a)$ satisfies

$r(s,a)$

a	0	1	2	3
s				
0	0	-1	-4	-7
1	4	-2	-5	×
2	4	-3	×	×
3	4	×	×	×

where × denotes a nonfeasible action. Choose $p = 0.5$, so that $P(D_t = 0) = P(D_t = 1) = 0.5$. Thus

$$P_d = \begin{bmatrix} 0.5 & 0.5 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.5 & 0.5 \end{bmatrix} = P_d^*$$

and $(r_d)^T = (-1, 4, -3, 4)$. Since $g^{d^\infty} = P_d^* r_d$, $(g^{d^\infty})^T = (1.5, 1.5, 0.5, 0.5)$.

As suggested by the proof of Theorem 8.3.2, we can alter this policy by choosing a decision rule which leads the process from states 2 and 3 to the closed class $\{0, 1\}$. Let γ denote the decision rule which orders one unit in state 0 and zero units otherwise. For that decision rule,

$$P_\gamma = \begin{bmatrix} 0.5 & 0.5 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \end{bmatrix}$$

so that

$$P_\gamma^* = \begin{bmatrix} 0.5 & 0.5 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.0 & 0.0 \end{bmatrix}.$$

Therefore $(g^{\gamma^\infty})^T = (1.5, 1.5, 1.5, 1.5)$ and $g^{\gamma^\infty} \geq g^{d^\infty}$. Observe that the Markov chain generated by γ^∞ is unichain. We leave it as an exercise to find the optimal policy. Our discussion above implies that it will have constant gain.

8.4 THE AVERAGE REWARD OPTIMALITY EQUATION—UNICHAIN MODELS

In this section we provide an optimality equation for unichain (and recurrent) Markov decision problems with average reward criterion. We analyze these models prior to multichain models, because for unichain models we may characterize optimal policies and their average rewards through a *single* optimality equation. The reason for this is that, in unichain models, *all* stationary policies have constant gain so that the MDP analog of (8.2.11),

$$\max_{d \in D} \{(P_d - I)g\} = 0,$$

holds for any constant g and provides no additional information regarding the structure of g. We will show that multichain models, be they communicating, weakly communicating, or noncommunicating, require the above equation in addition to the

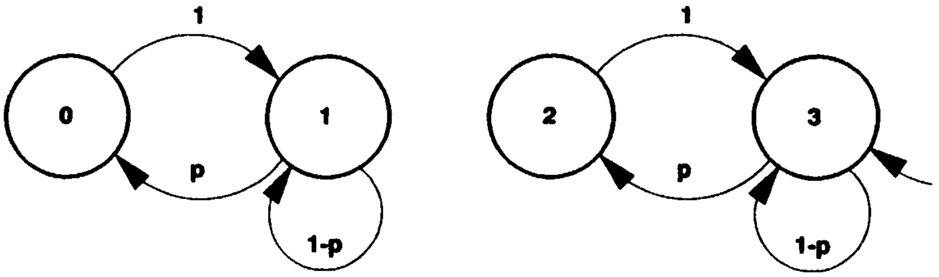

Figure 8.4.1 Transition diagram for a multichain queueing admission control policy.

equation

$$\max_{d \in D}\{r_d - g + (P_d - I)h\} = 0 \tag{8.4.1}$$

to determine optimal policies.

We provide an example of a model with unichain stationary policies.

Example 8.4.1. Consider the queueing admission control model of Sec. 3.7.1. Assume that the "possible service distribution" $f(s)$ has countable support (i.e., it assigns positive probability to all non-negative integer values). For example, $f(s)$ might be a Poisson or geometric probability mass function. Under this assumption, the state 0, which corresponds to an empty queue, may be reached in one step with positive probability from any system state. Hence all policies are unichain.

On the other hand, if $f(s)$ has finite support, the model may include multichain policies so that the analyses of this chapter need not apply. As an example, suppose $f(0) = 1 - p$, $f(1) = p$, $0 < p < 1$, and jobs arrive deterministically at the rate of one per period. Then any stationary policy which accepts a job in state 0, no jobs in states 1 or 3, and a job in state 2 contains at least two recurrent classes (Fig. 8.4.1).

The development in this section parallels that in Sec. 6.2 as closely as possible. We focus primarily on finite-state and action models; however, results in Sec. 8.4.1 hold in greater generality. Our proof of the existence of a solution to (8.4.1) relies on the finiteness of the set of stationary policies; we consider models with compact action sets in Sec. 8.4.4.

8.4.1 The Optimality Equation

When the gain has equal components, we denote its value by the scalar g or the vector ge. The optimality equation in a unichain average reward MDP may be expressed in component notation as

$$0 = \max_{a \in A_s}\left\{r(s,a) - g + \sum_{j \in S} p(j|s,a)h(j) - h(s)\right\} \tag{8.4.2}$$

or in matrix-vector and operator notation as

$$0 = \max_{d \in D}\{r_d - ge + (P_d - I)h\} \equiv B(g,h). \tag{8.4.3}$$

We regard B as a mapping from $R^1 \times V$ to V, where, as before, V denotes the space of bounded real-valued functions on S. Note that the maximum in (8.4.3) is with respect to the componentwise partial order and $D \equiv D^{\mathrm{MD}}$. When D consists of a single decision rule, this equation reduces to (8.2.12).

Before investigating properties of solutions of the optimality equation, we provide a *heuristic* derivation which we later formalize in Sec. 8.4.2. to demonstrate the existence of a solution. Begin with the optimality equation for the discounted reward

$$0 = \max_{d \in D} \{ r_d + (\lambda P_d - I) v_\lambda^* \}$$

and assume that v_λ^* has the partial Laurent series expansion

$$v_\lambda^* = (1 - \lambda)^{-1} g^* e + h + f(\lambda)$$

where, as before, $f(\lambda)$ denotes a vector which converges pointwise to 0 as $\lambda \uparrow 1$. By substituting this expression into the optimality equation, we obtain

$$\begin{aligned} 0 &= \max_{d \in D} \left\{ r_d + (\lambda P_d - I) \left[\frac{g^* e}{1 - \lambda} + h + f(\lambda) \right] \right\} \\ &= \max_{d \in D} \{ r_d - g^* e + (\lambda P_d - I) h + f(\lambda) \}. \end{aligned}$$

The second equation follows from the first by noting that, since g^* is constant, $P_d g^* e = g^* e$ for all $d \in D$. Taking the limit as $\lambda \uparrow 1$ suggests that (8.4.3) is the appropriate form for the optimality equation.

We now provide a different heuristic derivation based on the finite-horizon optimality equations. Eq. (8.2.5) suggests that

$$v_N^* = (N - 1) g^* e + h + o(1).$$

Rewrite the finite-horizon optimality equations (4.5.1) in this notation as

$$v_{N+1}^* = \max_{d \in D} \{ r_d + P_d v_N^* \}$$

and substitute in the above expression for v_N^* to obtain

$$N g^* e + h + o(1) = \max_{d \in D} \{ r_d + (N - 1) P_d g^* e + P_d h + o(1) \}.$$

Again, noting that $P_d g^* e = g^* e$ for all $d \in D$, rearranging terms and choosing N sufficiently large establishes (8.4.3).

We now provide bounds on g^* based on subsolutions and supersolutions of the optimality equation and show that, when the unichain optimality equation has a solution with bounded h, the average, lim inf average and lim sup average optimality criteria of Sec. 8.1.2 are equivalent. The following theorem is the average reward analog of Theorem 6.2.2 and one of the most important results for average reward models; part (c) gives the main result.

Theorem 8.4.1. Suppose S is countable.

a. If there exists a scalar g and an $h \in V$ which satisfy $B(g, h) \leq 0$, then

$$ge \geq g_+^*. \tag{8.4.4}$$

b. If there exists a scalar g and $h \in V$ which satisfy $B(g, h) \geq 0$, then

$$ge \leq \sup_{d \in D^{\mathrm{MD}}} g_-^{d^\infty} \leq g_-^*. \tag{8.4.5}$$

c. If there exists a scalar g and an $h \in V$ for which $B(g, h) = 0$, then

$$ge = g^* = g_+^* = g_-^*. \tag{8.4.6}$$

Proof. Since $B(g, h) \leq 0$, Proposition 6.2.1 implies that

$$ge \geq r_d + (P_d - I)h \tag{8.4.7}$$

for all $d \in D^{\mathrm{MR}}$. Let $\pi = (d_1, d_2, \dots) \in \Pi^{\mathrm{MR}}$. Then, (8.4.7) implies that

$$ge \geq r_{d_1} + (P_{d_1} - I)h.$$

Applying (8.4.7) with $d = d_2$ and multiplying it by P_{d_1} yields

$$ge = gP_{d_1}e \geq P_{d_1}r_{d_2} + (P_{d_1}P_{d_2} - P_{d_1})h. \tag{8.4.8}$$

Repeating this argument (or using induction), for any $n \geq 2$ shows that

$$ge \geq P_{d_1}P_{d_2} \cdots P_{d_{n-1}}r_{d_n} + P_{d_1}P_{d_2} \cdots P_{d_{n-1}}(P_{d_n} - I)h.$$

Summing these expressions over n and noting (8.1.3) shows that, for all $\pi \in \Pi^{\mathrm{MR}}$ and any N,

$$Nge \geq v_{N+1}^\pi + (P_N^\pi - I)h.$$

Since $h \in V$, $P_N^\pi h \in V$, so that $\lim_{N\to\infty} N^{-1}(P_N^\pi - I)h(s) = 0$ for each $s \in S$. Therefore,

$$ge \geq \limsup_{N\to\infty} \frac{1}{N} v_{N+1}^\pi = g_+^\pi$$

for all $\pi \in \Pi^{\mathrm{MR}}$. Extension to $\pi \in \Pi^{\mathrm{HR}}$ follows from Theorem 8.1.2, so (a) follows.

To establish part (b), note that $B(g, h) \geq 0$ implies there exists a $d^* \in \mathrm{D}^{\mathrm{MD}}$ for which

$$ge \leq r_{d^*} + (P_{d^*} - I)h.$$

Applying the above argument with P_{d^*} replacing P_{d_n} establishes that

$$ge \leq \liminf_{N \to \infty} \frac{1}{N} v_{N+1}^{(d^*)^\infty} = g_-^{(d^*)^\infty} \leq g_-^*,$$

from which part (b) follows.

Under (c), (a), and (b), hold so that $ge \leq g_-^* \leq g_+^* \leq ge$, from which (8.4.6) follows. □

The constant gain assumption was used to establish the identity on the left-hand side of (8.4.8). Actually all we required for the proof to be valid was that $g \geq P_{d_N} g$. This holds with equality for constant g. Without such an assumption we need further conditions to assure that $g \geq P_{d_N} g$ for arbitrary g. We return to this point when we analyze general models in Chap. 9.

The above theorem holds without any assumptions about model classification, but it is most useful when we know *a priori* that the optimal gain does not depend on the initial state, as for example, in a unichain model. The bounds in parts (a) and (b) are always valid, but they are not of much practical significance when the quantities on the right-hand side of (8.4.4) and (8.4.5) vary with s. More importantly, the hypothesis of part (c) will be vacuous unless the optimal gain is constant. The following example illustrates these points.

Example 8.4.2. Let $S = \{s_1, s_2\}$; $A_{s_1} = \{a_{1,1}, a_{1,2}\}$ and $A_{s_2} = \{a_{2,1}\}$; and $r(s_1, a_{1,1}) = 1$, $r(s_1, a_{1,2}) = 3$, and $r(s_2, a_{2,1}) = 2$. All transitions are deterministic and as indicated in Fig. 8.4.1 (Fig. 8.4.2). Obviously the stationary policy which uses action $a_{1,2}$ in s_1 and $a_{2,1}$ in s_2 is optimal. Further the corresponding Markov chain is multichain with $g^*(s_1) = 3$, and $g^*(s_2) = 2$. For arbitrary g and h

$$B(g,h)(s_1) = \max\{1 - g + h(s_2) - h(s_1), 3 - g\}$$

$$B(g,h)(s_2) = 2 - g.$$

Note that $h(s_1) = h(s_2) = 0$, $g = 3$ satisfies $B(g,h) \leq 0$, so, by Theorem 8.4.1a, $ge \geq g^*$. If we choose $g = 2$, $B(g,h) \geq 0$, so that, by part (b) of the above theorem, $ge \leq g^*$. Thus we have established the bounds $2e \leq g^* \leq 3e$ but no tighter bounds

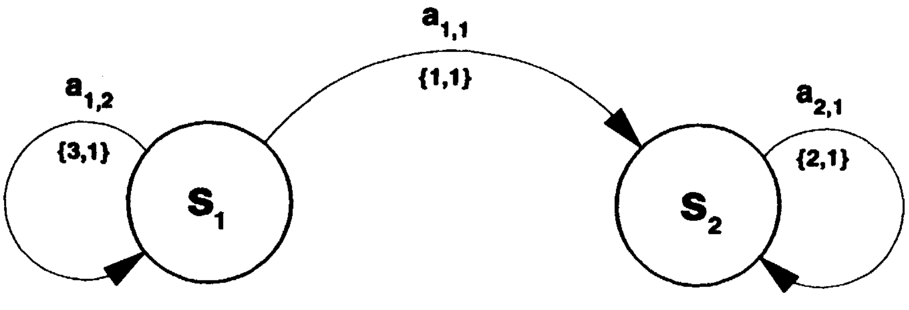

Figure 8.4.2 Symbolic representation of Example 8.4.2.

are available through the above theorem. Note further that there exists no scalar g for which $B(g, h) = 0$, so part (c) of Theorem 8.4.1 is vacuous.

If instead $r(s_1, a_{1,2}) = 2$, then $g^* = 2e$ even though the Markov chain generated by an optimal policy was multichain. In this case $g = 2$ and h as above satisfy $B(g, h) = 0$, so part (c) of Theorem 8.4.1 applies to this particular multichain model.

In the proof of Theorem 8.4.1 we used the assumption of bounded h to assure that $\lim_{N\to\infty} N^{-1}P_N^\pi h(s) = 0$ for all $s \in S$. For the proof to be valid we require only that this limit exists and equal 0. We express this formally as follows.

Corollary 8.4.2. Suppose that

$$\lim_{N\to\infty} N^{-1}E_s^\pi\{h(X_N)\} = 0 \tag{8.4.9}$$

for all $\pi \in \Pi^{\text{HR}}$ and $s \in S$. Then results (a) – (c) in Theorem 8.4.1 hold.

Note that if the maximum over D is not attained, and "sup" replaces "max" in (8.4.4), (8.4.6), and (8.4.8), we can prove a similar result. We leave establishing the results of Theorem 8.4.1 under these weaker assumptions as an exercise. This theorem also extends to general S.

8.4.2 Existence of Solutions to the Optimality Equation

In this section we show that the unichain optimality equation (8.4.3) has a solution. Our approach formalizes the heuristic argument at the beginning of the preceding section. Other methods for demonstrating this result include policy iteration (Sec. 8.6), compactness of the set of stationary policies (Secs. 8.4.4 and 8.10), and approaches based on convex analysis and fixed point theory. Refer to the Bibliographic Remarks section of this chapter for references for the last two approaches.

Theorem 8.4.3. Suppose S and A_s are finite Assumption 8.0.2 holds and the model is unichain.

a. Then there exists a $g \in R^1$ and an $h \in V$ for which

$$0 = \max_{d\in D}\{r_d - ge + (P_d - I)h\}.$$

b. If (g', h') is any other solution of the average reward optimality equation, then $g = g'$.

Proof. Choose a sequence of discount factors $\{\lambda_n\}$, $0 \le \lambda_n < 1$ with the property that $\lambda_n \uparrow 1$. By Theorem 6.2.10a, for each λ_n there exists a stationary discount optimal policy. Since D^{MD} is finite, we can choose a subsequence $\{\lambda'_n\}$ for which the same policy, δ^∞ is discount optimal for all λ'_n. Denote this subsequence by $\{\lambda_n\}$. Since δ^∞ is discount optimal for $\lambda = \lambda_n$, $v_{\lambda_n}^* = v_{\lambda_n}^{\delta^\infty}$. Therefore, for any $d \in D^{\mathrm{MD}}$,

$$\begin{aligned} 0 &= r_\delta + (\lambda_n P_\delta - I)v_{\lambda_n}^{\delta^\infty} = r_\delta + (\lambda_n P_\delta - I)v_{\lambda_n}^* \\ &= \max_{d' \in D}\left\{r_{d'} + (\lambda_n P_{d'} - I)v_{\lambda_n}^*\right\} \qquad (8.4.10) \\ &\geq r_d + (\lambda_n P_d - I)v_{\lambda_n}^* = r_d + (\lambda_n P_d - I)v_{\lambda_n}^{\delta^\infty}. \end{aligned}$$

By Corollary 8.2.4,

$$v_{\lambda_n}^{\delta^\infty} = (1 - \lambda_n)^{-1} g^{\delta^\infty} e + h^{\delta^\infty} + f(\lambda_n). \qquad (8.4.11)$$

Noting that h^{δ^∞} is bounded implies

$$\lambda_n P_d h^{\delta^\infty} = P_d h^{\delta^\infty} + (\lambda_n - 1)P_d h^{\delta^\infty} = P_d h^{\delta^\infty} + f'(\lambda_n) \qquad (8.4.12)$$

for all $d \in D$, where $f'(\lambda)$ converges to 0 as $\lambda \uparrow 1$. Substituting (8.4.11) into the first expression in the sequence of equations (8.4.10), noting (8.4.12) and that $(\lambda_n P_d - I)e = (\lambda - 1)e$, implies

$$\begin{aligned} 0 &= r_\delta + (\lambda_n P_\delta - I)\left\{(1 - \lambda_n)^{-1} g^{\delta^\infty} e + h^{\delta^\infty} + f(\lambda_n)\right\} \\ &= r_\delta - g^{\delta^\infty} e + (P_\delta - I)h^{\delta^\infty} + f''(\lambda_n), \end{aligned}$$

where $f''(\lambda)$ converges to 0 as $\lambda \uparrow 1$. Performing similar operations to the last expression in the sequence of equations (8.4.10) establishes that

$$\begin{aligned} &r_d + (\lambda_n P_d - I)\left\{(1 - \lambda_n)^{-1} g^{\delta^\infty} e + h^{\delta^\infty} + f(\lambda_n)\right\} \\ &\quad = r_d - g^{\delta^\infty} e + (P_d - I)h^{\delta^\infty} + f_d(\lambda_n) \qquad (8.4.13) \end{aligned}$$

for all $d \in D$, where $f_d(\lambda)$ denotes a vector which converges to zero as $\lambda \uparrow 1$. Therefore

$$r_\delta - g^{\delta^\infty} e + (P_\delta - I)h^{\delta^\infty} + f(\lambda_n) \geq r_d - g^{\delta^\infty} e + (P_d - I)h^{\delta^\infty} + f_d(\lambda_n).$$

Taking the limit as $\lambda_n \uparrow 1$ shows that

$$0 = r_\delta - g^{\delta^\infty} e + (P_\delta - I)h^{\delta^\infty} \geq r_d - g^{\delta^\infty} e + (P_d - I)h^{\delta^\infty}$$

for each $d \in D$, from which result (a) follows.

To prove part (b), note that $g = g' = g^*$ follows from Theorem 8.4.1(c). □

Some comments regarding the above result and its proof follow. Note that, in addition to establishing that the unichain average reward optimality equation has a solution, the above proof identifies this solution as the gain and bias of the stationary policy δ^∞. Consequently, by Theorem 8.4.1(c), $g^{\delta^\infty} = g^*$ and δ^∞ is average optimal. Note also that this analysis applies to models in which "sup" replaces "max" in the optimality equations.

We have shown that $(g^{\delta^\infty}, h^{\delta^\infty})$ satisfies the optimality equation, however, the solution is not unique since $(g^{\delta^\infty}, h^{\delta^\infty} + ke)$ is also a solution for any scalar k.

Schweitzer and Federgruen (1978a) show that in unichain models, this gives a complete characterization of the set of solutions of the optimality equation.

Note that the argument used to establish Theorem 8.4.3 does not extend directly to countable-state models because the partial Laurent series expansions in Corollary 8.4.2 are not available without additional assumptions.

We conclude this section with an example which illustrates the optimality equation and properties of its solution.

Example 8.4.3. We analyze an infinite-horizon average reward version of the example in Fig. 3.1.1. There is a single decision in state s_2 and two decisions in s_1, $a_{1,1}$ and $a_{1,2}$. Let δ use action $a_{1,1}$ and γ use action $a_{1,2}$. Both δ^∞ and γ^∞ are unichain with recurrent state s_2 and transient state s_1. The optimality equations are

$$\begin{aligned} 0 &= \max\{5 - g - 0.5h(s_1) + 0.5h(s_2),\ \ 10 - g - h(s_1) + h(s_2)\}, \\ 0 &= -1 - g. \end{aligned} \tag{8.4.14}$$

Since the second equation above implies that $g^* = -1$, we know without any further analysis that δ^∞ and γ^∞ are average optimal. We proceed to solve the optimality equation. We apply Corollary 8.2.7 to find $(g^{\delta^\infty}, h^{\delta^\infty})$ and $(g^{\eta^\infty}, h^{\gamma^\infty})$ by solving

$$ge + (I - P_d)h = r_d$$

subject to $P_\delta^* h = 0$ or $P_\gamma^* h = 0$. Noting that

$$P_\delta^* = P_\gamma^* = \begin{bmatrix} 0.0 & 1.0 \\ 0.0 & 1.0 \end{bmatrix}$$

implies that $h^{\delta^\infty}(s_2) = h^{\gamma^\infty}(s_2) = 0$. Consequently $g^{\delta^\infty} = g^{\gamma^\infty} = -1$, $h^{\delta^\infty}(s_1) = 12$ and $h^{\gamma^\infty}(s_1) = 11$. Observe that $(g^{\delta^\infty}, h^{\delta^\infty})$ satisfies (8.4.14), as does $(g^{\delta^\infty}, h^{\delta^\infty} + ke)$ for any constant k, but that $(g^{\gamma^\infty}, h^{\gamma^\infty})$ does not.

Hence, even though γ^∞ is average optimal, its gain and bias together do not satisfy the optimality equation. We elaborate on this point in Chap. 10.

8.4.3 Identification and Existence of Optimal Policies

In this section we show how to use the optimality equation to find a stationary optimal policy and discuss some results on the existence of optimal policies. We call a decision rule d_h *h-improving* if

$$d_h \in \underset{d \in D}{\arg\max}\{r_d + P_d h\}$$

or, equivalently,

$$r_{d_h} + P_{d_h} h = \max_{d \in D}\{r_d + P_d h\}.$$

This means that

$$r_{d_h} - ge + (P_{d_h} - I)h = \max_{d \in D}\{r_d - ge + (P_d - I)h\}. \tag{8.4.15}$$

Theorem 8.4.4. Suppose there exists a scalar g^*, and an $h^* \in V$ for which $B(g^*, h^*) = 0$. Then, if d^* is h^*-improving, $(d^*)^\infty$ is average optimal.

Proof. By hypothesis

$$0 = r_{d^*} - g^*e + (P_{d^*} - I)h^*.$$

From Corollary 8.2.7 $g^{(d^*)^\infty} = g^*$ so that Theorem 8.4.1(c) establishes the optimality of $(d^*)^\infty$. □

Example 8.4.3 shows that the converse of the above theorem need not hold. In it, γ^∞ is average optimal but not h^*-improving.

Theorem 8.4.3 established the existence of a solution to $B(g, h) = 0$ under the hypothesis of finite S and A_s. Combining this with the above result and Theorem 8.4.1 establishes the following existence results.

Theorem 8.4.5. Suppose S is finite and A_s is finite for each $s \in S$, $r(s, a)$ is bounded and the model is unichain. Then

a. there exists a stationary average optimal policy,

b. there exists a scalar g^* and an $h^* \in V$ for which $B(g^*, h^*) = 0$,

c. any stationary policy derived from an h^*-improving decision rule is average optimal, and

d. $g^*e = g_+^* = g_-^*$.

*8.4.4 Models with Compact Action Sets

In this section we establish the existence of solutions to the optimality equations, and the existence of optimal policies for unichain models with finite-state and compact action sets. The existence proof generalizes that of Theorem 8.4.3 and relies on continuity properties of the limiting matrix and deviation matrix in unichain models. We begin with assumptions and technical results.

Assumption 8.4.1. For each $s \in S$, $r(s, a)$ is a bounded, continuous function of a.

Assumption 8.4.2. For each $s \in S$ and $j \in S$, $p(j|s, a)$ is a continuous function of a.

The following result, which holds under weaker assumptions, relies on results in Appendix C which establish the continuity of matrix inverses and products.

Proposition 8.4.6. Let $\{P_n\}$ denote a sequence of unichain transition probability matrices and suppose

$$\lim_{n \to \infty} \|P_n - P\| = 0, \tag{8.4.16}$$

then

a. $\lim_{n \to \infty} \|P_n^* - P^*\| = 0$, and
b. $\lim_{n \to \infty} \|H_{P_n} - H_P\| = 0$.

Proof. For each n, P_n is unichain, so, by results in Appendix A, $P_n^* = q_n e^T$, where q_n is the unique solution of

$$q_n^T(I - P_n) = 0$$

subject to $q_n^T e = 1$. Let W_n denote the matrix which replaces the first column of $(I - P_n)$ by the column vector e, and W the same matrix formed from $I - P$. Thus q_n is the unique solution of

$$u^T W_n = z^T$$

where z denotes a column vector with 1 in the first component and 0 in the remaining components. Let q denote the unique solution of $q^T W = z^T$. Consequently, W_n^{-1} and W^{-1} exist, $q_n^T = z^T W_n^{-1}$, and $q^T = z^T W^{-1}$.

We show that q_n converges to q. By (8.4.16), $\lim_{n \to \infty} \|W_n - W\| = 0$, so, by Proposition C.5, $\lim_{n \to \infty} \|W_n^{-1} - W^{-1}\| = 0$. Therefore

$$\lim_{n \to \infty} q_n^T = \lim_{n \to \infty} z^T W_n^{-1} = z^T W^{-1} = q^T,$$

so (a) follows by noting that $P^* = qe^T$.

To establish (b), note that $H_{P_n} = (I - P_n + P_n^*)^{-1}(I - P_n^*)$. Since $(I - P - P^*)^{-1}$ exists, the result follows from part (a), Proposition C.5, and Lemma C.7. □

The main result of this section follows. After taking some technical considerations into account, its proof follows Theorem 8.4.3.

Theorem 8.4.7. Suppose S is finite, A_s is compact, the model is unichain, and Assumptions 8.4.1 and 8.4.2 hold.

a. Then there exists a $g \in R^1$ and an $h \in V$ for which

$$0 = \max_{d \in D} \{ r_d - ge + (P_d - I)h \}.$$

b. If (g', h') is any other solution of the average reward optimality equation, then $g = g'$.

c. There exists a $d^* \in D^{\text{MD}}$ for which $(d^*)^\infty$ is average optimal; further, if d is h-improving, then d^∞ is average optimal.

Proof. Choose a sequence of discount factors $\{\lambda_n\}$, $0 \le \lambda_n < 1$, with $\lambda_n \uparrow 1$. By Theorem 6.2.10(a) for each λ_n, there exists a stationary discount optimal policy $(\delta_n)^\infty$. Since $D^{\text{MD}} = \times_{s\in S} A_s$ is compact, we can choose a subsequence (n_k) for which $\delta_{n_k}(s)$ converges to a limit $\delta(s) \in D^{\text{MD}}$ for each $s \in S$. To simplify subsequent notation, denote the subsequence by $\{\delta_k\}$.

By Assumption 8.4.2, $p_{\delta_k}(j|s)$ converges to $p_\delta(j|s)$ for each $(s, j) \in S \times S$, so, by the finiteness of S, (8.4.16) holds. Consequently, by Proposition 8.4.6, $P^*_{\delta_k}$ converges in norm to P^*_δ and H_{δ_k} converges in norm to H_δ. As a result of Assumption 8.4.1, r_{δ_k} converges in norm to r_δ, so, by Lemma C.6,

$$\lim_{k\to\infty} g^{(\delta_k)^\infty} e = \lim_{k\to\infty} P^*_{\delta_k} r_{\delta_k} = P^*_\delta r_\delta = g^{\delta^\infty} e \tag{8.4.17}$$

and

$$\lim_{k\to\infty} h^{(\delta_k)^\infty} = \lim_{k\to\infty} H_{\delta_k} r_{\delta_k} = H_\delta r_\delta = h^{\delta^\infty}. \tag{8.4.18}$$

As a consequence of the optimality of $(\delta_k)^\infty$, for any $d \in D$,

$$0 = r_{\delta_k} + (\lambda_k P_{\delta_k} - I) v_{\lambda_k}^{(\delta_k)^\infty} \ge r_d + (\lambda_k P_d - I) v_{\lambda_k}^{(\delta_k)^\infty}.$$

By Corollary 8.2.4,

$$v_{\lambda_k}^{(\delta_k)^\infty} = (1 - \lambda_k)^{-1} g^{(\delta_k)^\infty} e + h^{(\delta_k)^\infty} + f(\lambda_k)$$

so that

$$\begin{aligned} 0 &= r_{\delta_k} - g^{(\delta_k)^\infty} e + (P_{\delta_k} - I) h^{(\delta_k)^\infty} + f(\lambda_k) \\ &\ge r_d - g^{(\delta_k)^\infty} e + (P_d - I) h^{(\delta_k)^\infty} + f_d(\lambda_k) \end{aligned}$$

for each $d \in D$, where $f_d(\lambda)$ denotes a vector which converges to 0 as $\lambda \uparrow 1$. Hence taking the limit as $\lambda_k \uparrow 1$ shows that

$$0 = r_\delta - g^{\delta^\infty} e + (P_\delta - I) h^{\delta^\infty} \ge r_d - g^{\delta^\infty} e + (P_d - I) h^{\delta^\infty}$$

for all $d \in D$, so that $B(g^{\delta^\infty}, h^{\delta^\infty}) = 0$ from which result (a) follows.

Part (b) follows as in Theorem 8.4.3. Since $r(s, a) + \sum_{j\in S} p(j|s, a) h(j)$ is continuous in a, and A_s is compact, $\arg\max_{a\in A_s}\{r(s, a) + \sum_{j\in S} p(j|s, a) h(j)\}$ is nonempty for each $s \in S$, so that part (c) follows from Theorem 8.4.4. □

Note that (8.4.17) and (8.4.18) establish the continuity of g^{d^∞} and h^{d^∞} in d, so that both of these quantities attain maxima on compact sets of decision rules.

8.5 VALUE ITERATION IN UNICHAIN MODELS

In this section we study value iteration for unichain average reward MDPs. We express the sequence of values generated by value iteration as

$$v^{n+1} = Lv^n, \tag{8.5.1}$$

where

$$Lv \equiv \max_{d \in D} \{r_d + P_d v\} \tag{8.5.2}$$

for $v \in V$. We assume throughout this section that the maximum is attained in (8.5.2). In discounted, positive, and negative models, we based value iteration on operators of this form, but analysis here requires more subtle arguments because we cannot appeal to contraction or monotonicity to establish convergence. In fact, in almost all practical applications, the sequence $\{v^n\}$ generated by (8.5.1) diverges or oscillates, so we must seek other ways for terminating algorithms and selecting optimal policies.

Our approach relies on material in Sec. 6.6, and we suggest reviewing it before proceeding. For a more general analysis of value iteration in average reward models, see Sec. 9.4. Throughout this section we assume finite S and A_s although results hold in greater generality.

8.5.1 The Value Iteration Algorithm

The following value iteration algorithm finds a stationary ε-optimal policy $(d^\varepsilon)^\infty$ and an approximation to its gain, when certain extra conditions are met.

Value Iteration Algorithm

1. Select $v^0 \in V$, specify $\varepsilon > 0$ and set $n = 0$.
2. For each $s \in S$, compute $v^{n+1}(s)$ by

$$v^{n+1}(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^n(j) \right\}. \tag{8.5.3}$$

3. If

$$sp(v^{n+1} - v^n) < \varepsilon, \tag{8.5.4}$$

go to step 4. Otherwise increment n by 1 and return to step 2.

4. For each $s \in S$, choose

$$d_\varepsilon(s) \in \arg\max_{a \in A_s} \left\{ r(s,\dot{a}) + \sum_{j \in S} p(j|s,a) v^n(j) \right\} \tag{8.5.5}$$

and stop.

The examples below illustrate the difficulties inherent in applying value iteration.

Example 8.5.1. Let $S = \{s_1, s_2\}$ and suppose there is a single decision rule d, with

$$r_d = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad P_d = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

If

$$v^0 = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \text{then} \quad v^n = P_d^n v^0 = \begin{cases} \begin{bmatrix} a \\ b \end{bmatrix} & n \text{ even} \\ \begin{bmatrix} b \\ a \end{bmatrix} & n \text{ odd}, \end{cases}$$

so unless $a = b$, $\lim_{n \to \infty} v^n$ does not exist and $\{v^n\}$ oscillates with period 2. Note that $\|v^{n+1} - v^n\| = |b - a|$ and $sp(v^{n+1} - v^n) = 2|b - a|$, so (8.5.4) is never satisfied in this example. Note also that each state is periodic with period 2.

Example 8.5.2. Consider the model in Fig. 3.1.1. In it, state s_2 is absorbing, $g^* = -1$, and $g^\pi = -1$ for any policy π. Choosing $v^0 = 0$, v^n equals

n	$v^n(s_1)$	$v^n(s_2)$	$sp(v^n - v^{n-1})$
0	0	0	NA
1	10	-1	11
2	9.5	-2	.5
3	8.75	-3	.25
4	7.875	-4	.125
5	6.9375	-5	.0625
⋮	⋮	⋮	⋮

so we see that, $v^n(s_1)$ approaches $12 - n$ and $v^n(s_2) = -n$.

Observe that if $\|v^{n+1} - v^n\| = 1$ for $n \geq 1$ but $sp(v^n - v^{n-1}) = 0.5^{n-1}$ for $n \geq 1$, so that stopping criterion (8.5.4) holds at $n = 5$. Note that v^n diverges linearly in n. More precisely,

$$v^n = ng^* + h,$$

where $h(s_1) = 12$ and $h(s_2) = 0$, as suggested in Example 8.4.3. Consequently, $g^* = v^{n+1}(s) - v^n(s)$ for $n \geq 1$ and $s \in S$.

The first example above shows that value iteration need not converge in models with periodic transition matrices, while the second example shows that the sequence may diverge but that $sp(v^{n+1} - v^n)$ may converge. In subsequent sections, we provide conditions which ensure the above algorithm terminates in a finite number of iterations, and that $(d_\varepsilon)^\infty$ is an ε-optimal policy.

8.5.2 Convergence of Value Iteration

In this section we provide conditions under which stopping criterion (8.5.4) holds for some finite n. Theorem 6.6.6 shows that

$$sp(v^{n+2} - v^{n+1}) \leq \gamma sp(v^{n+1} - v^n), \tag{8.5.6}$$

where

$$\gamma = \max_{s \in S, a \in A_s, s' \in S, a' \in A_{s'}} \left[1 - \sum_{j \in S} \min\{p(j|s,a), p(j|s',a')\} \right]. \tag{8.5.7}$$

Consequently, if $\gamma < 1$, (8.5.6) ensures that in a finite number of iterations, stopping criterion (8.5.4) will be satisfied. The following example shows that γ may equal 1 in a unichain aperiodic model, but that value iteration may still converge. It motivates a more general condition which ensures convergence.

Example 8.5.3. Let $S = \{s_1, s_2, s_3\}$; $A_{s_1} = \{a_{1,1}, a_{1,2}\}$, $A_{s_2} = \{a_{2,1}\}$, and $A_{s_3} = \{a_{3,1}\}$; $r(s_1, a_{1,1}) = 2$, $r(s_1, a_{1,2}) = 1$, $r(s_2, a_{2,1}) = 2$, and $r(s_3, a_{3,1}) = 3$; and $p(s_3|s_1, a_{1,1}) = 1$, $p(s_2|s_1, a_{1,2}) = 1$, $p(s_1|s_2, a_{2,1}) = 1$, and $p(s_1|s_3, a_{3,1}) = p(s_2|s_3, a_{3,1}) = p(s_3|s_3, a_{3,1}) = \frac{1}{3}$ (Fig. 8.5.1).

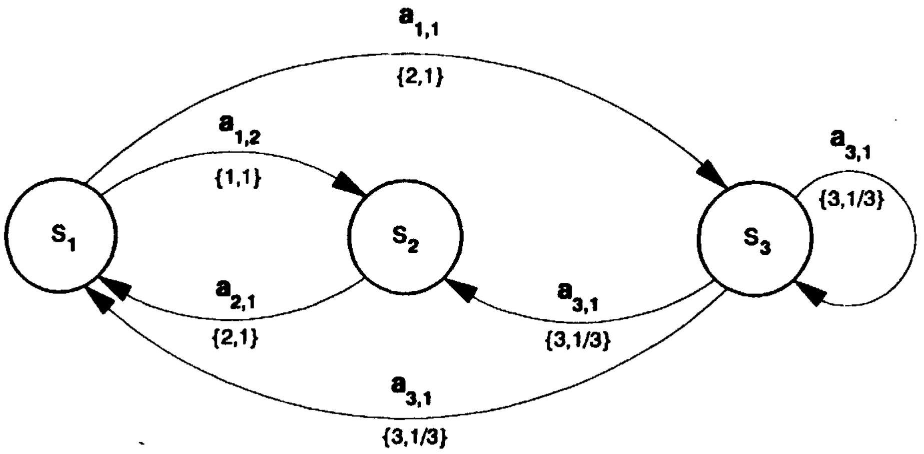

Figure 8.5.1 Symbolic representation of Example 8.5.3.

Let δ denote the decision rule which uses action $a_{1,1}$ in s_1, and η denote the decision rule which uses $a_{1,2}$ in s_1. Observe that

$$P_\delta = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}, \quad P_\eta = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix},$$

so that $\gamma_\delta = \gamma_\eta = \gamma = 1$ where γ_d is defined in (6.6.5). Note also that, under η, states s_1 and s_2 are periodic with period 2 and s_3 is transient, while under δ all states are recurrent and aperiodic.

Applying value iteration with $(v^0)^T = (0, 0, 0)$ and $\varepsilon = 0.01$ yields

n	$v^n(s_1)$	$v^n(s_2)$	$v^n(s_3)$	$sp(v^{n+1} - v^n)$	$\max_{s \in S}\{v^{n+1}(s) - v^n(s)\}$	$\min_{s \in S}\{v^{n+1}(s) - v^n(s)\}$
0	0.000	0.000	0.000			
1	2.000	2.000	3.000	1.000	3.000	2.000
2	5.000	4.000	5.333	1.000	3.000	2.000
3	7.333	7.000	7.778	0.667	3.000	2.333
4	9.778	9.333	10.370	0.259	2.592	2.333
5	12.370	11.778	12.827	0.147	2.592	2.445
6	14.827	14.370	15.325	0.135	2.592	2.457
7	17.325	16.827	17.841	0.059	2.516	2.457
8	19.841	19.325	20.331	0.026	2.516	2.490
9	22.331	21.841	22.832	0.026	2.516	2.490
10	24.832	24.331	25.335	0.013	2.503	2.490
11	27.335	26.832	27.833	0.005	2.503	2.498
12	29.833	29.335	30.333	0.005	2.503	2.498

Observe that the value iteration algorithm stops in step 3 with $n = 11$ and identifies the 0.01-optimal policy δ^∞ through step 4. Note also that $sp(v^{n+1} - v^n)$ is monotone but that it does not decrease at each iteration.

The above example shows that value iteration may converge with respect to the span seminorm even though (8.5.6) holds with $\gamma = 1$. Our approach for showing that this algorithm stops in a finite number of iterations relies on the concept of a J-stage span contraction. We say that an operator T: $V \to V$ is a *J-stage span contraction* if there exists an α, $0 \leq \alpha < 1$ and a non-negative integer J for which

$$sp(T^J u - T^J v) \leq \alpha \, sp(u - v) \tag{8.5.8}$$

for all u and v in V.

Proposition 8.5.1. Suppose T is a J-stage span contraction. Then for any $v^0 \in V$, the sequence $v^n = T^n v^0$ satisfies

$$sp(v^{nJ+1} - v^{nJ}) \leq \alpha^n \, sp(v^1 - v^0) \tag{8.5.9}$$

for all non-negative integers n.

Proof. Choose $v = v^0$ and $u = Tv^0$ and iterate (8.5.8). □

The above proposition provides a simple property of J-stage span contractions. We leave it as an exercise to generalize Theorem 6.6.2 which establishes existence of a span fixed point and convergence of $T^n v^0$ to it. The following result generalizes Theorem 6.6.6.

Theorem 8.5.2. Suppose there exists an integer $J \geq 1$ such that, for every pair of deterministic Markov policies π_1 and π_2,

$$\eta(\pi_1, \pi_2) \equiv \min_{(s,u) \in S \times S} \sum_{j \in S} \min\left\{P^J_{\pi_1}(j|s), P^J_{\pi_2}(j|u)\right\} > 0. \tag{8.5.10}$$

a. Then L defined in (8.5.2) is J-step contraction operator on V with contraction coefficient

$$\gamma' = 1 - \min_{\pi_1, \pi_2 \in \Pi^{\mathrm{MD}}} \eta(\pi_1, \pi_2). \tag{8.5.11}$$

b. For any $v^0 \in V$, let $v^n = L^n v^0$. Then, given $\varepsilon > 0$, there exists an N such that

$$sp(v^{nJ+1} - v^{nJ}) \leq \varepsilon$$

for all $n \geq N$.

Proof. The proof of Theorem 6.6.2 may be easily adapted to establish part (a). Part (b) follows by applying Proposition 8.5.1 with $\alpha = \gamma'$. □

Condition (8.5.10) means that starting in any pair of distinct states, policies π_1 and π_2 both reach at least one identical state with positive probability after J transitions. Since (8.5.10) must hold for any pair of policies, all stationary policies must be unichain and aperiodic under (8.5.10). This condition excludes periodic models such as that in Example 8.5.1 and multichain models.

Clearly it is not easy to verify (8.5.11) directly. The following theorem provides conditions which are easier to check and imply it. We leave it as an exercise to show that each of these conditions are distinct; that is, that any one of them does not imply any other.

Theorem 8.5.3. Suppose either

a. $0 \leq \gamma < 1$, where γ is given in (8.5.7),

b. there exists a state $s' \in S$ and an integer K such that, for any deterministic Markov policy π, $P^K_\pi(s'|s) > 0$ for all $s \in S$, or

c. all policies are unichain and $p(s|s, a) > 0$ for all $s \in S$ and $a \in A_s$.

Then (8.5.10) holds for all π_1 and π_2 in Π^{MD} and the conclusions of Theorem 8.5.2 follow.

Proof. If (a) holds, then (8.5.10) holds with $J = 1$. Since

$$\eta(\pi_1, \pi_2) \geq \min_{(s,u) \in S \times S} \left\{P^J_{\pi_1}(s'|s), P^J_{\pi_2}(s'|u)\right\},$$

(b) implies that (8.5.10) holds with $J = K$.

We now consider (c). Let $\pi_1 = (d_1, d_2, \dots)$ and $\pi_2 = (f_1, f_2, \dots)$ denote two deterministic Markovian policies, and choose $s_1 \neq s_2$ in S. Let

$$X_i(n) \equiv \left\{ j \in S \colon P^n_{\pi_i}(j|s_i) > 0 \right\}.$$

Let N denote the number of elements in S. We show by contradiction that $X_1(N) \cap X_2(N) \neq \varnothing$. Suppose $X_1(N) \cap X_2(N) = \varnothing$. Since $p(s|s,a) > 0$, $X_i(n) \subset X_i(n+1)$ for all n, so that $X_1(n) \cap X_2(n) = \varnothing$ for all $1 \leq n \leq N$. Consequently, for some $m < N$, $X_1(m) = X_1(m+1)$ and $X_2(m) = X_2(m+1)$. This means that $X_1(m)$ is closed under d_m and $X_2(m)$ is closed under f_m. Construct a decision rule δ by

$$\delta(s) = \begin{cases} d_m(s) & s \in X_1(m) \\ f_m(s) & s \in X_2(m) \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

Then P_δ has at least two distinct closed classes contradicting hypothesis (c). Therefore $X_1(N) \cap X_2(N) \neq \varnothing$, so that, for some j', $P^N_{\pi_1}(j'|s_1) > 0$ and $P^N_{\pi_2}(j'|s_2) > 0$, which shows that (8.5.10) holds. Since π_1 and π_2 are arbitrary, the result follows. □

Thus under any of the conditions in Theorem 8.5.3, value iteration achieves stopping criterion (8.5.4) for any $\varepsilon > 0$. Section 8.5.5 provides a method for transforming any unichain model into one which satisfies condition (c).

We illustrate Theorem 8.5.3 by expanding on Example 8.5.3.

Example 8.5.3 (ctd.). Inspection of Example 8.5.3 reveals that, when starting the algorithm with $v^0 = 0$, action $a_{1,1}$ always achieves the maximum in s_1, so that, for $n > 0$,

$$r_\delta(s) + \sum_{j \in S} p_\delta(j|s) v^n(j) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^n(j) \right\}$$

for all $s \in S$. Consequently, in this example value iteration corresponds to iterating the fixed policy δ^∞. Note that P_δ satisfies neither hypotheses (a) nor (c) of Theorem 8.5.3. Since

$$P^2_\delta = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 \\ \frac{4}{9} & \frac{1}{9} & \frac{4}{9} \end{bmatrix}, \qquad P^3_\delta = \begin{bmatrix} \frac{4}{9} & \frac{4}{9} & \frac{1}{9} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{7}{27} & \frac{16}{27} & \frac{4}{27} \end{bmatrix},$$

(b) holds with $K = 2$ and $s' = s_3$. Note also that P^3_δ satisfies (c).

Since P_η is periodic, (8.5.10), with $\pi_1 = \pi_2 = \eta^\infty$, does not hold for any J, so that this model does not satisfy (8.5.11).

The above example suggests the following result which we prove in greater generality in Sec. 9.4.

Theorem 8.5.4. Suppose that all stationary policies are unichain and that every optimal policy has an aperiodic transition matrix. Then, for all $v^0 \in V$ and any $\varepsilon > 0$, the sequence of $\{v^n\}$ generated by the value iteration algorithm satisfies (8.5.4) for some finite N.

8.5.3 Bounds on the Gain

The following proposition provides bounds on the optimal gain and ensures that, when value iteration stops, the policy derived through (8.5.5) is ε-optimal. Recall that δ is v-improving if $\delta \in \arg\max_{d \in D}\{r_d + P_d v\}$.

Theorem 8.5.5. Suppose the hypotheses of Theorem 8.4.5 hold, then for $v \in V$,

$$\min_{s \in S}[Lv(s) - v(s)] \le g^{d^\infty} \le g^* \le \max_{s \in S}[Lv(s) - v(s)], \qquad (8.5.12)$$

where d is any v-improving decision rule.

Proof. For any v-improving $d \in D$,

$$g^{d^\infty} = P_d^* r_d = P_d^*[r_d + P_d v - v] = P_d^*[Lv - v] \ge \min_{s \in S}[Lv(s) - v(s)].$$

Since $g^{d^\infty} \le g^*$, the two leftmost inequalities in (8.5.12) hold.

By Theorem 8.4.5, when S and A_s are finite, rewards are bounded, and the model is unichain, there exists a deterministic stationary optimal policy δ^∞, so that $g^{\delta^\infty} = g^*$. Therefore

$$g^* = g^{\delta^\infty} = P_\delta^* r_\delta = P_\delta^*[r_\delta + P_\delta v - v] \le P_d^*[Lv - v] \le \max_{s \in S}[Lv(s) - v(s)],$$

establishing the result. □

We apply the above bounds to value iteration and obtain an approximation to g^*.

Theorem 8.5.6. Suppose (8.5.4) holds and d_ε satisfies (8.5.5).

a. Then $(d_\varepsilon)^\infty$ is an ε-optimal policy.

b. Define

$$g' = \tfrac{1}{2}\left[\max_{s \in S}\left(v^{n+1}(s) - v^n(s)\right) + \min_{s \in S}\left(v^{n+1}(s) - v^n(s)\right)\right]. \qquad (8.5.13)$$

Then $|g' - g^*| < \varepsilon/2$ and $|g' - g^{(d_\varepsilon)^\infty}| < \varepsilon/2$.

Proof. Since $v^{n+1} = Lv^n$, applying Theorem 8.5.5 with $v = v^n$ shows that

$$\varepsilon > sp(v^{n+1} - v^n) \ge g^* - g^{(d_\varepsilon)^\infty},$$

which establishes part (a).

To prove part (b), note that, for scalars x, y, and z, if $x \le y \le z$ and $z - x < \varepsilon$, then

$$-\varepsilon/2 < \tfrac{1}{2}(x - z) \le y - \tfrac{1}{2}(x + z) \le \tfrac{1}{2}(z - x) < \varepsilon/2.$$

Therefore applying (8.5.12) with $v = v^n$ establishes the result. □

Thus under any conditions which ensure that $sp(v^{n+1} - v^n) < \varepsilon$ for some finite n, value iteration identifies an *ε-optimal* policy and an approximation to its value. We summarize this result as follows.

Theorem 8.5.7. Suppose (8.5.10) holds for all π_1 and π_2 in Π^{MD}, then, for any $\varepsilon > 0$, value iteration satisfies (8.5.4) in a finite number of iterations, identifies an optimal policy through (8.5.5), and obtains an approximation to g^* through (8.5.13). Further, for $n = 1, 2, \dots,$

$$sp(v^{nJ+1} - v^{nJ}) \le (\gamma')^n sp(v^1 - v^0). \tag{8.5.14}$$

Note that with the exception of the error bound, the results in Theorem 8.5.7 hold for any model in which g^* is constant, but require different methods of proof. We return to this point in Chap. 9.

Note also that (8.5.14) provides an estimate on the number of iterations required to obtain a specified degree of precision in estimating g^*, and establishes geometric convergence of value iteration, albeit along a subsequence. Note that, when $J = 1$, we have geometric convergence for the entire sequence.

We conclude this section by applying Theorem 8.5.6 to Example 8.5.3.

Example 8.5.3 (ctd.). From Theorem 8.5.6 we conclude that δ^∞ is 0.01-optimal (in fact, it is optimal). Applying (8.5.13), shows that when $n = 11$, $g^{(d_\varepsilon)^\infty} \approx g^* \approx g' = 2.505$. Observe also, that subsequences of $sp(v^{n+1} - v^n)$ converge geometrically, but that for several values of n this quantity is the same at two successive iterates.

8.5.4 An Aperiodicity Transformation

Before applying value iteration, we need to know that the model satisfies conditions which ensure its convergence. This occurs under the hypothesis of Theorem 8.5.2, or any of the conditions in Theorem 8.5.3 which imply it. We show here that, through a simple transformation, all policies can be made aperiodic so that condition (c) in Theorem 8.5.3 holds. *In practice, if there is some reason to believe that a model contains policies with periodic transition probability matrices, apply this transformation.*

Define a transformed MDP with components indicated by " ~ " as follows. Choose τ satisfying $0 < \tau < 1$ and define $\tilde{S} = S$, $\tilde{A}_s = A_s$ for all $s \in S$,

$$\tilde{r}(s,a) = \tau r(s,a) \qquad \text{for } a \in A_s \text{ and } s \in S,$$

and

$$\tilde{p}(j|s,a) = (1 - \tau)\delta(j|s) + \tau p(j|s,a) \qquad \text{for } a \in A_s \text{ and } s \text{ and } j \text{ in } S$$

where $\delta(j|s)$ is 1 if $s = j$, and 0 otherwise. The effect of this data transformation is that, for any decision rule d,

$$\tilde{P}_d = (1 - \tau)I + \tau P_d \qquad \text{and} \qquad \tilde{r}_d = \tau r_d,$$

so that, under it, all transition probability matrices have strictly positive diagonal entries and are aperiodic. We can interpret the transformed model as being generated by a continuous-time system in which the decision maker observes the system state every τ units of time, or a discrete-time model in which, at each decision epoch, the system remains in the current state with probability τ regardless of the decision chosen. The following proposition relates the transformed and untransformed models and applies regardless of chain structure.

Proposition 8.5.8. For any decision rule d,

$$\tilde{P}_d^* = P_d^* \qquad \text{and} \qquad \tilde{g}^{d^\infty} = \tau g^{d^\infty}.$$

Proof. We use Theorem A.5(c) to establish the result. Clearly, P_d and $\tilde{P}_d$ have the same class structure. Since

$$P_d^*\tilde{P}_d = (1 - \tau)P_d^* + \tau P_d^* P_d = P_d^*$$

and, by a similar argument,

$$\tilde{P}_d P_d^* = P_d^*,$$

(A.4) holds, so Theorem A.5(c) implies that $\tilde{P}_d^* = P_d^*$.

The second identity follows by noting that

$$\tilde{g}^{d^\infty} = \tilde{P}_d^* \tilde{r}_d = \tau P_d^* r_d = \tau g^{d^\infty}. \qquad \square$$

The following corollary assures us that we obtain the same optimal policy in either model.

Corollary 8.5.9. The sets of average optimal stationary policies for the original and transformed problem are identical, and $\tilde{g}^* = \tau g^*$.

We illustrate this transformation through the following example.

Example 8.5.1 (ctd.). We form the transformed model

$$\tilde{r}_d = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \tilde{P}_d = \begin{bmatrix} 1 - \tau & \tau \\ \tau & 1 - \tau \end{bmatrix}.$$

The delta coefficient of $\bar{P}_d$ equals $|1 - 2\tau|$, so when $0 < \tau < 1$, Proposition 6.6.1 implies that

$$sp(v^{n-1} - v^n) \leq |1 - \tau| sp(v^n - v^{n-1})$$

and value iteration converges with respect to the span seminorm. Note that in models with more than two states, even after applying the aperiodicity transformation, γ defined by (8.5.7) may equal 1.

8.5.5 Relative Value Iteration

Even when a model satisfies (8.5.11) or has been transformed by the approach of the preceding section, convergence of value iteration can be quite slow because γ' is close to 1. Since v^n diverges linearly in n, this might cause numerical instability.

The following relative value iteration algorithm avoids this difficulty but does not enhance the rate of convergence with respect to the span seminorm.

Relative Value Iteration Algorithm

1. Select $u^0 \in V$, choose $s^* \in S$, specify $\varepsilon > 0$, set $w^0 = u^0 - u^0(s^*)e$, and set $n = 0$.
2. Set $u^{n+1} = Lw^n$ and $w^{n+1} = u^{n+1} - u^{n+1}(s^*)e$.
3. If $sp(u^{n+1} - u^n) < \varepsilon$, go to step 4. Otherwise increment n by 1 and return to step 2.
4. Choose $d_\varepsilon \in \arg\max_{d \in D}\{r_d + P_d u^n\}$.

This algorithm renormalizes u^n at each iteration by subtracting $u^n(s^*)$ from each component. This avoids the pointwise divergence of v^n of ordinary value iteration but does not effect the sequence of maximizing actions or the value of $sp\,(u^{n+1} - u^n)$. To see this, we apply relative value iteration to Example 8.5.2.

Example 8.5.2 (ctd.). Choosing $s^* = s_2$ implies that $w^n(s_2) = 0$, $w^n(s_1) = 11$, $u^n(s_2) = -1$, and $u^n(s_1) = 10$ for all n. Consequently $sp(v^{n+1} - v^n) = sp(u^{n+1} - u^n) = sp(w^{n+1} - w^n) = 0$. Observe also that w^n converges to (in this case equals) an h which satisfies $B(g^*, h) = 0$.

The above example also shows that relative value iteration can be used to directly obtain an h which solves the optimality equation. Note, however, that by using value it ration we can obtain the same approximate solution by choosing $h(s) = v^n(s) - v^n(s^*)$ for n sufficiently large. Also we may vary s^* with n.

8.5.6 Action Elimination

In discounted models we combined bounds on v_λ^* with inequality (6.7.2) to obtain a procedure for identifying suboptimal actions. In average reward models we cannot directly use that approach; however, by using related methods, we can derive a useful one-step-ahead action elimination procedure. The following result is the average reward analog of Proposition 6.7.1.

Proposition 8.5.10. Suppose $B(g^*, h) = 0$ and

$$r(s,a') - g^* + \sum_{j \in S} p(j|s,a')h(j) - h(s) < 0. \tag{8.5.15}$$

Then any stationary policy which uses action a' in state s cannot be optimal.

To apply (8.5.15) to eliminate actions requires bounds for both g^* and h. Section 8.5.4 provides bounds on g^*, but no easily computable bounds are available for h, so we abandon this approach. Instead we provide a method for identifying actions which need not be evaluated when computing $v^{n+1}(s)$.

Define

$$L(s,a)v \equiv r(s,a) + \sum_{j \in S} p(j|s,a)v(j),$$

and recall that, for $v \in V$, $\Lambda(v) = \min_{s \in S} v(s)$ and $\Psi(v) = \max_{s \in S} v(s)$.

Proposition 8.5.11. Suppose v^n is known and, for $a' \in A_s$, there exist functions $u(s,a')$ and $l(s)$ for which

$$u(s,a') \geq L(s,a')v^n, \tag{8.5.16}$$

$$l(s) \leq Lv^n(s), \tag{8.5.17}$$

and

$$l(s) > u(s,a'). \tag{8.5.18}$$

Then

$$a' \notin \operatorname*{arg\,max}_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a)v^n(j) \right\}$$

Proof. By hypothesis,

$$L(s,a')v^n \leq u(s,a') < l(s) \leq Lv^n(s),$$

which implies the result. □

To use this result we require functions l and u, which satisfy (8.5.16)–(8.5.18) and may be evaluated with *little* effort *prior* to computing $L(s,a)v^n$ for any $a \in A_s$. The trick in developing an implementable procedure is for each state-action pair to store the value of $L(s,a)v^{n-k}$ the last time it was evaluated, and develop upper and lower bounds based on it. The following lemma provides such bounds.

Lemma 8.5.12. For $k = 1, 2, \ldots, n$,

$$\begin{aligned} L(s,a)v^n &\le L(s,a)v^{n-k} + \Psi(v^n - v^{n-k}) \\ &\le L(s,a)v^{n-k} + \sum_{j=0}^{k-1} \Psi(v^{n-j} - v^{n-j-1}), \end{aligned} \tag{8.5.19}$$

and

$$\begin{aligned} Lv^n(s) &\ge v^n(s) + \Lambda(v^n - v^{n-k}) \\ &\ge v^n(s) + \sum_{j=0}^{k-1} \Lambda(v^{n-j} - v^{n-j-1}). \end{aligned} \tag{8.5.20}$$

Proof. We derive (8.5.19); a derivation of (8.5.20) uses similar methods and is left as an exercise. We have

$$\begin{aligned} L(s,a)v^n(s) &= r(s,a) + \sum_{j \in S} p(j|s,a)v^n(j) + \sum_{j \in S} p(j|s,a)v^{n-k}(j) \\ &\quad - \sum_{j \in S} p(j|s,a)v^{n-k}(j) \\ &= r(s,a) + \sum_{j \in S} p(j|s,a)v^{n-k}(j) + \sum_{j \in S} p(j|s,a)\left[v^n(j) - v^{n-k}(j)\right] \\ &\le L(s,a)v^{n-k} + \sum_{j \in S} p(j|s,a)\Psi(v^n - v^{n-k}). \end{aligned} \tag{8.5.21}$$

The first inequality in (8.5.19) follows immediately from (8.5.21) while the second inequality follows by writing

$$\left[v^n(j) - v^{n-k}(j)\right] = \left[v^n(j) - v^{n-1}(j)\right] + \cdots + \left[v^{n-k+1}(j) - v^{n-k}(j)\right]$$

prior to taking the maximum in (8.5.21). □

We combine the two results above to obtain inequalities that can be used for action elimination.

Theorem 8.5.13. Suppose, for $a' \in A_s$ and some $k = 1, 2, \ldots, n$, that either

$$v^n(s) - L(s,a')v^{n-k} - \sum_{j=0}^{k-1} sp(v^{n-j} - v^{n-j-1}) > 0 \tag{8.5.22}$$

or

$$v^n(s) - L(s,a')v^{n-k} - sp(v^n - v^{n-k}) > 0, \tag{8.5.23}$$

then

$$a' \notin \underset{a \in A_s}{\arg\max} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a)v^n(j) \right\}. \tag{8.5.24}$$

Note that we can obtain sharper bounds if $\gamma < 1$, where γ is defined in (8.5.7).

A value iteration algorithm which includes the one-step action elimination procedure based on (8.5.22) follows. In it, $E_n(s)$ denotes the set of actions which need not be evaluated at iteration n because they satisfy (8.5.24).

Undiscounted Value Iteration with Action Elimination.

1. Select $v^0 \in V$, $\varepsilon > 0$ and set $n = 0$, $E_0(s) = \varnothing$ for $s \in S$ and $\Delta(s, a) = 0$ for all $a \in A_s$ and $s \in S$.
2. Set

$$v^{n+1}(s) = \max_{a \in A_s/E_n(s)} L(s,a)v^n. \tag{8.5.25}$$

3. If

$$sp(v^{n+1} - v^n) < \varepsilon,$$

go to step 6. Otherwise continue.

4. (Update the elimination criterion) For $a \in A_s/E_n(s)$, set

$$\Delta(s,a) = L(s,a)v^n + \{v^{n+1}(s) - v^n(s)\} + sp(v^{n+1} - v^n), \tag{8.5.26}$$

and, for $a \in E_n(s)$, set

$$\Delta(s,a) = \Delta(s,a) + sp(v^{n+1} - v^n).$$

5. (Action elimination) Replace $n + 1$ by n. For each $a \in A_s$ and $s \in S$, if

$$v^n(s) - \Delta(s,a') > 0,$$

include $a' \in E_n(s)$. Go to step 2.

6. For each $s \in S$, choose $d_\varepsilon(s)$ satisfying

$$d_\varepsilon(s) \in \operatorname*{arg\,max}_{a \in A_s/E_n(s)} L(s,a)v^n.$$

Some comments regarding this algorithm follow. The key feature of the algorithm is that, when computing $v^{n+1}(s)$ in (8.5.25), $L(s, a)$ is evaluated for non-eliminated actions only, namely those in $A_s/E_n(s)$. Since evaluating $L(s, a)$ requires $|S|$ multiplications and $|S|$ additions for each state action pair, this can result in large reductions in computational effort for problems with many actions. Updating the action elimination criteria in Step 4 and checking whether an action should be eliminated requires little extra work because these quantities have already been evaluated in Step 2. Note that $\Delta(s, a)$ stores the negative of the expression in (8.5.22). We update it in (8.5.26) by subtracting $v^n(s)$ and adding $v^{n+1}(s)$.