Introduction to Markov Decision Processes

Martin L. Puterman and Timothy C. Y. Chan

January 17, 2023

Chapter 1

Infinite Horizon Average Expected Reward

"The field of average cost optimization is as strange one. Counterexamples exist to almost all natural conjectures, yet these conjectures are the basis of a proper intuition and are valid if reformulated right or if natural conditions are imposed."
Peter Whittle, Mathematician and statistician, 1927-2021

This chapter focuses on infinite horizon Markov decision processes with the expected average criterion that we refer to as *average reward models*. We prefer the average reward criterion in non-terminating systems, such as controlled queues, in which decisions are made frequently and in perpetuity so that discounting future rewards or costs does not seem prudent.

As the quote above attests to, average reward models poses numerous analytical challenges. This is because the existence and structure of the average reward depends on the *limiting* properties[1] of underlying Markov chains. In contrast to discounted models in which Markov chain properties do not impact results, they are fundamental here. Examples of challenges faced include:

1. For stationary policies with periodic transition probabilities limits may not exist due to oscillation of the powers of transition probability matrices.

2. In finite state models, limits needed to define the average reward for history-dependent and Markovian policies need not exist.

[1]Limiting properties refer to the behavior of large powers of transition probability matrices.

3. In countable state models the limit of transition probabilities may exist but it need not be a transition probability matrix.

4. The form of the Bellman equation depends on whether the average reward is constant or state dependent. This is impacted by the class structure of underlying Markov chains.

In light of these challenges, we will focus primarily on finite state and actions models which are aperiodic and have constant (state-independent) average reward. We will emphasize this point throughout. Moreover we will often refer to the average reward as the *gain*. We encourage the reader to review Appendix **??** for background on Markov chains relevant for Markov decision process.

(We need to add all notation from this chapter to the list of symbols.)
(Do we want a chapter overview?)

1.1 Preliminaries

We make the following assumptions:

Finite state space: S is finite.

Stationary rewards: The (expected) reward function does not vary from epoch to epoch and does not depend on the subsequent state.[2] It will be written as $r(s, a)$, independent of n.

Bounded rewards: There is a finite W such that $|r(s, a)| \leq W$ for all $a \in A_s, s \in S$.

Stationary transition probabilities: The transition probabilities will be written $p(j|s, a)$, independent of n.

In this chapter, we will see that the form of the average reward depends on structural properties of Markov chains generated by stationary policies. Therefore we will spend some effort exploring and illustrating relevant Markov chain concepts. We will also introduce several constructs that are unique to Markov decision processes including *partial Laurent expansions*, *bias vectors* and *relative value vectors*. These are fundamental to the derivation of the Bellman equation and underly the three fundamental algorithms we will focus on: value iteration, policy iteration and linear programming. Analysis and use of Gauss-Seidel iteration, modified policy iteration and action elimination will be left to the reader.

[2] As in Chapter 5, (**??**), when the reward depends on the subsequent state, we take expected rewards under $p(j|s, a)$ prior to analysis.

1.2 The long run average reward or gain

For $\pi \in \Pi^{HR}$ we propose a definition of its *long run average reward* or *gain* to be

$$g^{\pi}(s) := \lim_{N\to\infty} \frac{1}{N} E^{\pi}\left[\sum_{n=1}^{N} r(X_n, Y_n)\middle| X_1 = s\right]. \tag{1.1}$$

We will show that for finite state models the limit in (1.1)

1. always exists and has a closed form representation when π is stationary, and
2. need not exist when π is history dependent or Markovian[3].

Therefore for non-stationary policies a more rigorous definition of the gain is required.

1.2.1 The gain of a stationary policy

For a stationary policy $\pi = d^{\infty}$ where $d \in D^{MR}$, (1.1) becomes

$$g^{d^{\infty}}(s) = \lim_{N\to\infty} \frac{1}{N} E^{d^{\infty}}\left[\sum_{n=1}^{N} r_d(X_n, Y_n)\middle| X_1 = s\right]. \tag{1.2}$$

It can be expressed in matrix notation as

$$\mathbf{g}^{d^{\infty}} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{P}_d^{n-1}\mathbf{r}_d \tag{1.3}$$

where $\mathbf{P}_d^0 = \mathbf{I}$. Doing so enables us to apply part a. of Theorem **??** in Appendix **??** to obtain the following representation of the gain. **(can we number appendix with Section A?)**

Theorem 1.1. Let S be finite and let $\pi = d^{\infty}$ where $d \in D^{MR}$. Then the limit in (1.1) exists and satisfies

$$\mathbf{g}^{d^{\infty}} = \mathbf{P}_d^* \mathbf{r}_d \tag{1.4}$$

where

$$\mathbf{P}_d^* := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{P}_d^{n-1}. \tag{1.5}$$

Note that when the Markov chain corresponding to d^{∞} is regular, $\mathbf{P}_d^n \to \mathbf{P}_d^*$ so that the gain may be also regarded as the *steady state reward*. A fundamental result in Markov chain theory establishes that exponential convergence of $\mathbf{P}_d^n \to \mathbf{P}_d^*$ for regular chains.

[3]Recall that a Markovian policy need not be stationary.

1.2.2 The gain of a non-stationary policy need not exist

The following complicated example shows that the limit in (1.1) need not exist for history dependent π.

Example 1.1. * This simple deterministic finite state example shows that the limit in (1.1) need not exist for non-stationary policies, be they history dependent or Markovian.[a]

Let $S = \{s_1, s_2\}$, $A_{s_i} = \{a_{i,1}, a_{i,2}\}$ for $i = 1, 2$, $p(s_1|s_1, a_{1,1}) = 1$, $p(s_2|s_1, a_{1,2}) = 1$, $p(s_2|s_2, a_{2,1}) = 1$ and $p(s_1|s_2, a_{2,2}) = 1$ and $r(s_1, a_{1,1}) = r(s_1, a_{1,1}) = 1$ and $r(s_2, a_{2,1}) = r(s_2, a_{2,2}) = 0$. (See Figure 1.1).

Consider the history dependent policy π which when starting in state s_1 chooses the sequence of actions

$$a_{1,2}, a_{2,1}, a_{1,2}, a_{2,1}, a_{1,1}, a_{1,2}, a_{2,2}, a_{2,1}, a_{1,1}, a_{1,1}, a_{1,1}, a_{1,2}, a_{2,2}, a_{2,2}, a_{2,2}, a_{2,1}, \ldots$$

and generates the sequence of rewards

$$1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, \ldots$$

When the system starts in state s_2, it chooses the actions

$$a_{2,2}, a_{2,1}, a_{1,2}, a_{2,1}, a_{1,1}, a_{1,2}, a_{2,2}, a_{2,1}, a_{1,1}, a_{1,1}, a_{1,1}, a_{1,2}, a_{2,2}, a_{2,2}, a_{2,2}, a_{2,1}, \ldots$$

that are identical to that in s_1 after the first decision epoch. For states not visited by this deterministic policy, action choice is arbitrary.

A formal analysis of this policy in Appendix 1.1 shows that there exist subsequences of its cumulative averages with different limits[b]. Hence the limit (1.1) **does not** exist for this history dependent policy.

As a consequence of Lemma **??** for s_1 and s_2 there exist Markovian deterministic policies which have the same state-action probabilities and generate the same sequence of rewards after the first decision epoch. Moreover (see Example (**??**)), after defining $d_1(s_1) = a_{1,2}$ and $d_1(s_2) = a_{2,2}$ the Markovian policies are identical. Hence we we have constructed a Markovian deterministic policy for which the limit in (1.1) does not exist.

Now we consider stationary policies. Simple calculations produce the results in the following table.

action in s_1	action in s_2	gain in s_1	gain in s_2
$a_{1,1}$	$a_{2,1}$	1	0
$a_{1,1}$	$a_{2,2}$	1	1
$a_{1,2}$	$a_{2,1}$	0	0
$a_{1,2}$	$a_{2,2}$	1/2	1/2

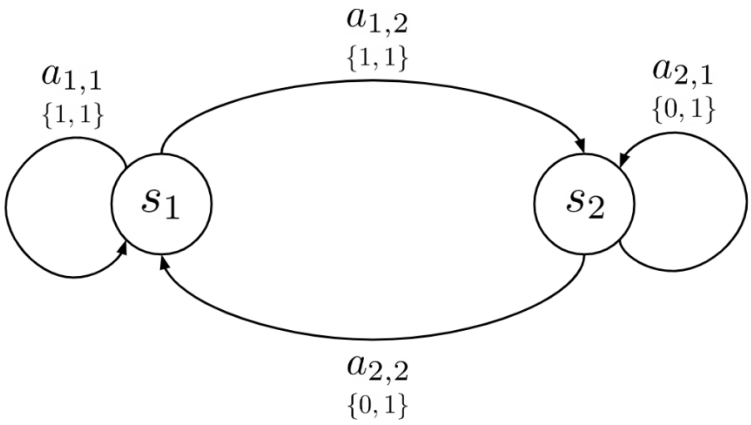

Figure 1.1: Symbolic representation of the model in Example 1.1

Since the largest one-step reward is 1, the stationary policy, $(d^*)^\infty$ that uses $a_{1,1}$ in s_1 and $a_{2,2}$ in s_2 is optimal. Observe also that for every stationary policies, the gain has the same value for both states.

[a] This example was adopted from Section A.4 in Sennott [1999]. Example 8.1.1 in Puterman [1994] provides a model with similar properties.

[b] A sequence converges if every subsequence has the same limit.

A more general definition of the gain

The above example illustrates the need for a more rigorous definition of the gain of a policy. For any $\pi = (d_1, d_2, \ldots) \in \Pi^{HR}$ let

$$v_N^\pi(s) := \left[\sum_{n=1}^{N} r(X_n, Y_n) \middle| X_1 = s\right] = \sum_{n=1}^{N} \mathbf{P}_\pi^{n-1} \mathbf{r}_{d_n}(s) \tag{1.6}$$

where $\mathbf{P}_\pi^0 = \mathbf{I}$ and $\mathbf{P}_\pi^n = \mathbf{P}_{d_1}\mathbf{P}_{d_2}\ldots\mathbf{P}_{d_n}$.

Now we define two quantities $g_-^\pi(s)$ and $g_+^\pi(s)$ which we refer to respectively as the *liminf average reward* and *limsup average reward* as follows:

$$g_-^\pi(s) := \liminf_{N\to\infty} \frac{1}{N} v_N^\pi(s) \quad \text{and} \quad g_+^\pi(s) := \limsup_{N\to\infty} \frac{1}{N} v_N^\pi(s). \tag{1.7}$$

Clearly $g_-^\pi(s) \le g_+^\pi(s)$. The analysis of the above example in Appendix 1.1 shows that this equality may be strict. The limit in (1.1) exists whenever $g_-^\pi(s) = g_+^\pi(s)$ in which case we define the *average reward*

$$g^\pi(s) := \lim_{N\to\infty} \frac{1}{N} v_N^\pi(s). \tag{1.8}$$

As a consequence of Lemma **??**, we have the following important result which simplifies subsequent analyses in this chapter in the same way as Theorem **??** in Chapter **??**. It says that given any history dependent policy, there exists a Markov randomized policy with the same liminf average reward, limsup average reward or average reward.

Theorem 1.2. Let $\pi \in \Pi^{HR}$, then for each $s \in S$, there exists a $\pi' \in \Pi^{MR}$ for which

$$g_-^{\pi'}(s) = g_-^\pi(s) \quad \text{and} \quad g_+^{\pi'}(s) = g_+^\pi(s). \tag{1.9}$$

Furthermore if $g_-^\pi(s) = g_+^\pi(s)$,

$$g^{\pi'}(s) = g^\pi(s). \tag{1.10}$$

1.2.3 Average optimal policies

We would like to be able to write that a policy π^* is average reward optimal whenever

$$g^{\pi^*}(s) \geq g^\pi(s) \tag{1.11}$$

for all $\pi \in \Pi^{\mathrm{HR}}$ and $s \in S$. Unfortunately as we have shown above, the limits implicit in defining the quantities in (1.11) need not exist, even when S is finite. Thus we cannot avoid using the quantities in (1.7) to define optimality. (Note the quote at the beginning of this chapter).

Hence we must be content (at least until we establish the existence of average optimal stationary policies) to define the following rigorous notion of optimality.

A policy π^* is *average reward optimal* if for all $s \in S$

$$g_-^{\pi^*}(s) \geq g_+^\pi(s) \tag{1.12}$$

for all $\pi \in \Pi^{\mathrm{HR}}$.

What this definition means is that a policy is average optimal if its smallest limit point is greater than the largest limit point of any other policy.

We return to Example 1.1 to illustrate this result. As noted in the analysis of that example, there exists a Markovian policy π' for which

$$g_-^{\pi'}(s) = 1/2 \quad \text{and} \quad g_+^{\pi'}(s) = 2/3$$

for $s \in S$. Moreover $g^{d^\infty}(s) = g_+^{d^\infty}(s) = g_-^{d^\infty}(s)$ for all stationary policies so that for the policy $(d^*)^\infty$ distinguished in Example 1.1

$$g_-^{(d^*)^\infty}(s) \geq g_+^{\pi'}(s)$$

for all $s \in S$ confirming the optimality of d^* under the more general definition.

Optimal value functions

To correspond to the above concepts of optimality we define the following optimal values for all $s \in S$:

$$g_-^*(s) = \sup_{\pi \in \Pi^{HR}} g_-^\pi(s) \quad \text{and} \quad g_+^*(s) = \sup_{\pi \in \Pi^{HR}} g_+^\pi(s) \tag{1.13}$$

and when $g_-^*(s) = g_+^*(s)$,

$$g^*(s) = \sup_{\pi \in \Pi^{HR}} g^\pi(s). \tag{1.14}$$

Hence in the latter case, a policy π is average or gain optimal whenever $g^\pi(s) = g^*(s)$ for all $s \in S$.

1.3 Bias of a stationary policy

In order to develop efficient algorithms for find average optimal policies, we introduce an additional quantity that we refer to as the *bias*. For a stationary randomized policy d^∞ define its bias by

$$h^{d^\infty}(s) := E\left\{ \sum_{n=1}^{\infty} \left[r_d(X_n) - g^{d^\infty}(X_n) \right] \middle| X_1 = s \right\} \tag{1.15}$$

When the limit above does not exist (i.e some closed class of the Markov chain generated by d^∞ is periodic) we define the bias by

$$h^{d^\infty}(s) := \lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} E\left\{ \sum_{n=1}^{j} \left[r_d(X_n) - g^{d^\infty}(X_n) \right] \middle| X_1 = s \right\} \tag{1.16}$$

When the limit in (1.15) exists the two definitions are equivalent however when the limit does not exist, we require the second definition.

Equivalently we can express the bias of a stationary policy in matrix-vector notation as

$$\mathbf{h}^{d^\infty} = \sum_{n=1}^{\infty} \left[\mathbf{P}_d^{n-1} \mathbf{r}_d - \mathbf{g}^{d^\infty} \right] = \sum_{n=1}^{\infty} \left[\mathbf{P}_d^{n-1} - \mathbf{P}^* \right] \mathbf{r}_d \tag{1.17}$$

where the second inequality follows from (1.4). When the limit in (1.17) does not exist, the equivalent matrix representation is

$$\mathbf{h}^{d^\infty} = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \sum_{n=1}^{j} \left[\mathbf{P}_d^{n-1}\mathbf{r}_d - \mathbf{g}^{d^\infty}\right] = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \sum_{n=1}^{j} \left[\mathbf{P}_d^{n-1} - \mathbf{P}^*\right] \mathbf{r}_d. \tag{1.18}$$

As a consequence of Proposition **??** in Appendix **??**, we have the following closed form representation for $\mathbf{h}^{d^\infty}$.

Theorem 1.3. Let $d \in D^{MR}$. Then

$$\mathbf{h}^{d^\infty} = \mathbf{H}_d \mathbf{r}_d \tag{1.19}$$

where $\mathbf{H}_d := \mathbf{H}_{P_d} = (\mathbf{I} - (\mathbf{P}_d - \mathbf{P}_d^*))^{-1}(\mathbf{I} - \mathbf{P}_d^*)$.

Although Theorems 1.1 and 1.3 provide closed form representations for $\mathbf{g}^{d^\infty}$ and $\mathbf{h}^{d^\infty}$, we will not use them for computation except in simple examples. Instead we will derive a system equations which can be solved to find these quantities. We emphasize that we have introduced the shorthand notation $\mathbf{H}_d$ in the above theorem to avoid the cumbersome notation $\mathbf{H}_{P_d}$.

1.3.1 Interpreting the bias

First we use the definition (1.15) directly. The nth term in the sum, $r_d(X_n) - g^{d^\infty}(X_n)$, represents the difference in the reward received in period n and the gain. Hence the bias represents the expected total sum of these differences.

We now appeal to the matrix representation for the bias in (1.17). When $\mathbf{P}_d$ is regular, part f. of Theorem **??** implies that $\mathbf{P}_d - \mathbf{P}_d^*$ converges to $\mathbf{0}$ exponentially fast so that terms in the sum become very small quickly. Hence the bias can be regarded as the total *initial* difference between the expected reward and the steady state reward. For this reason, the bias is sometimes referred to as the *transient reward*.

Example 1.2. We consider the model in Example **??** and analyze the stationary policy d^∞ where $d(s_1) = a_{1,1}$ and $d(s_2) = a_{2,2}$. For this policy

$$\mathbf{P}_d = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \quad \text{and} \quad \mathbf{r}_d = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

Since $\mathbf{P}_d$ is regular, $\mathbf{P}^n \to \mathbf{P}_d^*$. Define the largest component difference between $\mathbf{P}^n$ and $\mathbf{P}_d^*$ by

$$\Delta_n := \max_{(s,j)\in S\times S} |\mathbf{P}_d^n(j|s) - \mathbf{P}_d^*(j|s)|.$$

We see that $\Delta_2 = 0.1067$, $\Delta_3 = 0.0427$, $\Delta_4 = 0.0171\,, \ldots$ indicative of the exponential convergence of $\mathbf{P}_d^n$ to $\mathbf{P}_d^*$. Because the eigenvalues of $\mathbf{P}_d$ are 1 and 0.4 we observe exponential convergence at the rate of the second eigenvalue as noted in the Markov chain appendix.

Direct calculation shows that

$$\mathbf{P}_d^* = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix} \quad \text{and} \quad \mathbf{g}^{d^\infty} = \begin{bmatrix} 2\frac{2}{3} \\ \\ 2\frac{2}{3} \end{bmatrix}.$$

Observe that the rows of $\mathbf{P}_d^*$ are equal an so are the components of $\mathbf{g}^{d^\infty}$.

We now evaluate the bias of d^∞. Direct calculation shows that

$$\mathbf{H}_d = \begin{bmatrix} \frac{5}{9} & -\frac{5}{9} \\ \\ -\frac{10}{9} & \frac{10}{9} \end{bmatrix} \quad \text{and} \quad \mathbf{h}^{d^\infty} = \mathbf{H}_d \mathbf{r}_d = \begin{bmatrix} \frac{5}{9} \\ \\ -\frac{10}{9} \end{bmatrix}.$$

Table 1.1 below provides the the first few terms of the partial sums of (1.17)

$$\mathbf{S}_N := \sum_{n=1}^{N} \left[\mathbf{P}_d^{n-1} \mathbf{r}_d - \mathbf{g}^{d^\infty} \right].$$

Continuing in this way we see that $\mathbf{S}_N \to \mathbf{h}^{d^\infty} = (0.5556, -1.1111)^T$.

Observe that starting in s_1 the bias is positive since the expected reward $r_d(s_1)$ exceeds $g^{d^\infty}(s_1)$ and negative when starting in s_2 since $r_d(s_2)$ is less than $g^{d^\infty}(s_2)$.

n	$\mathbf{S}_n^T$
1	(0.33,-0.67)
2	(0.47,-0.93)
3	(0.52,-1.04)

Table 1.1: First three terms of $\mathbf{S}_n$

1.3.2 The vanishing discount rate approach

The most elegant analysis of finite state average reward models uses the relationship between the discounted reward and the average reward. We refer to this analysis as the *vanishing discount rate* approach. It yields the following key result. Its straightforward proof is informative as it illustrates several computations that will be used below.

Theorem 1.4. Let S be finite, $d \in D^{MR}$, and $0 \leq \lambda < 1$. Then

$$\mathbf{v}_\lambda^{d^\infty} = \frac{\mathbf{g}^{d^\infty}}{1-\lambda} + \mathbf{h}^{d^\infty} + \mathbf{f}(\lambda) \tag{1.20}$$

where $\mathbf{h}^{d^\infty} = \mathbf{H}_d \mathbf{r}_d$ and $\mathbf{f}(\lambda)$ is a vector which converges to $\mathbf{0}$ as $\lambda \uparrow 1$.

Proof. From (**??**) the expected total discounted reward of stationary policy d^∞ may be written as

$$\mathbf{v}_\lambda^{d^\infty} = \sum_{n=0}^{\infty} \lambda^n \mathbf{P}_d^n \mathbf{r}_d.$$

Adding and subtracting $\mathbf{P}_d^*$ inside the summation yields:

$$\begin{aligned}
\sum_{n=0}^{\infty} \lambda^n \mathbf{P}_d^n \mathbf{r}_d &= \sum_{n=0}^{\infty} \lambda^n (\mathbf{P}_d^* + \mathbf{P}_d^n - \mathbf{P}_d^*) \mathbf{r}_d \\
&= \sum_{n=0}^{\infty} \lambda^n \mathbf{P}_d^* \mathbf{r}_d + [\mathbf{I} - \mathbf{P}_d^* + \sum_{n=1}^{\infty} \lambda^n (\mathbf{P}_d^n - \mathbf{P}_d^*)] \mathbf{r}_d \\
&= \frac{\mathbf{P}_d^* \mathbf{r}_d}{1-\lambda} + [\sum_{n=0}^{\infty} \lambda^n (\mathbf{P}_d - \mathbf{P}_d^*)^n - \mathbf{P}_d^*] \mathbf{r}_d \qquad (1.21) \\
&= \frac{\mathbf{P}_d^* \mathbf{r}_d}{1-\lambda} + [(\mathbf{I} - \lambda(\mathbf{P}_d - \mathbf{P}_d^*))^{-1} - \mathbf{P}_d^*] \mathbf{r}_d \qquad (1.22) \\
&= \frac{\mathbf{P}_d^* \mathbf{r}_d}{1-\lambda} + \mathbf{H}_d \mathbf{r}_d + \left[[(\mathbf{I} - \lambda(\mathbf{P}_d - \mathbf{P}_d^*))^{-1} - \mathbf{P}_d^*] - \mathbf{H}_d \right] \mathbf{r}_d \qquad (1.23) \\
&= \frac{\mathbf{g}^{d^\infty}}{1-\lambda} + \mathbf{h}^{d^\infty} + \mathbf{f}(\lambda)
\end{aligned}$$

We briefly explain some key steps above. That $\mathbf{P}_d^n - \mathbf{P}_d^* = (\mathbf{P}_d - \mathbf{P}_d^*)^n$ for $n \geq 1$ in (1.21) follows from properties of $\mathbf{P}_d^*$ in part b. of Theorem **??**. Its derivation is left as an exercise. The existence and representation for the inverse in (1.22) for $\lambda \leq 1$ follows from Proposition **??** in the Appendix **??**. Equation (1.23) follows by adding and subtracting $\mathbf{H}_d$ in(1.22) and the final result follows from the representations for $\mathbf{g}^{d^\infty}$ and $\mathbf{h}^{d^\infty}$ in Theorems 1.1 and 1.3.

Clearly

$$\mathbf{f}(\lambda) := \left[[(\mathbf{I} - \lambda(\mathbf{P}_d - \mathbf{P}_d^*))^{-1} - \mathbf{P}_d^*] - \mathbf{H}_d \right]$$

converges to $\mathbf{0}$ completing the proof. □

The expression (1.20) is referred to as the *partial Laurent expansion*. When $\mathbf{f}(\lambda)$ is written as an infinite series, we obtain the *Laurent expansion* of the expected discounted reward. It is the basis for the concept of sensitive discount optimality[4] which will not be discussed in this book..

[4]See Chapter 10 in Puterman [1994]

Note that the above result provides an approximation for the bias as

$$\mathbf{h}^{d^\infty} \approx \mathbf{v}_\lambda^{d^\infty} - \frac{\mathbf{g}^{d^\infty}}{1-\lambda}. \tag{1.24}$$

This means that the bias of a stationary policy approximately equals the difference between its expected discounted reward and a system that accrues the average reward every period.

The following corollary uses the above result to relate the discounted reward to the average reward[5].

Corollary 1.1. Let S be finite. Then $d \in D^{MR}$

$$\lim_{\lambda \uparrow 1}(1-\lambda)\mathbf{v}_\lambda^{d^\infty} = \mathbf{g}^{d^\infty} \tag{1.25}$$

The above proof also provides the matrix relationship

$$\sum_{n=0}^{\infty} \lambda^n \mathbf{P}_d^n = \frac{\mathbf{P}_d^*}{1-\lambda} + \mathbf{H}_d + \mathbf{F}(\lambda) \tag{1.26}$$

where all of the components of the matrix $\mathbf{F}(\lambda)$ converge to $\mathbf{0}$ as $\lambda \uparrow 1$.

The following table continues Example 1.2 by illustrating the calculations implicit in Theorem 1.4 and Corollary 1.1. Observe that approximations are accurate to two decimal places for $\lambda = 0.99$ and three decimal places for $\lambda = 0.999$.

[5]Results of this kind are often called Tauberian theorems because they relate two different ways of summing a divergent infinite series (See Derman [1970] or Sennott [1999]). For a sequence of real numbers $\{a_n\}$ it is analogous to

$$\lim_{\lambda \uparrow 1}(1-\lambda)\sum_{n=0}^{\infty} a_n \lambda^n = \lim_{N \to \infty} \sum_{n=1}^{N} a_n$$

λ	$(1-\lambda)\mathbf{v}_\lambda^{d^\infty}$	$\mathbf{v}_\lambda^{d^\infty} - [(1-\lambda)^{-1}\mathbf{g}^{d^\infty} - \mathbf{h}^\infty]$
0.9	$(2.719, 2.563)^T$	$(-0.0347, 0.0694)^T$
0.95	$(2.694, 2.613)^T$	$(-0.0179, 0.0358)^T$
0.99	$(2.672, 2.656)^T$	$(-0.0037, 0.0074)^T$
0.999	$(2.667, 2.666)^T$	$(-0.0004, 0.0007)^T$

Table 1.2: Illustration of the approximation to $\mathbf{g}^{d^\infty} = (2.667, 2.667)^T$ based on Corollary 1.1 and the accuracy of the partial Laurent series approximation in Theorem 1.4 as a function of λ for policy in Example 1.2.

1.3.3 Relationship to the expected total reward

Now we explore the relationship between the average reward and the expected total reward of a stationary policy.

Theorem 1.5. Let $d \in D^{MR}$. Then

$$\mathbf{v}_N^{d^\infty} = N\mathbf{g}^{d^\infty} + \mathbf{h}^{d^\infty} + \mathbf{o}(N) \tag{1.27}$$

where $\mathbf{o}(N)$ is a vector that converges to $\mathbf{0}$ as $N \to \infty$.

Proof. Adding and subtracting $\mathbf{P}_d^*$ from the matrix representation for the expected total reward over N decision epochs yields

$$\begin{aligned}
\mathbf{v}_N^{d^\infty} &= \sum_{n=1}^{N} \mathbf{P}_d^{n-1}\mathbf{r}_d = \sum_{n=1}^{N} (\mathbf{P}_d^* + \mathbf{P}_d^{n-1} - \mathbf{P}_d^*)\mathbf{r}_d \\
&= N\mathbf{P}_d^*\mathbf{r}_d + \sum_{n=1}^{N} (\mathbf{P}_d^{n-1} - \mathbf{P}_d^*)\mathbf{r}_d \\
&= N\mathbf{P}_d^*\mathbf{r}_d + \sum_{n=1}^{N} (\mathbf{P}_d^{n-1} - \mathbf{P}_d^*)\mathbf{r}_d + \sum_{n=N+1}^{\infty} (\mathbf{P}_d^{n-1} - \mathbf{P}_d^*)\mathbf{r}_d - \sum_{n=N+1}^{\infty} (\mathbf{P}_d^{n-1} - \mathbf{P}_d^*)\mathbf{r}_d \\
&= N\mathbf{P}_d^*\mathbf{r}_d + \mathbf{H}_d\mathbf{r}_d - [\sum_{n=N+1}^{\infty} (\mathbf{P}_d^{n-1} - \mathbf{P}_d^*)]\mathbf{r}_d
\end{aligned} \tag{1.28}$$

where the last expression in (1.28) is $\mathbf{o}(N)$ because $\sum_{n=1}^{N}(\mathbf{P}_d^{n-1} - \mathbf{P}_d^*)$ converge to $\mathbf{H}_d$ as a consequence of Proposition **??**. □

This result means that for each $s \in S$, $v_N^{d^\infty}$ is *approximately linear* in N with slope $g^{d^\infty}(s)$ and intercept $h^{d^\infty}(s)$. Furthermore this theorem provides another interpretation for the bias as

$$\mathbf{h}^{d^\infty} \approx \mathbf{v}_N^{d^\infty} - N\mathbf{g}^{d^\infty}. \tag{1.29}$$

As an immediate consequence of Theorem 1.5 we have:

Corollary 1.2. Let $d \in D^{MR}$. Then

$$\mathbf{g}^{d^\infty} = \lim_{N\to\infty} \frac{1}{N}\mathbf{v}_N^{d^\infty}. \tag{1.30}$$

Observe that Theorem 1.27 also suggests the following approximation

$$\mathbf{g}^{d^\infty} \approx \mathbf{v}_N^{d^\infty} - \mathbf{v}_{N-1}^{d^\infty}. \tag{1.31}$$

We will find this approximation very useful when we analyze value iteration.

We again illustrate these results in the context of the decision rule analyzed in Example 1.2. Table 1.3 shows that $\frac{1}{N}\mathbf{v}_N^{d^\infty}$ does not accurately approximate $\mathbf{g}^{d^\infty}$ but $\mathbf{v}_N^{d^\infty} - \mathbf{v}_{N-1}^{d^\infty}$ does. In fact the approximation is accurate to three decimal places when $N = 11$. Note also that the approximation of the expected total reward based on (1.27) is very accurate even for small N.

N	$\frac{1}{N}\mathbf{v}_N^{d^\infty}$	$\mathbf{v}_N^{d^\infty} - \mathbf{v}_{N-1}^{d^\infty}$	$\mathbf{v}_N^{d^\infty} - [N\mathbf{g}^{d^\infty} - \mathbf{h}^\infty]$
5	$(2.777, 2.447)^T$	$(2.675, 2.644)^T$	$(-0.006, 0.011)^T$
10	$(2.722, 2.556)^T$	$(2.667, 2.666)^T$	$(-0.0001, 0.0001)^T$
20	$(2.694, 2.611)^T$	$(2.667, 2.667)^T$	*
50	$(2.678, 2.644)^T$	$(2.667, 2.667)^T$	*

Table 1.3: Illustration of the approximation to $\mathbf{g}^{d^\infty} = (2.667, 2.667)^T$ based on Corollary 1.2 and (1.31). Column 4 shows the accuracy of the approximation in Theorem 1.5 as a function of N for the policy in Example 1.2. The entries denoted by * are less than 10^{-8}.

1.3.4 Computing the gain and bias of a stationary policy

In this section we show how to obtain the gain and bias without computing $\mathbf{P}_d^*$ and $\mathbf{H}_d$. We derive a system of equations that when solved yields the gain and bias and moreover can be generalized to obtain the appropriate Bellman equations. We motivate them by appealing to the partial Laurent expansion of the expected total discounted reward and the approximation to the finite horizon expected reward that appeared in the two preceding sections. We will show that the gain and bias satisfy the following system of matrix equations[6]:

$$\mathbf{g} = \mathbf{P}_d\mathbf{g} \quad \text{and} \quad \mathbf{h} = \mathbf{r}_d - \mathbf{g} + \mathbf{P}_d\mathbf{h} \tag{1.32}$$

which sometimes is written as

$$(\mathbf{I} - \mathbf{P}_d)\mathbf{g} = \mathbf{0} \quad \text{and} \quad \mathbf{g} + (\mathbf{I} - \mathbf{P}_d)\mathbf{h} = \mathbf{r}_d.$$

The first representation above expressed in component notation becomes:

$$g(s) = \sum_{s\in S} p(j|s)g(j) \quad \text{and} \quad h(s) = r_d(s) - g(s) + \sum_{j\in S} p_d(j|s)h(j). \tag{1.33}$$

Heuristic derivation based on the partial Laurent series approximation

As a result of (**??**), the expected total discounted reward of stationary policy d^∞, $\mathbf{v}_\lambda^{d^\infty}$ satisfies

$$\mathbf{v} = \mathbf{r}_d + \lambda\mathbf{P}_d\mathbf{v}.$$

Substituting (1.20) into both sides of this equation yields

$$\begin{aligned}\frac{\mathbf{g}^{d^\infty}}{1-\lambda} + \mathbf{h}^{d^\infty} + \mathbf{f}(\lambda) &= \mathbf{r}_d + \lambda\mathbf{P}_d\left[\frac{\mathbf{g}^{d^\infty}}{1-\lambda} + \mathbf{h}^{d^\infty} + \mathbf{f}(\lambda)\right] \\ &= \mathbf{r}_d + \frac{\mathbf{P}_d\mathbf{g}^{d^\infty}}{1-\lambda} - \mathbf{P}_d\mathbf{g}^{d^\infty} + \lambda\mathbf{P}_d\mathbf{h}^{d^\infty} + \mathbf{f}(\lambda)\end{aligned}$$

where the last equality follows by noting $\frac{\lambda}{1-\lambda} = \frac{1}{1-\lambda} - 1$. For this last relationship to be valid in the limit as $\lambda \uparrow 1$ requires that $\mathbf{g}^{d^\infty} = \mathbf{P}_d\mathbf{g}^{d^\infty}$. When this is the case[7]

$$\begin{aligned}\mathbf{h}^{d^\infty} &= \mathbf{r}_d - \mathbf{P}_d\mathbf{g}^{d^\infty} + \lambda\mathbf{P}_d\mathbf{h}^{d^\infty} + \mathbf{f}(\lambda) \\ &= \mathbf{r}_d - \mathbf{g}^{d^\infty} + \lambda\mathbf{P}_d\mathbf{h}^{d^\infty} + \mathbf{f}(\lambda).\end{aligned}$$

[6]Many of you may be familiar with this system without the first equation. This is because in most models in the literature, $\mathbf{g}^{d^\infty}$ is a constant vector so this equation becomes redundant. We will discuss this issue below.

[7]We don't care about the precise form of $\mathbf{f}(\lambda)$, only that it converges to 0 as $\lambda \uparrow 1$ so analogously to "small o notation" we write $\mathbf{f}(\lambda)$ again for $\mathbf{f}(\lambda) - \lambda\mathbf{P}_d\mathbf{f}(\lambda)$.

Letting $\lambda \uparrow 1$ shows that $\mathbf{h}^{d^\infty} = \mathbf{r}_d - \mathbf{g}^{d^\infty} + \mathbf{P}_d \mathbf{h}^{d^\infty}$. Hence $\mathbf{g}^{d^\infty}$ and $\mathbf{h}^{d^\infty}$ must be solutions of (1.32).

Heuristic derivation based on the expected total reward approximation

From (**??**), the expected total reward of stationary policy d^∞, $\mathbf{v}_n^{d^\infty}$ satisfied

$$\mathbf{v}_{n+1} = \mathbf{r}_d + \mathbf{P}_d \mathbf{v}_n$$

for $n = 0, 1, \ldots$. Substituting (1.27) into both sides of this equation yields

$$(n+1)\mathbf{g}^{d^\infty} + \mathbf{h}^{d^\infty} + \mathbf{o}(n) = \mathbf{r}_d + \mathbf{P}_d \left[n\mathbf{g}^{d^\infty} + \mathbf{h}^{d^\infty} + \mathbf{o}(n) \right]$$

so that

$$n\mathbf{g}^{d^\infty} + \mathbf{h}^{d^\infty} = \mathbf{r}_d - \mathbf{g}^{d^\infty} + n\mathbf{P}_d \mathbf{g}^{d^\infty} + \mathbf{P}_d \mathbf{h}^{d^\infty} + \mathbf{o}(n).$$

For this to hold for all $n = 0, 1, \ldots$ requires that the expressions multiplied by n be equal. Hence $\mathbf{g}^{d^\infty} = \mathbf{P}_d \mathbf{g}^{d^\infty}$. If this holds it follows that $\mathbf{h}^{d^\infty} = \mathbf{r}_d - \mathbf{g}^{d^\infty} + \mathbf{P}_d \mathbf{h}^{d^\infty}$. Consequently $\mathbf{g}^{d^\infty}$ and $\mathbf{h}^{d^\infty}$ must be solutions of (1.32).

Making this formal

We now describe some important properties of the system of equations (1.32). We begin with an analysis of Example 1.2 to illustrate the subtleties involved in solving this system of equations.

Example 1.2 (ctd.) For the decision rule d^∞ analyzed above, (1.32) becomes

$$\begin{bmatrix} g(s_1) \\ g(s_2) \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} g(s_1) \\ g(s_2) \end{bmatrix}$$

and

$$\begin{bmatrix} h(s_1) \\ h(s_2) \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} - \begin{bmatrix} g(s_1) \\ g(s_2) \end{bmatrix} + \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} h(s_1) \\ h(s_2) \end{bmatrix}.$$

Expressing this system of equations as

$$(\mathbf{I} - \mathbf{P}_d)\mathbf{g} = \mathbf{0} \quad \text{and} \quad \mathbf{I}\mathbf{g} + (\mathbf{I} - \mathbf{P}_d)\mathbf{h} = \mathbf{r}_d \tag{1.34}$$

we can write them in terms of the combined matrix $\mathbf{B}$ implicitly defined by

$$\mathbf{B} \begin{bmatrix} \mathbf{g} \\ \mathbf{h} \end{bmatrix} := \begin{bmatrix} (\mathbf{I} - \mathbf{P}_d) & \mathbf{0} \\ \mathbf{I} & (\mathbf{I} - \mathbf{P}_d) \end{bmatrix} \begin{bmatrix} \mathbf{g} \\ \mathbf{h} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{r}_d \end{bmatrix}. \tag{1.35}$$

Substituting appropriate values, this matrix equation becomes

$$\begin{bmatrix} 0.2 & -0.2 & 0 & 0 \\ -0.4 & 0.4 & 0 & 0 \\ 1 & 0 & 0.2 & -0.2 \\ 0 & 1 & -0.4 & 0.4 \end{bmatrix} \begin{bmatrix} g(s_1) \\ g(s_2) \\ h(s_1) \\ h(s_2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 3 \\ 2 \end{bmatrix}. \tag{1.36}$$

Applying Gauss-Jordan elimination[a] reduces this equation to

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} g(s_1) \\ g(s_2) \\ h(s_1) \\ h(s_2) \end{bmatrix} = \begin{bmatrix} 2\frac{2}{3} \\ 0 \\ 2\frac{2}{3} \\ 1\frac{2}{3} \end{bmatrix}. \tag{1.37}$$

Hence $g(s_1) = g(s_2) = 2\frac{2}{3}$ and $h(s_1) - h(s_2) = 1\frac{2}{3}$ so from Example 1.2 we see that $\mathbf{g} = \mathbf{g}^{d^\infty}$. Moreover for arbitrary constant c,

$$\mathbf{h} = \begin{bmatrix} 1\frac{2}{3} \\ 0 \end{bmatrix} + c \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

where the vector $(1\frac{2}{3}, 0)^T$ corresponds to the the solution when $h(s_2) = 0$. So again from Example 1.2 we see that when $c = -\frac{10}{9}$, $\mathbf{h} = \mathbf{h}^{d^\infty}$.

[a] Some might refer to this as Gaussian elimination.

This example illustrates the following key features of the system of equations (1.32):

1. When $\mathbf{g}$ satisfies these equations, $\mathbf{g} = \mathbf{g}^{d^\infty}$.

2. These equations do not uniquely determine $\mathbf{h}$. As we see directly from (1.36) or from the observation that the second row of the reduced matrix (1.37) contains all zero entries, $\mathbf{B}$ is not of full rank[8].

3. The matrix $\mathbf{I} - \mathbf{P}_d$ plays a key role in finding $\mathbf{g}^{d^\infty}$ and $\mathbf{h}^{d^\infty}$. When $\mathbf{x}$ solves $\mathbf{P}_d\mathbf{x} = \mathbf{I}\mathbf{x}$, it is a right eigenvector of $\mathbf{P}_d$ corresponding to the eigenvalue 1. The eigenvalues of $\mathbf{P}_d$ are 1 and 0.4 and the (right) eigenvector corresponding to eigenvalue 1 is $(1, 1)^T$.

We formalize these observations in the following important theorem.

Theorem 1.6. Let $d \in D^{MR}$.
a. Then $\mathbf{g}^{d^\infty}$ and $\mathbf{h}^{d^\infty}$ are solutions of (1.34).
b. Suppose $\mathbf{g}$ and $\mathbf{h}$ are solutions of (1.34). Then $\mathbf{g} = \mathbf{g}^{d^\infty}$ and $\mathbf{h} = \mathbf{h}^{d^\infty} + \mathbf{u}$ where $(\mathbf{I} - \mathbf{P}_d)\mathbf{u} = 0$.
c. Suppose $\mathbf{g}$ and $\mathbf{h}$ are solutions of (1.34) and $\mathbf{P}_d^*\mathbf{h} = \mathbf{0}$. Then $\mathbf{g} = \mathbf{g}^{d^\infty}$ and $\mathbf{h} = \mathbf{h}^{d^\infty}$.

Proof. To establish that $\mathbf{g}^{d^\infty}$ satisfies the first equation in in (1.32) we have that

$$(\mathbf{I} - \mathbf{P}_d)\mathbf{g}^{d^\infty} = (\mathbf{I} - \mathbf{P}_d)\mathbf{P}_d^*\mathbf{r}_d = \mathbf{0}$$

where we use the result that $\mathbf{P}_d^* = \mathbf{P}_d^*\mathbf{P}_d$ to obtain the last equality.. To establish the second equality in (1.32), note that

$$\begin{aligned}\mathbf{g}^{d^\infty} + (\mathbf{I} - \mathbf{P}_d)\mathbf{h}^{d^\infty} &= \mathbf{P}_d^*\mathbf{r}_d + (\mathbf{I} - \mathbf{P}_d)\mathbf{H}_d\mathbf{r}_d \\ &= \mathbf{P}_d^*\mathbf{r}_d + (\mathbf{I} - \mathbf{P}_d^*)\mathbf{r}_d = \mathbf{r}_d\end{aligned}$$

where the next to last equality follows from equation (**??**) in Appendix **??** that $(\mathbf{I} - \mathbf{P}_d)\mathbf{H}_d = \mathbf{I} - \mathbf{P}_d^*$. Hence part a. follows.

Now we prove part b. Let $(\mathbf{g}_1, \mathbf{h}_1)^T$ and $(\mathbf{g}_2, \mathbf{h}_2)^T$ denote two solutions of (1.34) and $\Delta\mathbf{g} := \mathbf{g}_1 - \mathbf{g}_2$ and $\Delta\mathbf{h} := \mathbf{h}_1 - \mathbf{h}_2$. Then it's easy to see that

$$\Delta\mathbf{g} = \mathbf{P}_d\Delta\mathbf{g} \quad \text{and} \quad \Delta\mathbf{g} = (\mathbf{I} - \mathbf{P}_d)\Delta\mathbf{h}. \tag{1.38}$$

Multiplying both sides of the second equation by $\mathbf{P}_d$ and noting that the first equation gives

$$\Delta\mathbf{g} = \mathbf{P}_d\Delta\mathbf{g} = (\mathbf{P}_d^2 - \mathbf{P}_d)\Delta\mathbf{h}.$$

Repeating this argument n times we see that for all $n \geq 0$,

$$\Delta\mathbf{g} = (\mathbf{P}_d^{n+1} - \mathbf{P}_d^n)\Delta\mathbf{h}.$$

[8]In the example $\mathbf{B}$ has 3 independent rows and one redundant row and $\mathbf{I} - \mathbf{P}_d$ has one redundant row.

Adding all these expressions for $\mathbf{g}$ together yields

$$(n+1)\mathbf{g} = (\mathbf{P}_d^{n+1} - \mathbf{I})\mathbf{h}.$$

Now divide both sides by $n+1$ and take the limit as $n \to \infty$ yields

$$\Delta\mathbf{g} = \lim_{n\to\infty} \frac{1}{n+1}(\mathbf{P}_d^{n+1} - \mathbf{I})\Delta\mathbf{h} = \mathbf{0}.$$

Hence $\mathbf{g}_1 = \mathbf{g}_2 = \mathbf{g}^{d^\infty}$, the last equality following from the first part of the proof.

Substituting $\Delta\mathbf{g} = \mathbf{0}$ into the second equation in (1.38), shows that $(\mathbf{I}-\mathbf{P}_d)\Delta\mathbf{h} = \mathbf{0}$. Part a. showed that $\mathbf{h}^{d^\infty}$ is a solution of the second equation when $\mathbf{g} = \mathbf{g}^{d^\infty}$. So setting $\mathbf{h}_2 = \mathbf{h}^{d^\infty}$ implies $\mathbf{h}_1 = \mathbf{h}^{d^\infty} + \Delta\mathbf{h}$ which establishes part b.

The easiest way to establish part c. is to add $\mathbf{P}_d^*\mathbf{h} = \mathbf{0}$ to the right hand side of $\mathbf{r}_d = \mathbf{g}^{d^\infty} + (\mathbf{I} - \mathbf{P}_d)\mathbf{h}$ and note that $\mathbf{g}^{d^\infty} = \mathbf{P}_d^*\mathbf{r}_d$ so that

$$\mathbf{r}_d = \mathbf{g}^{d^\infty} + (\mathbf{I} - \mathbf{P}_d + \mathbf{P}_d^*)\mathbf{h} = \mathbf{P}_d^*\mathbf{r}_d + (\mathbf{I} - \mathbf{P}_d + \mathbf{P}_d^*)\mathbf{h}$$

Subtracting $\mathbf{P}_d^*\mathbf{r}_d$ from both sides and multiplying by the inverse of $(\mathbf{I}-\mathbf{P}_d+\mathbf{P}_d^*)$ yields

$$\mathbf{h} = (\mathbf{I} - \mathbf{P}_d + \mathbf{P}_d^*)^{-1}(\mathbf{I} - \mathbf{P}_d^*)\mathbf{r}_d = \mathbf{H}_d\mathbf{r}_d = \mathbf{h}^{d^\infty}$$

concluding the proof. □

We now discuss the many important implications of this result.

1. We will see below that the system of equations (1.32) provide the basis for the average reward Bellman (optimality) equation. Consequently understanding its properties is fundamental.

2. Theorem 1.6 establishes that (1.32) uniquely determines $\mathbf{g}^{d^\infty}$ and determines $\mathbf{h}^{d^\infty}$ up to a vector in the null space[9]of $\mathbf{I}-\mathbf{P}_d$. Part c. provides a way of identifying $\mathbf{h}^{d^\infty}$ but it requires determining $\mathbf{P}_d^*$. The discussion below will provide alternatives.

3. The equation $(\mathbf{I}-\mathbf{P}_d)\mathbf{g} = 0$ shows only that $\mathbf{g}$ is an element of the null space of $(\mathbf{I}-\mathbf{P}_d)$. Since the null space of $(\mathbf{I}-\mathbf{P}_d)$ is equivalent to the space of right eigenvectors of $\mathbf{P}_d$ corresponding to the eigenvalue 1, this means that this equation determines $\mathbf{g}$ up to m arbitrary constants where m represents the number of closed classes of $\mathbf{P}_d$. We will be most interested in the case when $m = 1$ where this system of equations simplifies further as the corollary below shows.

4. As an alternative to adding the condition $\mathbf{P}^*h = 0$ to uniquely determine $\mathbf{h}^{d^\infty}$ we can instead add the *third* equation

$$-\mathbf{h} + (\mathbf{P}_d - \mathbf{I})\mathbf{w}$$

[9]The vector $\mathbf{x} \neq \mathbf{0}$ is in the *null space* of the matrix $\mathbf{A}$ if $\mathbf{A}\mathbf{x} = \mathbf{0}$.

to the system (1.32). Solving this system of three matrix equations uniquely determines $\mathbf{g}^{d^\infty}$ and $\mathbf{h}^{d^\infty}$ and determines $\mathbf{w}$ up to an element of the null space of $\mathbf{I} - \mathbf{P}_d$. When $\mathbf{w}$ is uniquely determined by the condition $\mathbf{P}_d^* \mathbf{w} = \mathbf{0}$ or by adding a *fourth* equation, it corresponds to the third term in the Laurent series[10] expansion of $\mathbf{v}_\lambda^{d^\infty}$.

Simplifications for regular and unichain models

The following corollary shows how equations (1.32) simplify when the Markov chain corresponding to a Markovian decision rule has a single closed class, that is it is regular or unichain. It's proof is an immediate consequence of the previous theorem and the last remark above.

Corollary 1.3. Suppose $\mathbf{P}_d$ has a single closed class.
a. Then if $(\mathbf{I} - \mathbf{P}_d)\mathbf{g} = 0$, $\mathbf{g} = g\mathbf{e}$, for scalar g[a]
b. Suppose g and $\mathbf{h}$ satisfy

$$g\mathbf{e} + (\mathbf{I} - \mathbf{P}_d)\mathbf{h} = \mathbf{r}_d. \tag{1.39}$$

Then $g\mathbf{e} = \mathbf{g}^{d^\infty}$ and $\mathbf{h} = \mathbf{h}^{d^\infty} + k\mathbf{e}$ where k is a scalar.
c. Suppose g and $\mathbf{h}$ satisfy

$$g\mathbf{e} + (\mathbf{I} - \mathbf{P}_d)\mathbf{h} = \mathbf{r}_d$$

and $\mathbf{P}_d^* \mathbf{h} = 0$. Then $g\mathbf{e} = \mathbf{g}^{d^\infty}$ and $\mathbf{h} = \mathbf{h}^{d^\infty}$.

[a] Recall that $\mathbf{e}$ denotes a vector with all components equal to one.

Hence for unichain models, we can find $\mathbf{g}^{d^\infty}$ and $\mathbf{h}^{d^\infty}$ up to a constant by solving (1.39). This equation can be expressed in component notation as

The unichain policy evaluation equation

$$g + h(s) - \sum_{j \in S} p_d(j|s) h(j) = r_d(s) \tag{1.40}$$

for all $s \in S$.

[10] See Theorem 8.2.8 and Chapter 9 in Puterman [1994] for details on this concept.

Example 1.2(ctd.) We return to the model in Example 1.2. Since $\mathbf{P}_d$ consists of a single closed class, Corollary 1.3 implies that we need only solve the following under-determined system with three variables and two unknowns to find $\mathbf{g}^{d^\infty}$ and to find $\mathbf{h}^{d^\infty}$ up to an additive constant:

$$g + 0.2h(s_1) - 0.2h(s_2) = 3$$
$$g - 0.4h(s_1) + 0.4h(s_2) = 2.$$

Multiplying the first equation by 2 and adding it to the second equation shows that $g = 2\frac{2}{3}$. Substituting this value back into the first equation shows that $h(s_1) - h(s_2) = 1\frac{2}{3}$ in agreement with our previous analysis. Since

$$\mathbf{P}_d^* = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}$$

the condition $\mathbf{P}_d^*\mathbf{h} = \mathbf{0}$ implies that $\frac{2}{3}h(s_1) + \frac{1}{3}h(s_2) = 0$ or equivalently $h(s_2) = -2h(s_1)$. Hence $h^{d^\infty}(s_1) = \frac{5}{9}$ and $h^{d^\infty}(s_2) = -\frac{10}{9}$. Alternatively we can arbitrarily set $h(s_2) = 0$ to yield the solution $h(s_1) = 1\frac{2}{3}$.

Relative values

For a unichain model, we introduce a quantity we refer to as the relative value. To define it, specify a state s^* and a stationary policy d^∞.

The *relative value* of d^∞, denoted by $\mathbf{h}_{rel}^{d^\infty}$, is the solution of (1.39) subject to the condition $h(s^*) = 0$.

As we saw in the previous calculation,

$$h_{rel}^{d^\infty}(s) = h^{d^\infty}(s) - h^{d^\infty}(s^*). \tag{1.41}$$

In the above example, by choosing $s^* = s_2$ we found the relative values $h_{rel}^{d^\infty}(s_1) = 1\frac{2}{3}$ and $h_{rel}^{d^\infty}(s_2) = 0$. We will see below that relative values suffice for optimization so from this perspective we need not undertake the extra effort to evaluate $\mathbf{h}^{d^\infty}$.

1.4 Chain structure

In contrast to discounted models, in average reward models transition properties of Markov chains generated by stationary policies impact the form of the average reward and the structure of the evaluation and Bellman equations.

We say that a Markov decision process is:

- *Regular*[11] if the Markov chain corresponding to every Markovian deterministic decision rule is regular. That is, it consists of a single closed class and no transient states.
- *Unichain* if the Markov chain corresponding to every deterministic decision rule contains a single closed class and at least one decision rule contains a non-empty set of transient states.
- *Multi-chain* if at least one Markovian deterministic decision rule contains two or more closed classes.
- *Communicating* if for all states s and j in S, there exists a decision rule for which j is accessible from s, that is $p_d^n(j|s) > 0$ for some $n > 0$.
- *Weakly communicating* if the state space can be decomposed into two sets of state, one containing states that are transient under every stationary policy and the other which forms a communicating Markov decision process[12].

The first three categories depend on the chain structure of *every* Markovian deterministic decision rule. On the other hand, determining whether a model is communicating or weakly communicating requires only that *at least one* decision rule has the desired property. Note that a regular model is necessarily communicating, a unichain model is necessarily weakly communicating and a multi-chain model may or may not be communicating or weakly communicating.

This classification scheme must be taken into consideration because it determines what approach to use to analyze average reward Markov decision process. In practice (see Exercise 9) we may encounter real applications that are multi-chain and communicating (or weakly communicating).

We distinguish communicating models for the following reason.

Proposition 1.1. In a communicating model the optimal average reward must be constant.

Proof. Suppose there exists a policy with closed classes $C_1, C_2, \ldots, C_m$ with different average rewards on each. Assume further that C_1 has the greatest average reward g_1. Since the model is communicating, there exists a stationary policy under which there is a positive probability of reaching C_1 from each state in $C_2, \ldots, C_m$. Hence under this policy, all states in $C_2, \ldots, C_m$ are transient and have average reward g_1. □

Table 1.4 summarizes the impact of chain structure on evaluation equations, the average reward of a stationary policy and the optimal average reward. Note that communicating models have constant gain as shown above but require (1.32) because some policies may be multi-chain.

[11] Some books and papers refer to this condition as *recurrent* or *ergodic*.

[12] Some authors refer to such a model as weakly accessible

Model class	Evaluation equations	Gain of a stationary policy	Gain of optimal policy
regular	(1.39)	constant	constant
unichain	(1.39)	constant	constant
multi-chain	(1.32)	constant or non-constant	constant or non-constant
communicating	(1.32)	constant or non-constant	constant
weakly communicating	(1.32)	constant or non-constant	constant

Table 1.4: How model features vary with chain structure

The following two examples illustrate these properties.

Example 1.3. We again consider the two state model in Section **??**. In it there are four deterministic decision rules. The Markov chain corresponding to decision rule $d(s_1) = a_{1,1}$ and $d(s_2) = a_{2,1}$ has a unichain transition matrix with state s_1 transient and state $\{s_2\}$ a closed class. Also, all other deterministic decision rules correspond to regular Markov chains. Thus this model is unichain. Moreover it is easy to see that it is also communicating.

A variant: Suppose now we add a third action $a_{1,3}$ in s_1 with $p(s_1|s_1, a_{1,3}) = 1$, $p(s_2|s_1, a_{1,3}) = 0$ and arbitrary reward k (see Figure 1.2). Then as shown in Figure 1.3 the Markov chain corresponding to decision rule $d'(s_1) = a_{1,3}$ and $d'(s_2) = a_{2,1}$ has two closed classes. Hence with this additional action the model is multi-chain and communicating. This illustrates the interesting property that a model remains communicating when actions are added to it.

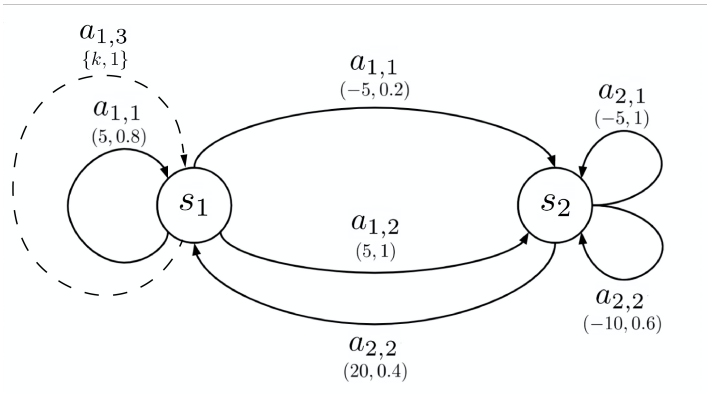

Figure 1.2: Model variant discussed in Example 1.3

(a) Chain structure under policy d^∞ (b) Chain structure under policy $(d')^\infty$

Figure 1.3: Symbolic representation of the two policies discussed in Example 1.3. Observe that under d^∞, s_1 is transient and s_2 is recurrent so that this policy generates a unichain model while under policy $(d')^\infty$, s_1 and s_2 each form closed classes so that the model is multi-chain.

Example 1.4. We now consider a modified version of the coffee delivering robot model of Section **??** in which the robot starts in the coffee room, cell 13, with a full cup of coffee and seeks to deliver it to the office, cell 1. In this model, $S = \{1, \ldots, 15\}$ where each state represents the location of the robot. Recall that actions specify an intended direction for the robot. Letting p denote the probability the robot moves in the intended direction and $(1-p)/(k-1)$ the probability it moves in each of the remaining $k-1$ directions including the possibility of remaining in the same cell.

In this model states 1 (deliver coffee) and 7 (fall down the stairs), each correspond to a closed class consisting of one absorbing state. When $p < 1$ the remaining set of states are transient under any decision rule since the robot reaches cells 1 or 7 with probability 1 under any stationary policy. When $p = 1$ decision rules can generate other closed classes. For example, by choosing the action R (right) in cell 2 and L (left) in cell 3 the robot will cycle between cells 2 and 3 forever and never reach cell 1 or cell 7.)

Hence in either case the model is multi-chain but when $p < 1$, the only closed classes are $\{1\}$ and $\{7\}$. Note that these closed classes each consists of a single reward-free state.

A variant: Suppose, as shown in Figure 1.4, we add an absorbing state Δ and an action a_Δ in states 1 and 7 which results in a transition to Δ with certainty so that $p(\Delta|1, a_\Delta) = p(\Delta|1, a_\Delta) = 1$. Moreover if the system reaches Δ it remains there forever and receives no reward. We can represent this by adding an action a' for which $p(\Delta|\Delta, a') = 1$. Moreover we set $r(1, a_\Delta, \Delta) = r(1, a_\Delta, \Delta) = r(\Delta, a', \Delta) = 0$. When $p < 1$, $\{\Delta\}$ is the only closed class, all other states are transient and the model is unichain.

We emphasize that when $p < 1$ in both the first formulation above and its

variant, the average reward of every policy is 0 so the average reward criterion does not distinguish policies. The expected total reward, which here is equivalent to the bias, is a more appropriate optimality criterion for this model.

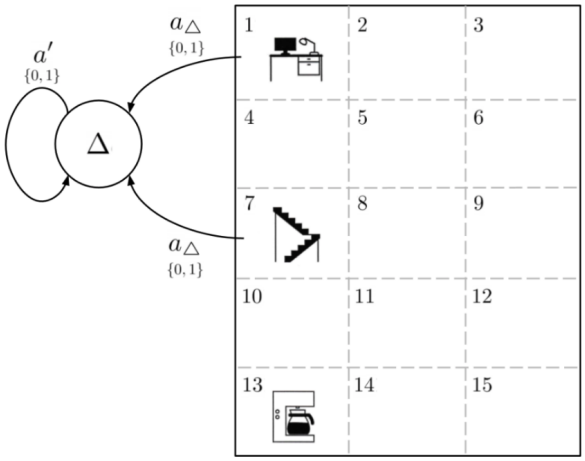

Figure 1.4: Variant of coffee delivering robot model with added absorbing state.

1.5 The Bellman Equation and its properties

As in earlier chapters, the Bellman equation plays a fundamental role in average reward Markov decision processes. Unfortunately, multi-chain models require a pair of nested Bellman equations while regular and unichain models require only a single (vector) equation. Because of this we consider only regular and unichain models in this section.

We show that a solution exists to the Bellman equation, that it corresponds to the optimal gain[13] and that it can be used to find optimal policies.

1.5.1 The Bellman equation in unichain models

The Bellman equations for a unichain model in component form are given by

$$h(s) = \max_{a \in A_s} \{ r(s,a) - g + \sum_{j \in S} p(j|s,a) h(j) \} \tag{1.42}$$

or equivalently in matrix-vector form as

[13] It doesn't determine the optimal bias. Instead it determines the gain of an average optimal policy and its bias up to a constant. A further optimality equation is required to determine the optimal bias. Moreover as Example 1.7 shows, there may exist average optimal policies with gain and bias that do not satisfy the Bellman equation.

$$\mathbf{h} = \underset{d \in D^{MD}}{\text{c-max}}\{\mathbf{r}_d - g\mathbf{e} + \mathbf{P}_d\mathbf{h}\}. \tag{1.43}$$

We often refer to this as the *average optimality equation* and use the shorthand AOE. Recall that c-max in (1.43) refers to the component-wise maximum as explicitly expressed in (1.42).

The following example illustrates these equation and some properties of solutions.

Example 1.5. The Bellman equations (1.42) for the unichain model in Example **??** become

$$h(s_1) = \max\{3 - g + 0.8h(s_1) + 0.2h(s_2), 5 - g + h(s_2)\} \tag{1.44}$$
$$h(s_2) = \max\{-5 - g + h(s_2), 2 - g + 0.4h(s_1) + 0.6h(s_2)\} \tag{1.45}$$

Observe that there are two equations with three variables so the solution is not unique.

In most instances, solving the Bellman equations requires value iteration, policy iteration or linear programming. In this example we will solve them by using the result (which we will soon demonstrate) that there exists a deterministic stationary policy that is optimal. Hence by enumerating the stationary policies and computing $g^{d^\infty} = \mathbf{P}_d^*\mathbf{r}_d$ for each, we find that $(d^*)^\infty$ with $d^*(s_1) = a_{1,2}$ and $d^*(s_2) = a_{2,2}$ is optimal with $g^{(d^*)^\infty} = 2\frac{6}{7} = 2.857$.

Solving the evaluation equations (1.40) subject to the condition $\mathbf{P}_{d^*}\mathbf{h} = \mathbf{0}$ we obtain the solution $g^{(d^*)^\infty} = 2\frac{6}{7}$, $h^{(d^*)^\infty}(s_1) = \frac{75}{49}$ and $h^{d^\infty}(s_2) = -\frac{30}{49}$. If instead we solve these equations under the condition $h(s_2) = 0$ we obtain $g^{(d^*)^\infty}$ together with the relative values $h_{rel}^{(d^*)^\infty}(s_1) = \frac{15}{7}$ and $h_{rel}^{(d^*)^\infty}(s_2) = 0$. Finally either version of $h(s_1)$ and $h(s_2)$ plus a constant k also satisfies this system of equations.

We now verify that these values satisfy the Bellman equations. It is easiest to do so for the relative values. Consider equation (1.44)

$$h(s_1) = \max\{3 - g + 0.8h(s_1) + 0.2h(s_2), 5 - g + h(s_2)\}.$$

Since $h_{rel}^{(d^*)^\infty}(s_1) = 5 - g^{(d^*)^\infty} + h_{rel}^{(d^*)^\infty}(s_2)$, we need only check the first expression in brackets. Since $3 - g^{(d^*)^\infty} + 0.8h_{rel}^{(d^*)^\infty}(s_1) + 0.2h_{rel}^{(d^*)^\infty}(s_2) = \frac{13}{7} < \frac{13}{7} = h_{rel}^{(d^*)^\infty}(s_1)$, the first equation holds.

Now consider the equation (1.45)

$$h(s_2) = \max\{-5 - g + h(s_2), 2 - g + 0.4h(s_1) + 0.6h(s_2)\}.$$

Since $h_{rel}^{(d^*)^\infty}(s_2) = 2 - g^{(d^*)^\infty} + 0.4h_{rel}^{(d^*)^\infty}(s_1) + 0.6h_{rel}^{(d^*)^\infty}(s_2)$, we need only check the first equation in brackets. Since $-5 - g^{(d^*)^\infty} + h_{rel}^{(d^*)^\infty}(s_2) = -7\frac{6}{7} < 0 = h_{rel}^{(d^*)^\infty}(s_2)$ the second equation holds. Hence we have found a solution to the Bellman equations.

We can motivate the derivation of the Bellman equation in the same way that we derived the evaluation equations in Section 1.3.4. That is we start with the discounted optimality equation and replace $\mathbf{v}$ by its partial Laurent expansion. Formalizing this argument provides an existence proof of the existence of a solution to the optimality equation. Alternatively we can start with the finite horizon optimality equation and replace the N-stage value by its representation in terms of the gain and bias.

1.5.2 Bounds on the optimal average reward

We show that a solution of the Bellman equation provides the optimal gain by providing inequalities that provide upper and lower bounds. We relegate the rather technical but insightful proof to the Appendix of this chapter.

Theorem 1.7. Supposed there exist a scalar g and a bounded vector $\mathbf{h}$ for which:
a.

$$\mathbf{h} \geq \underset{d \in D^{MD}}{\text{c-max}}\{\mathbf{r}_d - g\mathbf{e} + \mathbf{P}_d\mathbf{h}\}. \tag{1.46}$$

Then $g \geq g_+^*(s)$ for all $s \in S$.
b.

$$\mathbf{h} \leq \underset{d \in D^{MD}}{\text{c-max}}\{\mathbf{r}_d - g\mathbf{e} + \mathbf{P}_d\mathbf{h}\}. \tag{1.47}$$

Then $g \leq g_-^*(s)$ for all $s \in S$.
c.

$$\mathbf{h} = \underset{d \in D^{MD}}{\text{c-max}}\{\mathbf{r}_d - g\mathbf{e} + \mathbf{P}_d\mathbf{h}\}. \tag{1.48}$$

Then $g = g^*(s)$ for all $s \in S$.

Some comments on this result follow:

1. Part c. expresses the most important result for our development, namely that the existence of a solution to the Bellman equation (1.43) identifies the optimal average reward.

2. Parts a. and b. hold regardless of chain structure.

3. Part c. is vacuous when the optimal gain is not constant because no such g and $\mathbf{h}$ need exist as the following example shows.

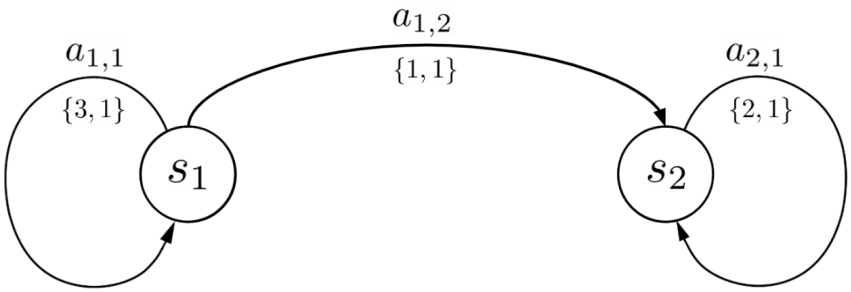

Figure 1.5: Model in Example 1.6

Example 1.6. Consider the deterministic model shown in Figure 1.5 with $S = \{s_1, s_2\}$, $A_{s_1} = \{a_{1,1}, a_{1,2}\}$ and $A_{s_2} = \{a_{2,1}\}$ with $r(s_1, a_{1,1}) = 3, r(s_1, a_{1,1}) = 1$ and $r(s_2, a_{2,1}) = 2$. Clearly the model is multi-chain and not communicating. The optimal policy is d^∞ with $d(s_1) = a_{1,1}$ and $d(s_2) = a_{2,1}$. Moreover its easy to see that $g^*(s_1) = 3$ and $g^*(s_2) = 2$.

Then $g = 3$ and $h(s_1) = h(s_2) = 0$ satisfy (1.46) which is given by

$$\begin{aligned} h(s_1) &\geq \max\{3 - g + h(s_1), 1 - g + h(s_2)\} \\ h(s_2) &\geq 2 - g + h(s_2) \end{aligned}$$

so that $g \geq g^*(s)$ for $s \in S$ as shown in part a. of the theorem. Moreover $g' = 2$ and $h'(s_1) = h'(s_2) = 0$ satisfy (1.47) of the theorem so that $g' \leq g^*(s)$ for $s \in S$. Hence we have established that $2\mathbf{e} \leq \mathbf{g}^* \leq 3\mathbf{e}$.

Equation (1.48) becomes

$$\begin{aligned} h(s_1) &= \max\{3 - g + h(s_1), 1 - g + h(s_2)\} \\ h(s_2) &= 2 - g + h(s_2) \end{aligned}$$

The second equation requires that $g = 2$ so that the first equation must be violated. Hence part c. of theorem does not apply.

1.5.3 Solutions of the Bellman equation

We now establish the existence of a solution to the unichain optimality equation. The proof is quite informative so we include it below. It formalizes the use of the partial Laurent series of the expected discounted reward to link the gain and bias of a policy to its discounted reward. The fundamental idea is that in a model with finitely many

deterministic stationary policies, we can find a sequence of discount factors converging to one for which the *same* policy denoted by δ^∞ is discount optimal[14]

Theorem 1.8. In a finite state and action unichain model;
a. There exist a unique scalar g and a vector $\mathbf{h}$ that satisfy the Bellman equation (1.43)
b. There exists a deterministic stationary optimal policy.

Proof. Choose a sequence of discount factors $\{\lambda_n'\}$ for which $\lambda_n' \uparrow 1$. For each n there exists a deterministic stationary discount optimal policy by Theorem **??**. Since the set of stationary policies is finite, we can select a subsequence of $\{\lambda_n'\}$, denoted by $\{\lambda_n\}$ for which the policy δ^∞ is optimal. Hence $\mathbf{v}_{\lambda_n}^{\delta^\infty} = \mathbf{v}_{\lambda_n}^*$ for $n \geq 1$.

This means that $\mathbf{v}_{\lambda_n}^{\delta^\infty}$ satisfies the discounted Bellman equation (**??**) which can be rewritten as

$$\mathbf{0} = \underset{d' \in D^{MD}}{\text{c-max}} \{\mathbf{r}_{d'} + (\lambda_n \mathbf{P}_{d'} - \mathbf{I})\mathbf{v}\}$$

for all $n \geq 1$. Thus it follows that

$$\mathbf{0} = \mathbf{r}_\delta + (\lambda_n \mathbf{P}_\delta - \mathbf{I})\mathbf{v}_{\lambda_n}^{\delta^\infty} = \underset{d' \in D^{MD}}{\text{c-max}} \{\mathbf{r}_{d'} + (\lambda_n \mathbf{P}_{d'} - \mathbf{I})\mathbf{v}_{\lambda_n}^{\delta^\infty}\} \geq \mathbf{r}_d + (\lambda_n \mathbf{P}_d - \mathbf{I})\mathbf{v}_{\lambda_n}^{\delta^\infty} \tag{1.49}$$

for all $d \in D^{MD}$.

From Theorem 1.4, there exists g^{δ^∞} and $\mathbf{h}^{\delta^\infty}$ for which

$$\mathbf{v}_{\lambda_n}^{\delta^\infty} = \frac{g^{\delta^\infty}\mathbf{e}}{1-\lambda_n} + \mathbf{h}^{\delta^\infty} + \mathbf{f}(\lambda_n) \tag{1.50}$$

where $\mathbf{f}(\lambda_n) \to \mathbf{0}$ as $n \to \infty$.

Now substitute (1.50) into the right hand side of (1.49) to obtain for any $d \in D^{MD}$ that

$$\begin{aligned} \mathbf{r}_d + (\lambda_n \mathbf{P}_d - \mathbf{I})\mathbf{v}_{\lambda_n}^{\delta^\infty} &= \mathbf{r}_d + (\lambda_n \mathbf{P}_d - \mathbf{I}) \left[\frac{g^{\delta^\infty}\mathbf{e}}{1-\lambda_n} + \mathbf{h}^{\delta^\infty} + \mathbf{f}(\lambda_n) \right] \\ &= \mathbf{r}_d - g^{\delta^\infty}\mathbf{e} + (\lambda_n \mathbf{P}_d - \mathbf{I})\mathbf{h}^{\delta^\infty} + \mathbf{f}'(\lambda_n) \end{aligned} \tag{1.51}$$

where $\mathbf{f}'(\lambda_n) \to \mathbf{0}$ as $n \to \infty$. Substituting this expression for δ and d into (1.49) and letting $n \to \infty$ yields

$$\mathbf{0} = \mathbf{r}_\delta - g^{\delta^\infty}\mathbf{e} + (\mathbf{P}_\delta - \mathbf{I})\mathbf{h}^{\delta^\infty} \geq \mathbf{r}_d - g^{\delta^\infty}\mathbf{e} + (\mathbf{P}_d - \mathbf{I})\mathbf{h}^{\delta^\infty}$$

for all $d \in D^{MD}$. From this we conclude that

$$\mathbf{0} = \underset{d' \in D^{MD}}{\text{c-max}} \{\mathbf{r}_{d'} - g^{\delta^\infty}\mathbf{e} + (\mathbf{P}_{d'} - \mathbf{I})\mathbf{h}^{\delta^\infty}\}. \tag{1.52}$$

Hence Part a. follows. Moreover since the pair $(g^{\delta^\infty}, \mathbf{h}^{\delta^\infty})$ satisfy the optimality equation, part c. of Theorem 1.7 implies that $g^{\delta^\infty} = g^*$ so δ^∞ is average optimal. □

[14]The concept of *Blackwell optimality* could be used here. Note that a policy π is Blackwell optimal if there exists a λ^* for which π is discount optimal for all $\lambda \in [\lambda^*, 1)$.

1.5.4 The Bellman equation and optimal policies

Finally we show how to use solutions to the Bellman equation to identify optimal policies.

Theorem 1.9. Suppose $(g^*, \mathbf{h}^*)$ are solutions of the Bellman equation and for all $s \in S$

$$d^*(s) \in \arg\max_{a \in A_s} \left[r(s,a) + \sum_{j \in S} p(j|s,a) h^*(j) \right] \tag{1.53}$$

or equivalently

$$d^* \in \arg\max_{d \in D^{MD}} [\mathbf{r}_d + \mathbf{P}_d \mathbf{h}^*]. \tag{1.54}$$

Then $(d^*)^\infty$ is average optimal.

Proof. We prove the result using vector notation. From (1.54)

$$\mathbf{h}^* = \max_{d \in D^{MD}} [\mathbf{r}_d - g^* \mathbf{e} + \mathbf{P}_d \mathbf{h}^*] = \mathbf{r}_{d^*} - g^* \mathbf{e} + \mathbf{P}_{d^*} \mathbf{h}^*.$$

As a consequence of Corollary 1.3, $g^* = g^{(d^*)^\infty}$ establishing the result. □

Some comments about this result follow:

1. We emphasize that in a finite action model, there always exists such a d^*. This result holds in greater generality but requires some technical considerations.

2. Surprisingly neither (1.53) nor (1.53) includes g^* directly. However it does so indirectly since the pair $(g^*, \mathbf{h}^*)$ satisfies the Bellman equation.

3. Some authors refer to any decision rule d which satisfies

$$d \in \arg\max_{d \in D^{MD}} [\mathbf{r}_d + \mathbf{P}_d \mathbf{h}]$$

   as $\mathbf{h}$-*improving*.

4. As (1.53) and (1.53) suggest, d^* need not be unique.

We conclude this section with the following curious example[15] that shows that there exist average optimal policies with gain and bias that do not satisfy the optimality equation.

[15] This appears in slightly modified form as Example 8.4.3 in Puterman [1994].

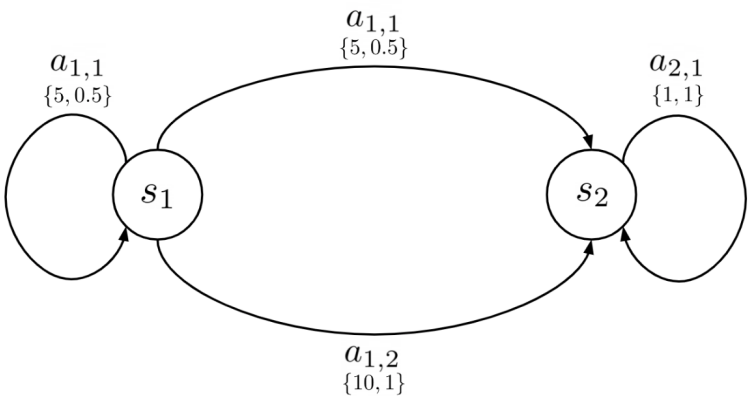

Figure 1.6: Symbolic representation of model in Example 1.7

Example 1.7. Let $S = \{s_1, s_2\}$; $A_{s_1} = \{a_{1,1}, a_{1,2}\}$, $A_{s_2} = \{a_{2,1}\}$; $p(s_1|s_1, a_{1,1}) = p(s_2|s_1, a_{1,1}) = 0.5$, $p(s_2|s_1, a_{1,2}) = 1$, $p(s_2|s_2, a_{2,1}) = 1$; and $r(s_1, a_{1,1}, s_1) = r(s_1, a_{1,1}, s_1) = 5$, $r(s_1, a_{1,2}, s_2) = 10$ $r(s_2, a_{2,1}, s_2) = 1$. Rewards need not be defined for zero probability transitions. See Figure 1.6[a].

Let δ denote the decision rule which uses action $a_{1,1}$ in s_1 and γ denote the decision rule which uses action $a_{1,2}$ in s_1. This model is unichain with s_2 absorbing and s_1 transient under both stationary policies. Clearly $g^{\delta^\infty} = g^{\gamma^\infty} = g^* = 1$ so that both stationary policies are average optimal.

We now find the gain and bias of each stationary policy by solving $\mathbf{h} = \mathbf{r}_d - g\mathbf{e} + \mathbf{P}_d\mathbf{h}$ subject to $\mathbf{P}_d^*\mathbf{h} = \mathbf{0}$. Since

$$\mathbf{P}_\delta^* = \mathbf{P}_\gamma^* = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$h^{\delta^\infty}(s_2) = h^{\gamma^\infty}(s_2) = 0$ so that it follows that $h^{\delta^\infty}(s_1) = 8$ and $h^{\gamma^\infty}(s_1) = 9$. Since the Bellman equation is given by

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

we see that $(g^{\gamma^\infty}, \mathbf{h}^{\gamma^\infty})$ satisfies the Bellman equation but $(g^{\delta^\infty}, \mathbf{h}^{\delta^\infty})$ does not. Why do you think this is so?

[a] A variant of this example is analyzed throughout Puterman [1994]

1.5.5 State-action value functions

Just as in the discounted case, we can cast value iteration in terms of state-action value functions. As in previous chapters, they will not be central to this chapter, but will be fundamental when we consider simulation-based methods. Discussion here will be brief.

Suppose the scalar $g*$ and $h^*(s)$ for all $s \in S$ solve the unichain optimality equation (1.40). Define the (optimal) state-action value functions $q^*(s,a)$ by by

$$q^*(s,a) := r(s,a) + \sum_{j \in S} p(j|s,a) h^*(j) \tag{1.55}$$

for all $a \in A_s$ and $s \in S$.

Since the Bellman equation can be expressed as

$$g^* + h^*(s) = \max_{a \in A_s} \{ r(s,a) + \sum_{j \in S} p(j|s,a) h^*(j) \} \tag{1.56}$$

it follows that

$$h^*(s) = \max_{a \in A_s} q^*(s,a) - g^*.$$

Hence, the solving the is equivalent to finding a scalar g^* and a function $q^*(s,a)$ defined for $s \in S$ and $a \in A_s$ which satisfy:

$$q(s,a) = r(s,a) - g + \sum_{j \in S} p(j|s,a) \max_{a \in A_s} q(j,a). \tag{1.57}$$

Choosing $d^*(s) \in \arg\max_{a \in A_s} q^*(s,a)$ identifies an optimal stationary policy.

Written as an expected value, (1.57) is equivalent to

$$q(s,a) = r(s,a) + E^{d^*} \left[\max_{a \in A_s} q(X_2, a) \,\middle|\, X_1 = s \right], \tag{1.58}$$

where $(d^*)^\infty$ is an optimal policy. The previous equation extends to models in which $r(s,a,j)$ is the model primitive as follows:

$$q(s,a) = E^{d^*} \left[r(X_1, a, X_2) + \max_{a \in A_s} q(X_2, a) \,\middle|\, X_1 = s \right]. \tag{1.59}$$

Observe that these two expressions differ from the usual Bellman equation in that the maximization is *inside* the expectation. While this difference is unimportant in this chapter, it will have a significant impact in Chapter **??**, where expectations are estimated by random draws from the transition probability distribution.

1.6 Value Iteration

Average reward value iteration refers to an algorithm based on the recursion expressed in component notation as

$$v^{n+1}(s) = \max_{a \in A_s} \{r(s,a) + \sum_{j \in S} p(j|s,a) v^n(j)\} \tag{1.60}$$

or equivalently in matrix-vector form as

$$\mathbf{v}^{n+1} = \text{c-}\max_{d \in D^{MD}} \{\mathbf{r}_d + \mathbf{P}_d \mathbf{v}^n\} := L\mathbf{v}^n. \tag{1.61}$$

Several challenges emerge to create a convergent algorithm from these recursions.

1. The average reward g does appear in the above expressions; we require a method to identify the optimal gain from the sequence of iterates.
2. The iterates of an algorithm based on these recursions may diverge; we require a method to obtain a convergent sequence from them.
3. The iterates of an algorithm based on these recursions may oscillate. We must identify when this is the case and develop remedies to address this concern.
4. Span based stopping criteria may not result in algorithm termination in multi-cahin models.

1.6.1 Examples

We consider two examples example that illustrate the challenges when using value iteration in average reward models.

A straight-forward example

We begin with our standard example to illustrate the divergence of iterates of average reward value iteration.

Example 1.8. We apply value iteration to the model in Example **??**. We initialize it at $v^0(s_1) = v^0(s_2) = 0$ to obtain the numerical results in Table 1.5. Observe that the iterates diverge but that $sp(\mathbf{v}^{n+1} - \mathbf{v}^n)$ converges to zero exponentially.

Observe further that for $s \in S$, $v^{10}(s) - v^9(s) = 2.857 = g^*$ which agrees with the result found in Example 1.5. See (1.31) for why this might be the case. We make this precise below.

n	$v^n(s_1)$	$v^n(s_2)$	$sp(\mathbf{v}^{n+1} - \mathbf{v}^n)$
0	0.00000	0.00000	na
1	5.00000	2.00000	3.0000000
2	7.40000	5.20000	0.8000000
3	10.20000	8.08000	0.0800000
4	13.08000	10.92800	0.0320000
5	15.92800	13.78880	0.0128000
6	18.78880	16.64448	0.0051200
7	21.64448	19.50221	0.0020480
8	24.50221	22.35912	0.0008192
9	27.35912	25.21635	0.0003277
10	30.21635	28.07346	0.0001311

Table 1.5: Iterates of value iteration for average reward version of Example 1.8

1.6.2 A periodic example

The following example provides an "edge-case" that causes value iteration to oscillate.

Example 1.9. Consider a two-state periodic model with $S = \{s_1, s_2\}$ and a single decision rule $d \in D^{MD}$ with reward and transition probabilities given by

$$\mathbf{r}_d = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{P}_d = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Set $\mathbf{v}^0 = \begin{bmatrix} a \\ b \end{bmatrix}$. Since $\mathbf{v}^n = \mathbf{P}_d^n \mathbf{v}^0$

$$\mathbf{v}^n = \begin{bmatrix} a \\ b \end{bmatrix} \quad \text{for } n \text{ even and} \quad \mathbf{v}^n = \begin{bmatrix} b \\ a \end{bmatrix} \quad \text{for } n \text{ odd.}$$

Hence unless $a = b$, the iterates oscillate. Moreover $sp(\mathbf{v}^{n+1} - \mathbf{v}^n) = 2|a - b|$. Since

$$\mathbf{P}_d^* = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

$g^{d^\infty} = 0$. From these observations we conclude that in a periodic model value iteration oscillates and need not converge with respect to the span semi-norm.

Removing periodicity

Periodicity is the most significant challenge to developing a convergent value iteration algorithm. In practice most applications will be aperiodic. However if you encounter a model with periodic policies, you can transform it into an aperiodic model by a simple transformation that preserves values (up to constant) and optimality.

To do so, for $0 < \tau < 1$ define a new model denoted by $\widetilde{\ }$ by

$$\widetilde{r(s,a)} = \tau r(s,a) \quad \text{and} \quad \widetilde{(p|j,s)} = (1-\tau)\delta(j|s) + \tau p(j|s,a) \tag{1.62}$$

for $s \in S$, $a \in A_s$ and $j \in S$ where $\delta(j|s) = 1$ if $j = s$ and 0 otherwise.

Alternatively this can be expressed in matrix-vector notation as

$$\widetilde{\mathbf{r}}_d = \tau \mathbf{r}_d \quad \text{and} \quad \widetilde{\mathbf{P}}_d = (1-\tau)\mathbf{I} + \tau \mathbf{P}_d \tag{1.63}$$

for $d \in D^{MD}$.

We now illustrate this transformation in the context of Example 1.9.

Example 1.9 ctd. **(recheck calculations below)** We apply the transformation to the above example to obtain

$$\widetilde{\mathbf{r}}_d = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{P}}_d = \begin{bmatrix} 1-\tau & \tau \\ \tau & 1-\tau \end{bmatrix}.$$

As a result of adding a self-transition with probability $1-\tau$, the transformed model is aperiodic. (We leave this as an exercise to prove this formally).

Since in this example $\mathbf{v}^n = \mathbf{P}_d^n \mathbf{v}_0$, we find by appealing to eigenvalue decomposition Theorem **??** in the Appendix that

$$\mathbf{P}_d^n = \begin{bmatrix} \frac{1}{2}[1+(1-2\tau)^n] & \frac{1}{2}[1-(1-2\tau)^n] \\ \\ \frac{1}{2}[1-(1-2\tau)^n] & \frac{1}{2}[1+(1-2\tau)^n] \end{bmatrix}.$$

Consequently

$$\mathbf{v}^{n+1} - \mathbf{v}^n = \mathbf{P}_d^n(\mathbf{P}_d - \mathbf{I})\mathbf{v}_0 = \tau \mathbf{P}_d^n \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{v}^0 = \tau(1-2\tau)^n \begin{bmatrix} b-a \\ a-b \end{bmatrix}.$$

Therefore $sp(\mathbf{v}^{n+1} - \mathbf{v}^n) = 2\tau(1-2\tau)^n|b-a|$ converges to 0 as $n \to \infty$ so that value iteration is convergent with respect to the span semi-norm.

The following result shows that that this transformation preserves solutions of the optimality equation and consequently optimal policies.

Proposition 1.2. Suppose for that the pair $(g^*, \mathbf{h}^*)$ satisfy (1.43).

a. Then the pair $(\tau g^*, \mathbf{h}^*)$ satisfy

$$\mathbf{h} = \underset{d \in D^{MD}}{\text{c-max}} \{\widetilde{r}_d - g\mathbf{e} + \widetilde{\mathbf{P}}_d \mathbf{h}\}. \tag{1.64}$$

b. The same policy is optimal for both models.

Proof. Rewriting (1.43) we obtain

$$\mathbf{h}^* + g^*\mathbf{e} = \underset{d \in D^{MD}}{\text{c-max}} \{\mathbf{r}_d + \mathbf{P}_d \mathbf{h}^*\}.$$

Multiply this equation by τ and add $(1-\tau)\mathbf{h}^*$ to both sides to obtain

$$\mathbf{h}^* + \tau g^*\mathbf{e} = \underset{d \in D^{MD}}{\text{c-max}} \{\tau \mathbf{r}_d + (1-\tau)\mathbf{I}\mathbf{h}^* + \tau \mathbf{P}_d \mathbf{h}^*\} = \underset{d \in D^{MD}}{\text{c-max}} \{\widetilde{\mathbf{r}}_d + \widetilde{P}_d \mathbf{h}^*\}$$

which establishes the result. □

Clearly the above result holds regardless how we specify $\mathbf{h}*$. Moreover replacing D^{MD} by a single policy establishes:

Corollary 1.4. Suppose for some $d \in D^{MD}$ that the pair $(g^{d^\infty}, \mathbf{h}^{d^\infty})$ satisfy (1.39), then the pair $(\tau g^{d^\infty}, \mathbf{h}^{d^\infty})$ satisfy

$$\mathbf{h} = \widetilde{r}_d - g\mathbf{e} + \widetilde{\mathbf{P}}_d \mathbf{h}.$$

Since we can apply this transformation to any model, we assume throughout the remainder of this section that *all policies are aperiodic* On occasion, we will restate it for emphasis.

1.6.3 A value iteration algorithm

To turn a recursion such as (1.61) or (1.60) into an algorithm requires initialization, a stopping criterion and a procedure to determine an ϵ-optimal policy and its value. We state it separately in vector and component notation however we will implement it in its component form.

Average reward value iteration: vector notation

1. **Initialize:** Select $\mathbf{v} \in V$ and specify stopping criterion.

2. **Iterate:** Do until stopping criteria met:

$$\mathbf{v} \leftarrow L\mathbf{v}$$

3. **Choose an ϵ-optimal policy:** Select

$$d^{\epsilon} \in \underset{d \in D^{\mathrm{MD}}}{\arg\,\text{c-max}} \{\mathbf{r}_d + \mathbf{P}_d \mathbf{v}\}.$$

4. **Approximate the optimal average reward:** Set

$$g^* = \frac{1}{2}\left(\max_{s \in S} \{Lv(s) - v(s)\} + \min_{s \in S} \{Lv(s) - v(s)\} \right). \tag{1.65}$$

Average reward value iteration: component notation

1. **Initialize:** Set $n = 0$, specify $\epsilon > 0$ and $v^0(s)$ for all $s \in S$.

2. **Iterate:** Compute $v^{n+1}(s)$ for all $s \in S$:

$$v^{n+1}(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^n(j) \right\}. \tag{1.66}$$

3. **Apply stopping criterion:** If

$$\text{sp}(\mathbf{v}^{n+1} - \mathbf{v}^n) < \epsilon, \tag{1.67}$$

proceed to step 4, else $n \leftarrow n + 1$ and return to step 2.

4. **Choose an ϵ-optimal policy:** For all $s \in S$, choose

$$d^\epsilon(s) \in \arg\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j|s,a) v^n(j) \right\}$$

.

5. **Approximate the optimal average reward:** Set

$$g^\epsilon = \frac{1}{2}[\max_{s \in S}\{v^{n+1}(s) - v^n(s)\} + \min_{s \in S}\{v^{n+1}(s) - v^n(s)\}]$$

.

Convergence of value iteration in unichain models

In the absence of contraction properties that were applicable in the discounted case, a proof that value iteration converges and eventually achieves the span based stopping criterion requires extensive subtle analysis. For ease of exposition, we establish the convergence of value iteration under the assumption that the transition probabilities of every stationary policy are **unichain**. Almost all of the results hold under the weaker condition of communicating or weakly communicating models which ensure that the optimal gain is constant.

The crucial idea is that in an aperiodic model,

$$\lim_{n \to \infty} [v^n(s) - ng^* - h^*(s)] \tag{1.68}$$

exists where g^* and $h^*(s)$ are solutions of the unichain optimality equation (1.42)[16].Note

[16]This result is valid in multi-chain models but requires that (g^*, h^*) be solutions of a pair of nested

that (1.5) establishes this result for each stationary policy but not for a solution of the optimality equation. We discuss the proof in Appendix 1.11.3.

Once we have established that the limit in (1.68) exists, the following important result follows easily.

Proposition 1.3. Assume a unichain aperiodic model in which $\{v^n(s)\}$ is generated by (1.60). Then for any $v^0(s)$

a. For all $s \in S$

$$\lim_{n\to\infty}\{v^{n+1}(s) - v^n(s)\} = g^*. \tag{1.69}$$

b. Assume $g^*(s)$ is constant. Then

$$\lim_{n\to\infty} \mathrm{sp}(\mathbf{v}^{n+1} - \mathbf{v}^n) = 0. \tag{1.70}$$

Proof. We prove part a. first. Fix $s' \in S$ and denote the limit in (1.5) by K. Then given $\epsilon > 0$ there exists an N such that for $n \geq N$ for which

$$|v^n(s') - ng^* - h^*(s') - K| < \epsilon.$$

Hence

$$\begin{aligned}
&|v^{n+1}(s') - v^n(s') - g^*| \\
&\quad = |[v^{n+1}(s') - (n+1)g^* - h^*(s') - K] - [v^n(s') - ng^* - h^*(s') - K]| \\
&\quad \leq |v^{n+1}(s') - (n+1)g^* - h^*(s') - K| + |v^n(s') - ng^* - h^*(s') - K| \\
&\quad < 2\epsilon.
\end{aligned}$$

Since s' was arbitrary, part a. follows.

We now prove b. Let

$$s^+ = \arg\max_{s\in S}\{v^{n+1}(s) - v^n(s)\} \text{ and } s^- = \arg\min_{s\in S}\{v^{n+1}(s) - v^n(s)\}.$$

Since g^* is constant, follows from part a. that

$$\begin{aligned}
\mathrm{sp}(\mathbf{v}^{n+1} - \mathbf{v}^n) &= \left(v^{n+1}(s^+) - v^n(s^+) - g^*\right) - \left(v^{n+1}(s^-) - v^n(s^-) - g^*\right) \\
&< |v^{n+1}(s^+) - v^n(s^+) - g^*| + |v^{n+1}(s^-) - v^n(s^-) - g^*| < \epsilon
\end{aligned}$$

Hence b. follows. □

The following results provides bounds on the optimal gain. They can be used to identify an ϵ-optimal policy when using value iteration and also estimate the optimal gain at termination. Moreover they may be useful for estimating the optimal gain when using simulation methods.

---

optimality equations.

Proposition 1.4. Assume a unichain model. Then for any $v(s)$,

$$\min_{s'\in S}\{Lv(s') - v(s')\} \le g^{d^\infty} \le g^* \le \max_{s'\in S}\{Lv(s') - v(s')\} \tag{1.71}$$

where

$$d(s) \in \arg\max_{a\in A_s}\left\{r(s,a) + \sum_{j\in S}\lambda p(j|s,a)v(j)\right\} \tag{1.72}$$

Proof. We first prove the right hand inequality. Suppose d satisfies (1.72) and let R_d denote its set of recurrent states. Then since $\mathbf{P}_d^* = \mathbf{P}_d^*\mathbf{P}_d$

$$\begin{aligned} g^{d^\infty}\mathbf{e} = \mathbf{P}_d^*\mathbf{r}_d = \mathbf{P}_d^*\left[\mathbf{r}_d + \mathbf{P}_d\mathbf{v} - \mathbf{v}\right] &= \mathbf{P}_d^*\left[L\mathbf{v} - \mathbf{v}\right] \\ &\ge \min_{s\in R_d}\{Lv(s) - v(s)\} \ge \min_{s\in S}\{Lv(s) - v(s)\}. \end{aligned}$$

Note that the minimum in the first inequality above is over R_d because columns corresponding to transient states of $\mathbf{P}_d$ will have 0's in columns of $\mathbf{P}_d^*$ corresponding to transient states.

From Theorem 1.8, there exists a g^* and a decision rule d^* for which

$$\begin{aligned} g^*\mathbf{e} = \mathbf{P}_{d^*}^*\mathbf{r}_{d^*} = \mathbf{P}_{d^*}^*\left[\mathbf{r}_{d^*} + \mathbf{P}_{d^*}\mathbf{v} - \mathbf{v}\right] &= \mathbf{P}_{d^*}^*\left[L\mathbf{v} - \mathbf{v}\right] \\ &\le \max_{s\in R_{d^*}}\{Lv(s) - v(s)\} \le \max_{s\in S}\{Lv(s) - v(s)\}. \end{aligned}$$

□

As an immediate consequence of the above propositions we have the main result of this section. The proof of parts b. and c. follow immediately from Proposition 1.4.

Theorem 1.10. Convergence of value iteration

Assume an aperiodic unichain model. Then

a. Value iteration converges and the stopping criterion (1.77) holds for n sufficiently large.

b. For every $n > 0$

$$\min_{s\in S}\{v^{n+1}(s) - v^n(s)\} \le g^{d^\infty} \le g^* \le \max_{s\in S}\{v^{n+1}(s) - v^n(s)\} \tag{1.73}$$

where

$$d(s) \in \arg\max_{a\in A_s}\left\{r(s,a) + \sum_{j\in S}\lambda p(j|s,a)v^n(j)\right\} \tag{1.74}$$

b. If $\text{sp}(\mathbf{v}^{n+1} - \mathbf{v}^n) < \epsilon$ for some $\epsilon > 0$ then

$$\left| \max_{s \in S} \{v^{n+1}(s) + v^n(s)\} + \min_{s \in S} \{v^{n+1}(s) - v^n(s)\} - 2g \right| < \epsilon \tag{1.75}$$

for $g = g^*$ or $g = g^{d^\infty}$ where d satisfies (1.74)

We comment about several key aspects of this proposition and theorem.

1. Part a. gives the most significant result. Namely in a model in which $g^*(s)$ is constant, the value iteration algorithm converges and the span-based stopping criterion is achieved.

2. The proof of part b. of Proposition 1.3 holds under the weaker assumption that the optimal gain is constant[17]. This is true when:

   (a) Every policy has a regular or unichain transition matrix.

   (b) The model is communicating.

   (c) The model is weakly communicating.

   Thus this result holds in considerable generality.

3. Part a. of Proposition 1.3 justifies the approximation

$$g^*(s) \approx v^{n+1}(s) - v^n(s).$$

   Part c. of the above theorem makes it precise and justifies the calculation in the step "Approximate the optimal average reward" in the value iteration algorithm statements.

4. Part b. of the above theorem shows how to obtain an ϵ-optimal policy d^∞ when the condition

$$\text{sp}(\mathbf{v}^{n+1} - \mathbf{v}^n) < \epsilon$$

   holds and provides the basis for the estimate in part c.

Referring back to Example 1.8, we see from Table 1.5 that $\text{sp}(\mathbf{v}^{n+1} - \mathbf{v}^n)$ converges to zero. At termination, using (1.65) we obtain the estimate $g^* = 2.857$.

[17] Without the constant gain assumption the above proof of part b. of Proposition 1.3 establishes only that

$$\text{sp}(\mathbf{v}^{n+1} - \mathbf{v}^n) = \text{sp}(\mathbf{g}^*)$$

so that (1.70) would not hold. Hence in a multi-chain model the span-based stopping criterion would not apply. See Puterman [1994] p.477 for an example.

1.6.4 Relative value iteration

Observe that in Example 1.8 the iterates of value iteration diverge. While this does not impact theory, it might become problematic in practice, especially in large models. To overcome this and as well provide a basis for a simulation approach, we often use *relative value iteration* instead. Relative value iteration distinguishes a state s' and normalizes by subtracting $v^n(s')$ from $v^n(s)$ after each iteration. We make this formal as follows:

Relative value iteration: component notation

1. **Initialize:** Choose a state s'. Set $n = 0$, specify $\epsilon > 0$ and specify $w^0(s)$ with the property that $w^0(s') = 0$.

2. **Iterate:** Compute $v^{n+1}(s)$ for all $s \in S$:

$$v^{n+1}(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) w^n(j) \right\}. \tag{1.76}$$

3. **Normalize:** For all $s \in S$, set

$$w^{n+1}(s) = v^{n+1}(s) - v^{n+1}(s').$$

4. **Apply stopping criterion:** If

$$\mathrm{sp}(\mathbf{w}^{n+1} - \mathbf{w}^n) < \epsilon, \tag{1.77}$$

   proceed to step 5, else $n \leftarrow n + 1$ and return to step 2.

5. **Choose an ϵ-optimal policy:** For all $s \in S$, choose

$$d^\epsilon(s) \in \arg\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j|s,a) w^n(j) \right\}$$

   .

6. **Approximate the optimal average reward:** Set

$$g^\epsilon = \frac{1}{2}[\max_{s \in S}\{v^{n+1}(s) - w^n(s)\} + \min_{s \in S}\{v^{n+1}(s) - w^n(s)\}]$$

   .

We now apply relative value iteration to Example 1.8 normalizing so that $w(s_2) = 0$.

n	$w^n(s_1)$	$w^n(s_2)$	$sp(\mathbf{w}^{n+1} - \mathbf{w}^n)$	Estimate of g^*
0	0.00000	0	na	na
1	3.00000	0	3.0	3.5
2	2.20000	0	0.8	2.8
3	2.12000	0	0.08	2.84
4	2.15200	0	0.032	2.864
5	2.13900	0	0.0128	2.8544
6	2.14432	0	0.00512	2.85824
7	2.14227	0	0.002048	2.85670
8	2.14309	0	0.000819	2.85732
9	2.14276	0	0.0003278	2.85707
10	2.14290	0	0.0001311	2.85717

Table 1.6: Iterates of relative value iteration for Example 1.8

Comparing this to Table 1.5 we observe that the span expressed in terms of $\mathbf{w}^n$ agrees with that expressed in terms of $\mathbf{v}^n$ (Why?). Moreover we see that the estimates of g^* converge quickly. Note that $w^n(s_1)$ converges to $h^*_{rel}(s_1) = \frac{15}{7}$ (see Example 1.5).

Thus we can conclude that with obvious modifications, Theorem 1.10 applies to relative value iteration and moreover, $w^n(s)$ converges to $h^*_{rel}(s)$.

1.7 Policy Iteration

In this section we describe the unichain policy iteration algorithm, illustrate its application in an example, prove its convergence and comment on some salient features.

1.7.1 The unichain policy iteration algorithm (UPIA)

We express policy iteration in vector-matrix notation to provide a high-level perspective on the key steps of the algorithm.

The Unichain Policy Iteration Algorithm: vector notation

1. **Initialization:** Choose $d' \in D^{\mathrm{MD}}$.
2. **Evaluation:** Obtain g' and $\mathbf{h}'$ by solving
$$\mathbf{h} = \mathbf{r}_{d'} - g\mathbf{e} + (\mathbf{I} - \mathbf{P}_{d'})\mathbf{h} \tag{1.78}$$
3. **Improvement:** Select
$$d'' \in \arg\operatorname*{c\text{-}max}_{d \in D^{\mathrm{MD}}} \{\mathbf{r}_d + \mathbf{P}_d \mathbf{h}'\} \tag{1.79}$$
and set $d'' = d'$ if possible.
4. **Termination:** If $d'' = d'$, stop. Otherwise $d' \leftarrow d''$ and return to step 2.

We restate UPIA in component notation to better describe and implementable algorithm. This description emphasizes that the improvement step is implemented state by state.

The Unichain Policy Iteration Algorithm: component notation

1. **Initialization:** For each $s \in S$, choose a'_s for some $a'_s \in A_s$.
2. **Evaluation:** Obtain g and $h'(s)$ for all $s \in S$ by solving the system of linear equations
$$h(s) = r(s, a'_s) - g + \sum_{j \in S} p(j|s, a'_s) h(j). \tag{1.80}$$
3. **Improvement:** For each $s \in S$ select
$$a''_s \in \arg\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) h'(j) \right\} \tag{1.81}$$
and set $a''_s = a'_s$ if possible.
4. **Termination:** If $a''_s = a'_s$ for all $s \in S$, stop. Otherwise for all $s \in S$, $a'_s \leftarrow a''_s$ and return to step 2.

some comments follow:

1. The evaluation step uniquely determines g and determines $h(s)$ up to a constant. This arbitrary constant does not impact decision rule or action choice in the

subsequent improvement step, however we require a specification to implement it numerically. We prefer setting $h(s') = 0$ for some distinguished state s' so that we solve the model for relative values.

2. Note that the value of g does not enter into the improvement step. Therefore $h(s)$ contains all of the all of the information required to identify a possible improved decision rule

3. Note that this algorithm is not impacted by periodicity since solution is of the evaluation equation is exact.

Example 1.10. We now solve the two-state example from Section **??** using the UPIA. It is easy to see that this model is unichain because state s_2 is absorbing under action $a_{2,1}$.

In implementing it we set $h(s_2) = 0$ for convenience. We emphasize that this choice is arbitrary.

1. Choose $a'_{s_1} = a_{1,2}$ and $a'_{s_2} = a_{2,1}$.

2. The evaluation equations become

$$\begin{aligned} h(s_1) =& 5 - g + h(s_2) \\ h(s_2) =& -5 - g + h(s_2). \end{aligned}$$

Setting $h(s_2) = 0$ shows that $g = -5$ and $h(s_1) = 10$.

3. In the improvement step we have

$$\begin{aligned} a''_{s_1} &= \arg\max_{i=1,2} \left\{ r(s_1, a_{1,i}) + \sum_{j=1,2} p(s_j|s_1, a_{1,i}) h(s_j) \right\} \\ &= \arg\max\{3 + 0.8h(s_1) + 0.2h(s_2), 5 + 0.9h(s_2)\} \\ &= \arg\max\{11, 5\} = a_{1,1} \end{aligned}$$

and

$$\begin{aligned} a''_{s_2} &= \arg\max_{i=1,2} \left\{ r(s_2, a_{2,i}) + \sum_{j=1,2} p(s_j|s_2, a_{2,i}) h(s_j) \right\} \\ &= \arg\max\{-5 + h(s_2), 2 + 0.4h(s_1) + 0.6h(s_2))\} \\ &= \arg\max\{-5, 6\} = a_{2,2} \end{aligned}$$

4. Since $a''_s \neq a'_s$ for all $s \in S$, replace a'_s by a''_s for all $s \in S$ and return to the evaluation step.

5. Evaluating this decision rule using (1.80) together with $h(s_2) = 0$ yields $g = 2\frac{2}{3} = 2.6667$ and $h(s_1) = \frac{5}{3} = 1.6667$. Hence g has increased.

6. Applying the improvement step we find that

$$a''_{s_1} = \arg\max\{4\frac{1}{3}, 5\} = a_{1,2}$$
$$a''_{s_2} = \arg\max\{-5, 2\frac{2}{3}\} = a_{2,2}.$$

7. Since $a''_s \neq a'_s$ for all $s \in S$, replace a'_s by a''_s for all $s \in S$ and return to the evaluation step.

8. Evaluating this decision rule using (1.80) together with $h(s_2) = 0$ yields $g = 2\frac{6}{7} = 2.857$ and $h(s_1) = \frac{15}{7} = 2.143$. Hence g has increased.

9. Applying the improvement step we find that

$$a''_{s_1} = \arg\max\{29.735, 30.147\} = a_{1,2}$$
$$a''_{s_2} = \arg\max\{20.146, 27.941\} = a_{2,2}.$$

10. Since $a''_s = a'_s$ for all $s \in S$, stop.

The above calculations showed that it required two improvement steps to find the optimal policy and an additional improvement step to confirm its optimality. This is because we chose the worst policy to initiate the algorithm. A better starting value would have led to faster convergence.

Observe that the algorithm terminates with:

1. the optimal policy d^*,

2. the optimal gain,

3. the relative values of d^* (or the bias of d^* if we had instead used the side condition $P_{d^*}\mathbf{h} = \mathbf{0}$).

4. the solution of the Bellman equation.

Moreover we saw that g increased from iteration to iteration. This need not be the case in general, at some iterations g may remain the same and h increase. This observation will be the basis for a convergence proof for unichain policy iteration.

1.7.2 Convergence of Unichain Policy Iteration

In this section, we establish that policy iteration finds an optimal policy in a finite number of iterations. We do this by proving the following two results:

1. If there is a strict improvement in a recurrent state of an improving decision rule, then the gain increases.

2. If there is a strict improvement in a transient state of an improving decision rule and no change in any recurrent state, then the gain remains the same but the bias increases.

The consequence of this is that the iterates of UPIA are monotone in the above sense[18]. Note that there are several other ways of proving this result[19], we believe this is the most informative.

Some technical results*

The following proposition is the basis for the first statement above.

Proposition 1.5. Suppose for some $d \in D^{MD}$, g and $\mathbf{h}$ satisfy

$$\mathbf{0} = \mathbf{r}_d - g\mathbf{e} + \mathbf{P}_d\mathbf{h} - \mathbf{h} \tag{1.82}$$

a. Then for any $\delta \in D^{MD}$

$$g^{\delta^\infty}\mathbf{e} = g\mathbf{e} + \mathbf{P}^*_\delta[\mathbf{r}_\delta - g\mathbf{e} + (\mathbf{P}_\delta - \mathbf{I})\mathbf{h}]. \tag{1.83}$$

b. If

$$r(s,\delta(s)) - g + \sum_{j\in S} p(j|s,\delta(s))h(j) - h(s) > 0 \tag{1.84}$$

for some s that is recurrent under δ. Then $g^{\delta^\infty} > g$.

c. If

$$r(s,\delta(s)) - g + \sum_{j\in S} p(j|s,\delta(s))h(j) - h(s) = 0 \tag{1.85}$$

for all s that are recurrent under δ. Then $g^{\delta^\infty} = g$.

Proof. Since $g^{\delta^\infty}\mathbf{e} = \mathbf{P}^*_\delta\mathbf{r}_\delta$, $\mathbf{P}^*_\delta(\mathbf{P}_\delta - \mathbf{I}) = \mathbf{0}$ and $\mathbf{P}^*_\delta\mathbf{e} = \mathbf{e}$, it follows that

$$\begin{aligned} g^{\delta^\infty} =& \mathbf{P}^*_\delta\mathbf{r}_\delta + g\mathbf{e} - g\mathbf{e} + \mathbf{P}^*_\delta(\mathbf{P}_\delta - \mathbf{I})\mathbf{h} \\ =& g\mathbf{e} + \mathbf{P}^*_\delta[\mathbf{r}_\delta - g\mathbf{e} + (\mathbf{P}_\delta - \mathbf{I})\mathbf{h}] \end{aligned}$$

from which part a. follows.

[18] We call this phenomenon *lexicographic* (or dictionary-ordering) monotonicity. What this means that if we write out the vector $(g, \mathbf{h})$ then either the first component g increases or g remains the same and $\mathbf{h}$ increases.

[19] One can use the partial Laurent expansion or the representation for $\mathbf{v}_n = ng\mathbf{e} + \mathbf{h} + \mathbf{o}(n)$.

Parts b. and c. follow immediately from Lemma **??** in the Appendix which states that the components of a limiting matrix $\mathbf{P}^*$ satisfy

$$p^*(j|s) \begin{cases} > 0 \text{ for j recurrent} \\ = 0 \text{ for j transient} \end{cases} .$$

□

Some comments follow:

1. The above proposition combines vector and component notation. We have chosen to do so to make results as clear as possible.

2. Some authors refer to (1.83) as a *comparison* "lemma" because it relates the change in value of the average reward to the one step change of using a different policy.

3. In the context of policy iteration, we will be concerned with choosing δ in part a. according to
$$\delta \in \arg\max_{d' \in D^{MD}} \{\mathbf{r}_{d'} + \mathbf{P}_{d'}\mathbf{h}\}. \tag{1.86}$$

4. Since $\mathbf{p}^*_\delta(j|s) = 0$ when j is transient under δ changes on transient states of δ do not effect the average reward.

5. In a regular model, that is When all states are recurrent under all stationary policies, the above result is all that is required to establish the convergence of policy iteration.

6. This result can be generalized to multi-chain models. See Lemma 9.2.5 in Puterman [1994].

The following proposition provides the basis for the second statement at the beginning of this section.

Proposition 1.6. Suppose for some $d \in D^{MD}$, g and $\mathbf{h}$ satisfy(1.82) and in addition $\mathbf{P}^*_d\mathbf{h} = \mathbf{0}$.

a. Then for any $\delta \in D^{MD}$,
$$\mathbf{h}^{\delta^\infty} = (\mathbf{I} - \mathbf{P}^*_\delta)\mathbf{h} + \mathbf{H}_\delta[\mathbf{r}_\delta - g\mathbf{e} + \mathbf{P}_\delta\mathbf{h} - \mathbf{h}] \tag{1.87}$$

b. Suppose $\delta(s) = d(s)$ for s recurrent under d. Then $\mathbf{h}^{\delta^\infty}_R = \mathbf{h}_R$ and
$$\mathbf{h}^{\delta^\infty}_T = \mathbf{h}_T + [\mathbf{I} - (\mathbf{P}_\delta)_{TT}]^{-1}[\mathbf{r}_\delta - g\mathbf{e} + \mathbf{P}_\delta\mathbf{h} - \mathbf{h}]_T \tag{1.88}$$

where the subscript T (R) denotes the restriction of the vector to its transient (recurrent) states of $\mathbf{P}_\delta$ and the $(\mathbf{P}_\delta)_{TT}$ denotes the sub-matrix of P_δ corresponding to transitions between transient states.

c. Suppose

$$r(s,\delta(s))-g+\sum_{j\in S} p(j|s,\delta(s))h(j)-h(s) \begin{cases} = 0 \text{ for s recurrent under } \delta \text{ and} \\ > 0 \text{ for s transient under } \delta \end{cases} \tag{1.89}$$

Then $g^{\delta^\infty} = g$ and $h^{\delta^\infty}(s) \geq h(s)$ for all $s \in S$ with $h^{\delta^\infty}(s') > h(s')$ for some $s' \in S$.

d. Suppose

$$r(s,\delta(s)) - g + \sum_{j\in S} p(j|s,\delta(s))h(j) - h(s) = 0 \tag{1.90}$$

for all $s \in S$. Then $g^{\delta^\infty} = g$ and $\mathbf{h}^{\delta^\infty} = \mathbf{h}$.

Proof. Since $\mathbf{h}^{\delta^\infty} = \mathbf{H}_\delta \mathbf{r}_\delta$, $\mathbf{H}_\delta \mathbf{e} = \mathbf{0}$ and $\mathbf{H}_\delta(\mathbf{P}_\delta - \mathbf{I}) = \mathbf{P}^*_\delta - \mathbf{I}$ it follows that

$$\begin{aligned} \mathbf{h}^{\delta^\infty} &= \mathbf{H}_\delta \mathbf{r}_\delta + \mathbf{H}_\delta[(\mathbf{P}_\delta - \mathbf{I})\mathbf{h} - (\mathbf{P}_\delta - \mathbf{I})\mathbf{h}] \\ &= (\mathbf{I} - \mathbf{P}^*_\delta)\mathbf{h} + \mathbf{H}_\delta \left[\mathbf{r}_\delta - g\mathbf{e} + (\mathbf{P}_\delta - \mathbf{I})\mathbf{h}\right] \end{aligned}$$

which establishes part a.

Since $\delta(s) = d(s)$ when s is recurrent under d, $\mathbf{P}^*_d = \mathbf{P}^*_\delta$. It follows from the assumption $\mathbf{P}_d\mathbf{h} = \mathbf{0}$ that $\mathbf{P}^*_\delta \mathbf{h} = \mathbf{0}$. From Lemma **??**, $(\mathbf{H}_\delta)_{RT} = \mathbf{0}$ and $(\mathbf{H}_\delta)_{TT} = (\mathbf{I} - (\mathbf{P}_\delta)_{TT})^{-1}$. Hence part b. follows by substituting these three observations into (1.87).

Part c. of the proposition follows from part b. by noting that $(\mathbf{P}_d)_{RR} = (\mathbf{P}_\delta)_{RR}$ applying using the result that $(\mathbf{I} - (\mathbf{P}_\delta)_{TT})^{-1} \geq \mathbf{I}$ from part c. of Lemma **??** in the Appendix.

Part d. follows immediately from part b. □

Some comments about this proposition follow:

1. The additional hypothesis that $\mathbf{P}_d\mathbf{h} = \mathbf{0}$ implies that $\mathbf{h} = \mathbf{h}^{\delta^\infty}$. It is not required to derive (1.87) but is fundamental for deriving (1.88).

2. Note that the increase in $\mathbf{h}$ in part c. occurs on transient states of $\mathbf{P}_\delta$.

3. The above propositions do not account for the eventuality that

$$r(s,\delta(s)) - g + \sum_{j\in S} p(j|s,\delta(s))h(j) - h(s)$$

is positive on some states and negative in others. When δ is derived from d in policy iteration improvement step, this is impossible.

4. It might at first appear that these propositions do not consider the the possibility that δ represents an improvement in **both** recurrent and transient states. However as a consequence of Proposition 1.5, this would result in an increase in the average reward. This is not of concern since it will not impact the proof of convergence of policy iteration below.

Convergence of policy iteration

The following theorem is the main result in this section.

Theorem 1.11. Suppose all deterministic stationary policies are unichain. Then the unichain policy iteration algorithm converges in a finite number of iterations to a solution $(g^*, \mathbf{h}^*)$ of the unichain Bellman equation. Moreover if

$$d^* \in \arg \underset{d \in D^{MD}}{\text{c-max}} \{\mathbf{r}_d + \mathbf{P}_d \mathbf{h}\} \tag{1.91}$$

then $(d^*)^\infty$ is average optimal.

Proof. Our notation follows the vector notation version of the unichain policy iteration algorithm above. Let $d' \in D^{MD}$ clearly

$$\underset{d \in D^{MD}}{\text{c-max}} \{\mathbf{r}_d - g^{(d')^\infty}\mathbf{e} + \mathbf{P}_d \mathbf{h}^{(d')^\infty} - \mathbf{h}^{(d')^\infty}\} \geq \mathbf{r}_{d'} - g^{(d')^\infty}\mathbf{e} + \mathbf{P}_{d'} \mathbf{h}^{(d')^\infty} - \mathbf{h}^{(d')^\infty} = \mathbf{0}.$$

Thus when d'' is chosen in the improvement step using (1.79),

$$\mathbf{r}_{d''} - g^{(d')^\infty}\mathbf{e} + \mathbf{P}_{d''} \mathbf{h}^{(d')^\infty} - \mathbf{h}^{(d')^\infty} \geq \mathbf{0}.$$

If equality holds in the above expression, we have found a solution of the unichain Bellman equation completing the proof. If not, and there is strict inequality in some recurrent state of $\mathbf{P}_{d''}$, $g^{(d'')^\infty} > g^{(d')^\infty}$ from Proposition 1.5. If there is no improvement in a recurrent state and an improvement in a transient state, then $d''(s) = d'(s)$ for all $s \in S$ and Proposition 1.6 implies $g^{(d'')^\infty} = g^{(d')^\infty}$ and $h^{(d'')^\infty}(s') \geq h^{(d')^\infty}$ with **strict** inequality in some state $s' \in S$. Since there are only finitely many stationary deterministic policies, only a finite number of such increases are possible. Thus the result follows. □

Some comments follow:

1. The above result shows UPIA finds a solution of the opitmality equation, an optimal policy and its average reward.

2. Depending on the specification used to uniquely determine $\mathbf{h}$, it finds either the bias or relative values of an optimal policy.

3. If UPIA is implemented with $h(s') = 0$ for some distinguished state s', the above proof still holds because Proposition 1.6 still shows the bias increases when there is a strict improvement only in a transient state.

4. The above theorem also provides a constructive proof of the existence of a solution on the unichain Bellman equation.

5. UPIA need not find the optimal bias. The following example shows that it finds the optimal bias among all policies with the same recurrent class as the policy it chooses at termination.

Example 1.11. Let $S = \{s_1, s_2\}$, $A_{s_1} = \{a_{1,1}, a_{1,2}\}$, $A_{s_2} = \{a_{2,1}\}$; $r(s_1, a_{1,1}) = 2$, $r(s_1, a_{1,2}) = 0$, $r(s_2, a_{2,1}, s_1) = 6$, $r(s_2, a_{2,1}, s_2) = 0$; $p(s_1|s_1, a_{1,1}) = 1$, $p(s_2|s_1, a_{1,2}) = 1$, $p(s_1|s_2, a_{2,1}) = p(s_1|s_2, a_{2,1}) = 0.5$. Note that $r(s_2, a_{2,1}) = 3$ and the model is unichain. See Figure 1.7.

Since actions are only chosen in state s_1, we can specify decision rules by their choice of action in s_1. Let d denote the deterministic decision rule that uses action $a_{1,1}$ and f denote the deterministic decision rule that uses action $a_{1,2}$.

We initiate policy iteration with policy f^∞. Evaluating this policy, we find that $g^{f^\infty} = 2$, $h^{f^\infty}(s_1) = -\frac{4}{3}$ and $h^{f^\infty}(s_2) = \frac{2}{3}$. At the improvement step we find in s_1

$$\max_{a \in A_{s_1}} \{r(s_1, a) + p(s_1|s_1, a)h^{f^\infty}(s_1) + p(s_2|s_1, a)h^{f^\infty}(s_2)\}$$
$$= \max\{2 + h^{f^\infty}(s_1), h^{f^\infty}(s_2)\} = \max\{\frac{2}{3}, \frac{2}{3}\}.$$

Thus UPIA terminates with policy f^∞. But further calculations show that $g^{d^\infty} = 2$, $h^{d^\infty}(s_1) = 0$ and $h^{d^\infty}(s_2) = 2$ so that the bias of d^∞ exceeds that of f^∞. Hence policy iteration finds an average optimal policy but not a policy with optimal bias among average optimal policies. This is because d and f have different sets of recurrent states.

1.8 Linear Programming

We include a discussion of the linear programming formulation of the average reward Markov decision process because:

1. The formulation is more elegant than that in the discounted case because the dual variables represent the probabilities associated with a randomized policy, and

2. The model allows direct inclusion of meaningful and practical constraints.

We first analyze linear programming for regular models and generalize results for unichain models. The multi-chain model remains beyond the scope of this book.

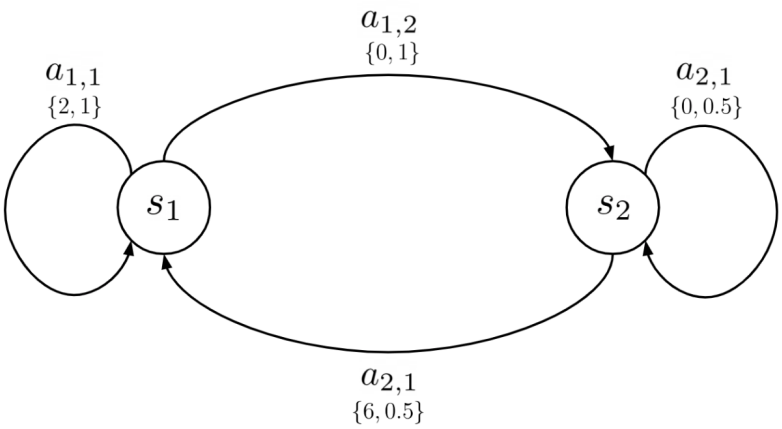

Figure 1.7: Symbolic representation of model in Example 1.11 illustrating that UPIA does not find a bias optimal policy.

1.8.1 Linear programming formulation

We describe a linear program (LP) for regular and unichain models in which every policy has a constant average reward so that the optimal average reward is necessarily constant. As a consequence of part a. of Theorem 1.7 expressed in component notation then if there exist a scalar g and vector $h(s)$ for which for all $s \in S$ and $a \in A_s$

$$h(s) \geq r(s,a) - g + \sum_{j \in S} p(j|s,a)h(j) \tag{1.92}$$

then $g \geq g^*$.

Hence g^* equals the smallest g for which there exists a vector $h(s)$ satisfying (1.92) for all $s \in S$ and $a \in A_s$. This observation gives rise to the following primal linear program:

Primal linear program

Find a scalar g and $h(s)$ for all $s \in S$ so as to

$$\text{Minimize} \quad g$$

and satisfy

$$g + h(s) - \sum_{j \in S} p(j|s,a)h(j) \geq r(s,a) \quad \text{for all}\, a \in A_s,\ s \in S. \tag{1.93}$$

Following our approach in the discounted case we formulate the following dual linear program.

Dual linear program

Find $x(s,a) \geq 0$ for all $s \in S$ and $a \in A_s$ so as to

$$\text{Maximize} \sum_{s\in S}\sum_{a\in A_s} r(s,a)x(s,a) \tag{1.94}$$

and satisfy

$$\sum_{a\in A_j} x(j,a) - \sum_{s\in S}\sum_{a\in A_s} p(j|s,a)x(s,a) = 0 \quad \text{for all } j \in S \tag{1.95}$$

and

$$\sum_{s\in S}\sum_{a\in A_s} x(s,a) = 1. \tag{1.96}$$

Summing (1.95) over $j \in S$ shows that

$$\begin{aligned}\sum_{j\in S}\sum_{a\in A_j} x(j,a) - &\sum_{j\in S}\sum_{s\in S}\sum_{a\in A_s} p(j|s,a)x(s,a) \\ &= \sum_{j\in S}\sum_{a\in A_j} x(j,a) - \sum_{s\in S}\sum_{a\in A_s} x(s,a) = 0.\end{aligned}$$

Hence one of the constraints in (1.95) is redundant and the set of constraints of the dual linear programming have at most $|S|$ linearly independent rows.

We now provide linear programming formulations of the two-state model which appear as Example **??** of Chapter 2.

Example 1.12. The primal LP for the two-state model becomes:

$$\text{Minimize} \quad g$$

subject to

$$\begin{aligned} g + h(s_1) - 0.8h(s_1) - 0.2h(s_2) &\geq 3 \\ g + h(s_1) - h(s_2) &\geq -5 \\ g + h(s_2) - h(s_2) &\geq -5 \\ g + h(s_2) - 0.4h(s_1) - 0.h(s_2) &\geq 2 \end{aligned}$$

Note that the primal LP has 2 variables (one for each state) and 4 constraints (one for each state and action pair) and no non-negativity constraints. Furthermore after simplifying the first and fourth inequalities, we see that all of the inequalities

contain the quantity $h(s_1) - h(s_2)$ implying that $h(s_1)$ and $h(s_2)$ cannot be independently determined. Moreover the third inequality reduces to $g \geq -5$ and does not involve $h(s_1) - h(s_2)$.

The dual LP is given by

$$\text{Maximize} \quad 3x(s_1, a_{1,1}) - 5x(s_1, a_{1,2}) - 5x(s_2, a_{2,1}) + 2x(s_2, a_{2,2})$$

subject to

$$\begin{aligned} x(s_1, a_{1,1}) + x(s_1, a_{1,2}) - 0.8x(s_1, a_{1,1}) - 0.4x(s_2, a_{2,2}) &= 0 \\ x(s_2, a_{2,1}) + x(s_2, a_{2,2}) - 0.2x(s_1, a_{1,1}) - x(s_1, a_{1,2}) - x(s_2, a_{2,1}) - 0.6x(s_2, a_{2,2}) &= 0 \\ x(s_1, a_{1,1}) + x(s_1, a_{1,2}) + x(s_2, a_{2,1}) + x(s_2, a_{2,2}) &= 1 \end{aligned}$$

and $x(s_1, a_{1,1}) \geq 0, x(s_1, a_{1,2}) \geq 0, x(s_2, a_{2,1}) \geq 0, x(s_2, a_{2,2}) \geq 0$.

Note that the dual has 4 variables (one for each state action pair), 2 constraints (one for each state) an additional constraint corresponding to the sum of the dual variables equal one and non-negativity constraints. Further the above constraints can be simplified by combining terms.

As in the discounted case, we analyze these models through the dual LP and provide a nice interpretation for the dual variables.

1.8.2 Regular models

In this section we establish for regular models that:

1. There is a one-to-one relationship between randomized stationary policies and feasible solutions of the dual linear program,
2. There is a one-to one relationship between deterministic stationary policies and basic feasible solutions of the dual linear program, and
3. The dual linear program has a basic feasible solution corresponding to an optimal deterministic policy.

As a result of part c. of Theorem **??**, in a regular model, we can find the stationary distribution $q_d(s)$ of the Markov chain corresponding to Markovian deterministic decision rule d by solving

$$\sum_{j \in S} q(s) p_d(j|s) = q(s) \tag{1.97}$$

$$\sum_{s \in S} q(s) = 1 \tag{1.98}$$

$$q(s) \geq 0 \tag{1.99}$$

for all $s \in S$. where $p_d(j|s) = p(j|s, d(s))$. Moreover when $\mathbf{P}_d$ is regular, each component of $q_d(s) > 0$ for all $s \in S$.

The following theorem relates feasible[20] solutions of the dual LP to the stationary probability that a Markovian randomized decision rule occupies state $s \in S$ and chooses action $a \in A_s$. We remind you that $w_d(a|s)$ denotes the probability that Markovian randomized decision rule d chooses action a in state s and $p_d(j|s) = \sum_{a \in A_s} p(j|s, a) w_d(a|s)$.

Theorem 1.12. Suppose the model is regular.

a. For each $d \in D^{MR}$ define

$$x_d(s, a) := w_d(a|s) q_d(s). \tag{1.100}$$

Then $x_d(s, a)$ is a feasible solution to the dual linear program.

b. i. Let x(s,a) be a feasible solution to the dual linear program. Then for each $s \in S$,

$$\sum_{a \in A_s} x(s, a) > 0. \tag{1.101}$$

ii. Define the decision rule d_x implicitly by

$$w_{d_x}(a|s) := \frac{x(s, a)}{\sum_{a \in A_s} x(s, a)}. \tag{1.102}$$

Then $d_x \in D^{MR}$ and $x_{d_x}(s, a) = w_{d_x}(a|s) q_{d_x}(s)$ is a feasible solution to the dual linear program.

iii. For each $s \in S$ and $a \in A_s$,

$$x_{d_x}(s, a) = x(s, a). \tag{1.103}$$

Proof. Since $\sum_{a \in A_s} p(j|s, a) w_d(a|s) = p(j|s, d(s)) = p_d(j|s)$ and $\sum_{a \in A_s} w_d(a|s) = 1$, substituting $x_d(s, a)$ into the left hand side of (1.95) and noting (1.97) implies for all $s \in S$ that

$$\begin{aligned} \sum_{a \in A_j} w_d(a|s) q_d(s) - \sum_{s \in S} \sum_{a \in A_s} p(j|s, a) w_d(a|s) q_d(s) \\ = q_d(s) - \sum_{s \in S} p_d(j|s) q_d(s) = 0. \end{aligned}$$

Moreover it is easy to see that $x_d(s, a)$ satisfies (1.98) and (1.99) so a. follows.

[20] Recall that a *feasible solution* of a linear program is one that satisfies all of its constraints.

To prove b., define $f(s) := \sum_{a \in A_s} x(s,a)$ and let $S' \subseteq S$ denote the set of $s \in S$ for which $f(s) > 0$. We now show that $S' = S$. Write

$$\begin{aligned}\sum_{s \in S'} \sum_{a \in A_s} p(j|s,a) x(s,a) &= \sum_{s \in S'} \sum_{a \in A_s} p(j|s,a) x(s,a) \frac{f(s)}{f(s)} \\ &= \sum_{s \in S'} \sum_{a \in A_s} p(j|s,a) \frac{x(s,a)}{\sum_{a \in A_s} x(s,a)} f(s) \\ &= \sum_{s \in S'} \sum_{a \in A_s} p_{d_x}(j|s) f(s)\end{aligned}$$

where $d_x(s)$ is defined in (1.102). Hence for all $j \in S'$, from (1.95)

$$f(j) - \sum_{s \in S'} p_{d_x}(j|s) f(s) = 0$$

and

$$1 = \sum_{s \in S} \sum_{a \in A_s} x(s,a) = \sum_{s \in S} f(s) = \sum_{s \in S'} f(s)$$

because $f(s) = 0$ for $s \in S - S'$. This implies that $f(s)$ is the stationary distribution of d_x, that is $f(s) = q_{d_x}(s)$ for all $s \in S'$. Since all stationary policies are regular, $q_{d_x}(s) > 0$ for all $s \in S$. Hence $S' = S$ proving part i. of b.

As a consequence of part 2i., $w_{d_x}(s,a) \geq 0$ is well defined for all $s \in S$ and $a \in A_s$, non-negative and $\sum_{a \in A_s} w_{d_x}(a|s) = 1$. Therefore $w(a|s)$ is a probability distribution on A_s for each $s \in S$ so that d_x is a Markovian randomized decision rule. Hence form part a. $x_{d_x}(s,a)$ satisfies the dual linear program.

Finally

$$x(s,a) = x(s,a) \frac{f(s)}{\sum_{a \in A_s} x(s,a)} = w_{d_x}(a|s) q_{d_x}(s) = x_{d_x}(s,a)$$

where the penultimate equality was established in the proof of part bi. □

Some comments follow. The first two are the most significant.

1. From an application perspective, (1.102) is fundamental. It shows how to derive a randomized stationary policy from a feasible solution of the dual linear program.

2. As a consequence of part biii. of the above theorem, we can interpret $x(s,a)$ as the joint stationary probability that the system corresponding to d_x occupies state s **and** chooses action a in steady state.

3. The proof of part b. above is quite subtle. It exploits the property that in a regular Markov chain, the stationary distribution is strictly positive in all states.

Next we explore the relationship between basic feasible solutions[21] and deterministic stationary policies. The next result follows immediately from the previous theorem and the definition of basic feasible solutions.

Theorem 1.13. Suppose the model is regular.

a. Suppose $d \in D^{MD}$. Then

$$x_d(s,a) := \begin{cases} q_d(s) & \text{when } d(s) = a \\ 0 & \text{when } d(s) \neq a \end{cases} \tag{1.104}$$

is a basic feasible solution of the dual linear program.

b. Let $x(s,a)$ denote a basic feasible solution of the dual linear program.

i. Then $x(s,a) > 0$ for exactly one $a \in A_s$ for each $s \in S$.

ii. For each $s \in S$, $d_x(s) = a$ corresponds to $x(s,a) > 0$ so that $d_x \in D^{MD}$.

Proof. Part a. follows immediately from part *a.* of the preceding theorem.

Since as shown above, the dual has at most $|S|$ linear independent rows a basic feasible solution has at most $|S|$ positive components. Since part bi. of Theorem 1.12 shows that $\sum_{a \in A_s} x(s,a) > 0$ for each $s \in S$, each basic feasible solution has exactly one positive $x(s,a)$ for each $s \in S$ so part bi. holds. Hence from (1.102), $w_{d_x}(a|s) = 1$ for exactly one $a \in A_s$ for each $s \in S$ from which part bii. follows. □

This theorem represents the next building block in analyzing the dual linear program in regular average reward models. We add that:

1. The assumption of a regular model is crucial to this result. We use it to show that $\sum_{a \in A_s} x(s,a) > 0$ for each $s \in S$. In the next section we will consider models with transient states at which $x(s,a) = 0$ for all $a \in A_s$.

2. This theorem (part bii.) shows how to identify deterministic policies from basic feasible solutions. Namely for each state, choose the action for $x(s,a) > 0$.

3. Part a. shows that the solution of the dual gives the stationary probability distribution for the Markov chain corresponding to d_x.

Finally we relate optimal solutions to the dual with optimal deterministic stationary policies.

[21] From the perspective of the following result, we use the result that in a linear program with m constraints and n variables with $n > m$, a *basic feasible solution* has at most m non-zero entries. A formal definition of this concept appears in Appendix **??**.

Theorem 1.14. Suppose the model is regular and $r(s,a)$ is bounded for all $s \in S$, and $a \in A_s$.

Then there exists an optimal basic feasible solution $x^*(s,a)$ to the dual linear program. Moreover $d^\infty_{x^*}$ is an optimal deterministic stationary policy where $d_{x^*}(s) = a$ corresponding to $x^*(s,a) > 0$.

Proof. Since $r(s,a)$ is bounded, Theorem **??** implies that it has an optimal solution. Since any Markovian deterministic decision rule corresponds to a basic feasible solution, Theorem **??** implies it has an optimal basic feasible solution. Denote one such solution by $x^*(s,a)$. Then by the optimality of $x^*(s,a)$ and the relationship between Markovian decision rules in part a. of Theorem 1.12

$$\begin{aligned} g^{(d_{x^*})^\infty} = \sum_{s\in S}\sum_{a\in A_s} r(s,a)x_{d_{x^*}}(s,a) &= \sum_{s\in S}\sum_{a\in A_s} r(s,a)x^*(s,a) \\ &\geq \sum_{s\in S}\sum_{a\in A_s} r(s,a)x_d(s,a) = g^{d^\infty} \end{aligned}$$

for any $d \in D^{MR}$ so that $(d^*)^\infty$ is an optimal policy. □

Note that this theorem guarantees the existence of an optimal stationary deterministic policy but it does not exclude the possibility that there exist multiple optimal policies in which case there exist optimal randomized policies.

Moreover Theorem **??** guarantees the existence of an optimal solution $(g^*, \mathbf{h}^*)$ to the primal LP with with $g^* = \sum_{s\in S}\sum_{a\in A_s} r(s,a)x^*(s,a)$. Hence the above analysis provides an alternative approach to establishing the existence of an optimal solution to the average reward Bellman equations.

1.8.3 Unichain models

The following is the main result for a unichain model. In contrast to the previous section we give only the most important result (without proof).

Theorem 1.15. Suppose the model is unichain and $r(s,a)$ is bounded for all $s \in S$, and $a \in A_s$.

a. Then there exists a bounded optimal basic feasible solution $x^*(s,a)$ to the dual linear program.

b. Let S_{x^*} denote the set of $s \in S$ for which $x^*(s,a) > 0$ for some $a \in A_s$. Then

the policy $d^{\infty}_{x^*}$ defined by

$$d_{x^*}(s) = \begin{cases} a & \text{if } x^*(s,a) > 0 \quad s \in S_{x^*} \\ \text{arbitrary} & \qquad\qquad\quad s \in S - S_{x^*} \end{cases} \tag{1.105}$$

is optimal and S_{x^*} equals its set of recurrent states.

We illustrate this result by solving the dual linear program formulated in Example 1.12.

Example 1.13. Case 1: We rewrite the dual LP as

$$\text{Maximize} \quad 3x(s_1,a_{1,1}) - 5x(s_1,a_{1,2}) - 5x(s_2,a_{2,1}) + 2x(s_2,a_{2,2})$$

subject to

$$\begin{aligned} 0.2x(s_1,a_{1,1}) + x(s_1,a_{1,2}) - 0.4x(s_2,a_{2,2}) &= 0 \\ -0.2x(s_1,a_{1,1}) - x(s_1,a_{1,2}) + 0.4x(s_2,a_{2,2}) &= 0 \\ x(s_1,a_{1,1}) + x(s_1,a_{1,2}) + x(s_2,a_{2,1}) + x(s_2,a_{2,2}) &= 1 \end{aligned}$$

and $x(s_1,a_{1,1}) \geq 0, x(s_1,a_{1,2}) \geq 0, x(s_2,a_{2,1}) \geq 0, x(s_2,a_{2,2}) \geq 0$. Note that $x(s_2,a_{2,1})$ does not appear in the first or second constraint.

Solving this using the simplex algorithm we find that $x(s_1,a_{1,1}) = 0.6667$, $x(s_1,a_{1,2}) = 0$, $x(s_2,a_{2,1}) = 0$ and $x(s_2,a_{2,2}) = 0.3333$ with an objective function value of 2.6667. By Theorem 1.14 this corresponds to stationary policy $d_{x^*}(s_1) = a_{1,1}$ and $d_{x^*}(s_1) = a_{2,2}$.

Case 2: Suppose instead that $r(s_2,a_{2,1}) = 4$. In this case, the optimal solution is $x(s_1,a_{1,1}) = 0, x(s_1,a_{1,2}) = 0, x(s_2,a_{2,1}) = 1, x(s_2,a_{2,2}) = 0$ with an the objective function value of 4. Thus we see that the optimal action in state s_2 is $a_{2,1}$ but since $x(s_1,a_{1,1}) = x(s_1,a_{1,2}) = 0$ part bi. of Theorem 1.12 does not hold, so that (1.102) **does not** specify action choice in state s_1 which is transient when action $a_{2,1}$ is chosen in state s_2.

In Case 1, $S_{x^*} = S$ so the optimal policy has no transient states while in Case 2, $S^{x^*} = \{s_2\}$ so that s_1 is the only recurrent sate under the optimal policy. In the latter case, the optimal average reward is determined only by actions on the recurrent class of each policy so that action choice on transient states is immaterial. What action would you prefer in state s_1?

1.8.4 Constrained Markov decision processes

Because variables in the dual linear program represent the stationary probability that the system is in state $s \in S$ and chooses action $a \in A_s$, it provides a natural framework

for adding constraints to the model. We consider adding one or more constraints of the form:

$$\sum_{s\in S}\sum_{a\in A_s} c(s,a)x(s,a) \leq C. \tag{1.106}$$

The following examples show why such constraints could be useful.

Example 1.14. Consider the queuing service rate control model of Section **??**. Suppose we wish to constrain the percentage of time that the fastest service rate a_f is used to be at most α. Then (1.106) becomes

$$\sum_{s\in S} x(s,a_f) \leq \alpha$$

corresponding to $c(s,a) = 1$ if $a = a_f$ and 0 otherwise.

Example 1.15. Consider the inventory model of Section **??** in which we wish to limit the percentage of time that an order is backlogged[a] to be less than or equal to α . Then (1.106) becomes

$$\sum_{s\leq 0}\sum_{a\in A_s} x(s,a) \leq \alpha$$

corresponding to $c(s,a) = 1$ if $s \leq 0$ and 0 otherwise.

[a]Recall that *backlogging* means an inventory level of 0 or below.

We now show the impact of adding a constraint in the two-state example.

Example 1.16. suppose we add the constraint that the system cannot occupy state s_1 to the linear program analyzed as Case 1 in Example 1.13. This is captured through the constraint

$$x(s_1,a_{1,1}) + x(s_1,a_{1,2}) \leq 0.5.$$

Solving the model with this extra constraint yields $x^c(s_1,a_{1,1}) = 0.375$, $x^c(s_1,a_{1,2}) = 0.125$, $x^c(s_1,a_{2,1}) = 0$ and $x^c(s_1,a_{1,1}) = 0.5$. Since there are 3 non-zero variables, this is not a basic feasible solution for the unconstrained problem, but of course it is a basic feasible solution for the constrained problem. Through (1.102) it generates

the randomized policy $(d_{x^c})^\infty$ with the following action selection probabilities:

$$w_{d_{x^c}}(a_{1,1}|s_1) = \tfrac{0.375}{0.375+0.125} = 0.75, \qquad w_{d_{x^c}}(a_{1,2}|s_1) = \tfrac{0.125}{0.375+0.125} = 0.25,$$

$$w_{d_{x^c}}(a_{2,1}|s_2) = 0, \qquad w_{d_{x^c}}(a_{2,2}|s_2) = 1.$$

Moreover $g^{(d_{x^c})^\infty} = 1.5$ which is considerably less than its value of 2.667 in the unconstrained model. Also, note that when $\alpha \geq 0.6667$ the model has the same solution as the unconstrained model since this constraint is not binding.

The take away from this example is that in a constrained Markov decision process with one constraint the optimal stationary policy may be randomized. In particular, when the constraint is binding as when $\alpha = 0.5$, the policy randomizes in a single state. In general when there are K constraints, the optimal policy will randomize in at most K states.

1.9 Optimality of structured policies

Since the iterates of value iteration diverge, we cannot use the approach of Section **??** directly. Instead, we will base our analysis on the iterates of relative value iteration. To do so, distinguish a particular state s' and form iterates $w^n(s)$ based on the recursions:

$$v^{n+1}(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) w^n(j) \right\} \quad \text{and} \quad w^{n+1}(s) = v^{n+1}(s) - v^{n+1}(s')$$

By arguments above, when the optimal gain is constant such as in a regular, unichain, or communicating model, relative value iteration converges to $w^*(s)$. Since S is finite this is equivalent to convergence in norm so that Theorem **??** on the optimality of structured policies applies with $\mathbf{w}^n$ replacing $\mathbf{v}^n$.

We leave it as an exercise to apply this result to equipment replacement model in Section **??**.

1.10 Queuing Service Rate Control

We consider a infinite horizon average cost version of the discrete time service rate control model described in Section **??** and previously analyzed in the finite horizon (Section **??**) and discounted (Section **??**) cases. We encourage you to read this section carefully as we point out several challenges when solving average reward models.

1.10.1 The model

The controller chooses among three service probabilities $a_1 = 0.2$, $a_2 = 0.4$ and $a_3 = 0.6$ and arrival probability $b = 0.2$. (Note that we require, $b \leq 0.4$, otherwise the row sums

of transition probabilities when using a_3 will exceed 1.)

We assume further that the delay cost is $f(s) = s^2$ and the cost per period of serving at rate a_k is $m(a_k) = 5k^3$. To simplify notation let $c(s, a_k) := f(s) + m(a_k)$. **(note these are diff cost function settings from previous chapter)(I suggest cleaning up Chapter 4 model, its too verbose)**

We truncate the state space at N and assume that the transition probabilities in state N for $k = 1, 2, 3$ are given by:

$$p(j|N, a_k) = \begin{cases} a_k & j = N - 1 \\ 1 - a_k & j = N. \end{cases}$$

Since every state is accessible from every other state (prove it), the model is regular so that the average reward of every policy is constant. Hence the Bellman equation for this model becomes:

$$h(s) = \min_{k=1,2,3} \begin{cases} c(0, a_k) - g + (1 - b)h(0) + bh(1) & s = 0 \\ c(s, a_k) - g + a_k h(s - 1) + (1 - b - a_k)h(s) + bh(s + 1)) & 0 < s < N \\ c(N, a_N) - g + a_k h(N - 1) + (1 - a_k)h(N) & s = N \end{cases}$$

Since $f(0) = 0$ and the choice of service rate has no impact on transition behavior in state 0, the optimal decision in this state is to choose the least expensive service rate, $m(a_1)$. Hence the Bellman equation in state 0 becomes

$$h(0) = m(a_1) - g + (1 - b)h(0) + bh(1).$$

1.10.2 Value iteration

The beauty of this model is that as a result of the simple transition structure, we can avoid writing out the entire transition matrix when implementing value iteration. We find that coding is simplified when we express value iteration recursion in two steps.

First we define $q^n(s, a)$ by

$$q^n(s, a_k) = \begin{cases} c(0, a_k) + (1 - b)v^n(0) + bv^n(1) & s = 0 \\ c(s, a_k) + a_k v^n(s - 1) + (1 - b - a_k)v^n(s) + bv^n(s + 1) & 1 \le s \le N \\ c(N, a_k) + a_k v^n(N - 1) + (1 - a_k)v^n(N) & s = N. \end{cases}$$

Then

$$v^{n+1}(s) = \min_{k=1,2,3} q^n(s, a) \quad \text{for all } s \in S.$$

We implemented value iteration for $N = 20$, $N = 50$ and $N = 200$ with $v^0(s) = 0$ for all $s \in S$ and $\epsilon = 0.0001$. When $N = 20$, convergence required 260 iterations, when $N = 50$, convergence required 350 iterations and $N = 200$ required 796 iterations. In all three cases, $g = 19.4247$. The left-hand image in Figure 1.8 shows the ϵ-optimal policy.

We also implemented relative value setting $w^n(s) = v^n(s) - v^n(0)$ after each iteration. Convergence required the same number steps as above, but as noted in the text $w^n(s)$ converged. The right-hand image in Figure 1.8 depicts $w^n(s)$ at termination.

We solve the model when $N = 51$ using policy iteration in the next section. Since policy iteration finds an optimal policy we will see that the ϵ-optimal policy is indeed optimal and moreover $w^n(s) - h_{rel}(s)$ differ at most by 0.002. Thus relative value iteration well approximates the relative value function.

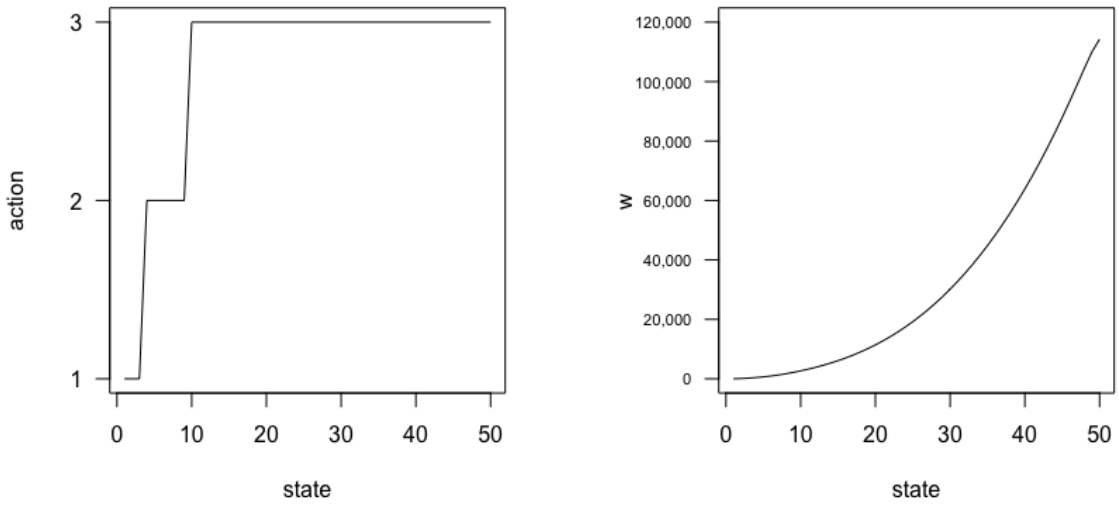

Figure 1.8: The ϵ-optimal stationary policy d^∞ (left) identified by value iteration and relative value iteration and the iterate $w^n(s)$ of relative value iteration at termination for the queuing service rate control model with $N = 50$. Observe that the ϵ-optimal policy is monotone and uses action a_1 for $s = 0, 1, 2$, action a_2 when $s = 3, 4, \ldots, 8$ and action a_3 for $s \geq 9$. Also note that $w^n(s)$ is convex increasing.

1.10.3 Policy iteration

We now solve several instances of this model using policy iteration and discuss results. For each problem size N, we initiate computation with the "silly" stationary policy: use action a_1 in state 0, a_2 in state 1, a_3 in state 2 and repeat this pattern up to state N.

We implemented policy by adding the constraint $h(0) = 0$ so as to uniquely specify $h(s)$ for all $s \in S$. Consequently the evaluation equation can be expressed in matrix notation as

$$\begin{bmatrix} \mathbf{1} & \mathbf{I} - \mathbf{P}_d \\ 0 & \mathbf{u} \end{bmatrix} \begin{bmatrix} g \\ \mathbf{h} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_d \\ 0 \end{bmatrix} \tag{1.107}$$

where **1** denotes an $|S| \times 1$ matrix with all entries equal to 1, 0 a scalar 0 and **u** a $1 \times |S|$ matrix with a 1 in the first component and the remaining entries equal to zero. Note the the last row of this equation corresponds to the constraint $h(0) = 0$[22].

We solved the model for $N = 5, 10, 15, 20, 25$ and 50 and observed that:

1. In all cases policy iteration required 3 iterations to terminate, the relative value function was convex increasing and the optimal policy was increasing in a_k (Figure 1.9).

2. The optimal policy for this model agrees with the optimal policy for a discounted model with $\lambda = 0.999$. This observation is consistent with the logic used to prove Theorem 1.8[23].

3. Note also that the optimal average reward equalled 19.4247 for all $N \geq 20$ suggesting that actions in higher states have have little impact on cost. This is because under the optimal policy, the steady state probability that there are 10 or more jobs in the system is less than 0.001. Consequently the simulation methods discussed below will not accurately estimate relative values in higher occupancy states because they will be visited infrequently.

We also solved a queuing control model with 6 actions and 5000 states using policy iteration. It required 4 iterates and 3.28 minutes on a Mac Book Air.

1.10.4 Linear programming

Formulation of the linear programming model for solution by a solver[24] requires specification of :

1. a matrix representing constraints,

2. the right hand side vector, and

3. objective function coefficients.

Moreover, since our goal is to minimize costs, the dual becomes a minimization instead of a maximization problem[25].

In a model with $S = \{0, 1, \ldots, N-1\}$, there are $3N$ dual variables so that the constraint matrix is $(N+1) \times 3N$. The rows of the matrix correspond to the N constraints involving the transition probabilities plus an additional constraint that the

[22] We found that expressing the matrix in this format convenient for using the solver in R or other software.

[23] This observation is a consequence of Blackwell optimality, namely that there in a finite state and action model there exists a stationary policy that is optimal for all λ sufficiently close to 1.

[24] We use the package lpSolver in R.

[25] As always we can formulate it as a maximization problem with negative rewards.

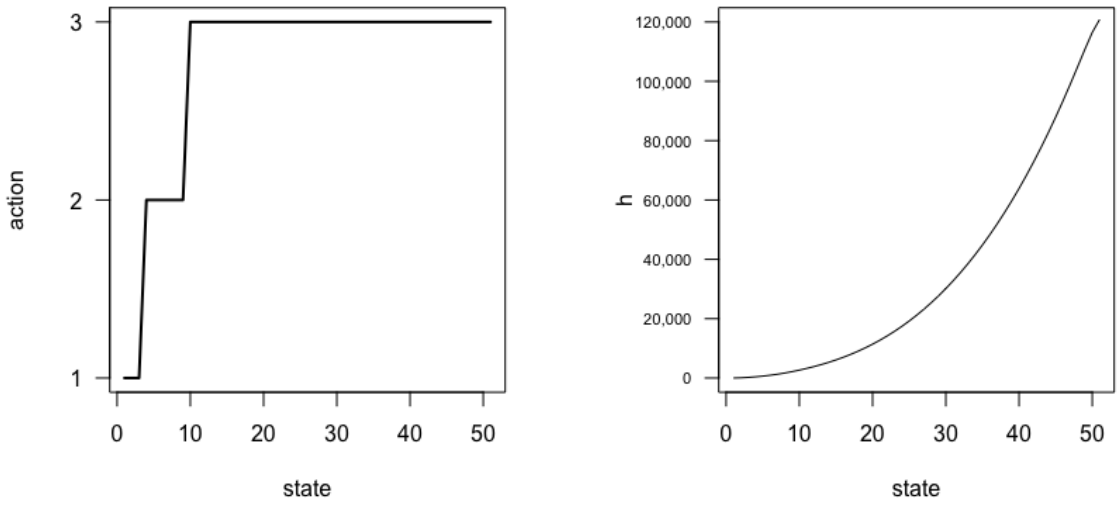

Figure 1.9: The optimal stationary policy d^∞ (left) and relative value function $h(s)$ (right) for the queuing service rate control model with $N = 50$ found using policy iteration. Observe that the optimal policy agrees with the ϵ-optimal policy found using value iteration. Moreover, $h_{rel}(s) \approx w^n(s)$ found using relative value iteration.

sum of the dual variables equals 1. Although one of the first N constraints is redundant, we leave it in letting the solver address it.

We find it convenient to to order the dual variables as $x(0, a_1), \ldots, x(N-1, a_1)$,$x(0, a_2), \ldots, x(N-1, a_2)$, $x(0, a_3), \ldots, x(N-1, a_3)$ so we can construct the constraint matrix by concatenating the matrices corresponding to each action. Let $\mathbf{P}_{a_k}$ denote $N \times N$ transition matrix that uses action a_k in every state. That is

$$\mathbf{P}_{a_k} = \begin{bmatrix} 1-b & b & 0 & 0 & \ldots & 0 \\ a_k & 1-b-a_k & b & 0 & \ldots & 0 \\ 0 & a_k & 1-b-a_k & b & \ldots & 0 \\ \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot & \cdot \\ 0 & \ldots & \ldots & 0 & a_k & 1-a_k \end{bmatrix}.$$

Thus the portion of the constraint matrix representing transition probabilities can be written as

$$\begin{bmatrix} (\mathbf{I}-\mathbf{P}_{a_1})^T & (\mathbf{I}-\mathbf{P}_{a_2})^T & (\mathbf{I}-\mathbf{P}_{a_3})^T \end{bmatrix}.$$

The objective function coefficients expressed as a vector is given by

$$\begin{bmatrix} c(0,a_1) & \ldots & c(N-1,a_1) & c(0,a_2) & \ldots & c(N-1,a_2) & c(0,a_3) & \ldots & c(N-1,a_3) \end{bmatrix}$$

and the right hand side vector equals $(0 \ldots 0 \quad 1)$.

Solving the model with $N = 21$ and noting (1.102) gives the same solution as that obtained by value iteration and policy iteration. Note that when using this solution, the system use action a_3 with probability less than 0.0002. This is because most of the time the system remains in low occupancy states.

If instead we solve the problem with the arrival probability equal 0.35, we find that the average optimal policy $(d^*)^\infty$ with

$$d^*(s) = \begin{cases} a_1 & s = 0, 1 \\ a_2 & s = 2, 3, 4 \\ a_3 & s \geq 5. \end{cases}$$

Because arrivals occur more often the system uses higher service rates at lower occupancy levels than when $b = 0.2$. In fact the system now uses action a_3 with probability 0.195 and optimal average cost equal to 60.27.

To illustrate the use of constraints, suppose we add the constraint that *the system use action a_3 at most* 15% *of the time in steady state.* This corresponds to the constraint

$$\sum_{s \in S} x(s, a_3) \leq 0.15.$$

Solving the model with this constraint has $x(s, a_1) > 0$ for $s = 0, 1$, $x(s, a_2) > 0$ for $s = 2, \ldots, 6$ and $x(s, a_3) > 0$ for $s \geq 6$. Thus since $x(6, a_2) = 0.0158$ and $x(6, a_3) = 0.053$ the optimal policy to the constrained model is similar to the above policy but instead randomizes action choice when $s = 6$. Using (1.102) gives

$$w_{d^*}(a_2|6) = \frac{0.016}{0.016 + 0.053} = .229 \quad w_{d^*}(a_3|6) = 0.771.$$

The optimal average reward to the constrained problem is 60.46 only slightly greater that the unconstrained solution. Since this constraint is binding the system uses action a_3 with probability 0.15.

If instead we use 10% instead of 15%, the optimal policy changes and uses action a_1 with certainty when $s = 0$, a_2 with certainty when $s = 1, \ldots, 6$ and randomizes action choice between a_2 and a_3 when $s = 7$ and chooses a_3 with certainty when $s \geq 8$. Note that $w_{d^{x^*}}(a_2|7) = 0.795$ and the optimal average cost equals 62.85.

When solving larger problems (say $N = 51$) using linear programming we found it challenging to identify the optimal policy because it was challenging to identify true zeroes from those due to underflow, even when we scaled the right hand side of the equality constraint significantly.

Thus we found that the main advantage of using linear programming was the ability to add constraints but this was outweighed by

1. the challenge of identifying the optimal policy from the solution even in moderate sized problems,

2. the work to generate the constraint matrix and

3. need to have an in depth understanding of linear programming to apply it successfully.

1.10.5 Concluding remarks on computation

As current and future trends suggest that Markov decision process users will want to solve large problems, we believe that relative value iteration or policy iteration would be the most attractive approaches for average reward models. Note that although we do not consider it here, modified policy iteration might be the most promising approach since it should require less effort than (relative) value iteration because it avoids many unnecessary maximizations and avoids generating the transition probability matrix that is required to evaluate a policy using policy iteration. We have included a discussion and illustration of linear programming because of its elegant theory and historical significance. Moreover it becomes useful in approximate dynamic programming (Chapter **??**).

1.11 Chapter Appendices

1.11.1 Analysis of Example 1.1*

This rather complicated deterministic finite state example demonstrates that the limit in (1.1) need not exist for a history dependent policy. Let $S = \{s_1, s_2\}$, $A_{s_i} = \{a_{i,1}, a_{i,2}\}$ for $i = 1, 2$, $p(s_1|s_1, a_{1,1}) = 1, p(s_2|s_1, a_{1,2}) = 1, p(s_2|s_2, a_{2,1}) = 1$ and $p(s_1|s_2, a_{2,2}) = 1$ and $r(s_1, a_{1,1}) = r(s_1, a_{1,1}) = 1$ and $r(s_2, a_{2,1}) = r(s_2, a_{2,2}) = 0$.

We describe a policy for which the limit in (1.1) does not exist by creating a series of rewards that has many different subsequential limits. Starting in state s_1 we can construct a policy that chooses a sequence of actions

$$a_{1,2}, a_{2,1}, a_{1,2}, a_{2,1}, a_{1,1}, a_{1,2}, a_{2,2}, a_{2,1}, a_{1,1}, a_{1,1}, a_{1,1}, a_{1,2}, a_{2,2}, a_{2,2}, a_{2,2}, a_{2,1}, \ldots$$

that generates the sequence of rewards

$$1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, \ldots$$

We analyze this formally by defining the expressions:

u_k denotes kth element of this sequence of 1s and 0s.

$q_0 = 0$ and q_i, $i \geq 1$ denotes the number of 1s in the ith block of 1's and 0's.

$s_n = \sum_{i=0}^{n} q_i$ denotes the number of 1s in the sequence up to and including block $n \geq 0$.

$w_n = \sum_{k=1}^{n} u_k/k$ denotes the average of the first n terms for $n \geq 1$ in the sequence.

Using this notation, the pattern of this sequence is as follows. The $(n+1)$st block consists of s_n 1's followed by s_n 0's. In the example above, $q_0 = 0, q_1 = 1, q_2 = 1, q_3 = 2, q_4 = 4, q_5 = 8, \ldots$ and $s_0 = 0, s_1 = 1, s_2 = 2, s_3 = 4, s_4 = 8, s_5 = 16, \ldots$.

We choose two subsequences of $\{u_n\}$ that yield different limits. First choose the subsequence corresponding to the last 0 in each block, that is $\{u_{2s_n} : n \geq 1\}$ or equivalently $u_2, u_4, u_8, \ldots$. Since for this subsequence, $w_n = s_n/2s_n$, its limit is 1/2. This corresponds to the smallest subsequential limit.

Next choose the subsequence corresponding to the last 1 in each block, that is terms $\{u_{2s_n+s_{n+1}} : n \geq 0\}$ or equivalently $u_1, u_3, u_6, u_{12}, \ldots$. The corresponding sequence of w_n's are $1, 1/3, 2/3, 2/3, \ldots$ and the general term is $w_n' = (s_n + s_{n+1})/(2s_n + s_{n+1})$ so that the limit of w_n' =2/3. This corresponds to that largest subsequential limit.

Hence

$$1/2 = \liminf_{n\to\infty} w_n < \limsup_{n\to\infty} w_n = 2/3$$

so that $\lim_{n\to\infty} w_n$ does not exist.

Hence there exists a history dependent policy (corresponding to this sequence of 0s and 1s for which the limit in (1.1) does not exist.

1.11.2 Appendix: Proof of Theorem 1.7

The following proof illustrates several key Markov decision process concepts especially how the Bellman equation which is expressed in terms of Markovian deterministic decision rules, accounts for for randomized history-dependent policies.

Proof. We first prove part a. Let $\pi = (d_1, d_2, \ldots)$ denote an arbitrary policy in Π^{MR}. We will show that for any $N \geq 1$ that

$$Ng\mathbf{e} \geq \sum_{n=1}^{N} \mathbf{P}_\pi^{n-1}\mathbf{r}_{d_n} + (\mathbf{P}_\pi^N - \mathbf{I})\mathbf{h} \tag{1.108}$$

where $\mathbf{P}_\pi^0 = \mathbf{I}$ and $\mathbf{P}_\pi^n = \mathbf{P}_{d_1} \ldots \mathbf{P}_{d_n}$. From this it follows that

$$g\mathbf{e} \geq \limsup_{N\to\infty} \frac{1}{N}\left[\sum_{n=1}^{N} \mathbf{P}_\pi^{n-1} r_{d_n} + (\mathbf{P}_\pi^N - \mathbf{I})\mathbf{h}\right] = \limsup_{N\to\infty} \frac{1}{N}\mathbf{v}_N^\pi = \mathbf{g}_+^\pi$$

where the middle equality follows from the boundedness of $\mathbf{h}$. Thus $g \geq g_+^\pi(s)$ for all $s \in S$ and $\pi \in \Pi^{MR}$. As a consequence of Theorem 1.10, this result is valid for all $\pi \in \Pi^{HR}$ so that $g \geq g_+^*(s)$ for all $s \in S$.

It remains to show that (1.108) holds. As a consequence of Lemma **??**

$$\underset{d\in D^{MD}}{\text{c-max}}\{\mathbf{r}_d - g\mathbf{e} + \mathbf{P}_d\mathbf{h}\} = \underset{d\in D^{MR}}{\text{c-max}}\{\mathbf{r}_d - g\mathbf{e} + \mathbf{P}_d\mathbf{h}\}.$$

Hence rearranging terms in (1.46) and noting the above implies that

$$g\mathbf{e} \geq \mathbf{r}_d + \mathbf{P}_d\mathbf{h} - \mathbf{h} \tag{1.109}$$

for any $d \in D^{MR}$.

Then multiplying

$$g\mathbf{e} \geq \mathbf{r}_{d_2} + \mathbf{P}_{d_2}\mathbf{h} - \mathbf{h}$$

by $\mathbf{P}_{d_1}$ and noting that $\mathbf{P}_{d_1}\mathbf{e} = \mathbf{e}$, it follows that

$$g\mathbf{e} \geq \mathbf{P}_{d_1}\mathbf{r}_{d_2} + \mathbf{P}_{d_1}(\mathbf{P}_{d_2} - \mathbf{I})\mathbf{h}.$$

Adding this together with (1.109) applied to d_1 implies that

$$\begin{aligned} 2g\mathbf{e} &\geq \mathbf{r}_{d_1} + \mathbf{P}_{d_1}\mathbf{r}_{d_2} + (\mathbf{P}_{d_1}\mathbf{P}_{d_2} - \mathbf{P}_{d_1} + \mathbf{P}_{d_1} - \mathbf{I})\mathbf{h} \\ &= \sum_{n=1}^{2} \mathbf{P}_\pi^{n-1}\mathbf{r}_{d_n} + (\mathbf{P}_\pi^2 - \mathbf{I})\underline{h} \end{aligned}$$

Repeating this argument for any $n \geq 2$ or applying induction establishes that (1.109) holds completing the proof of part a.

The proof of part b. is much easier. From (1.47) there exists a $d^* \in D^{MD}$ for which

$$g\mathbf{e} \leq \mathbf{r}^{d^*} + (\mathbf{P}_{d^*} - \mathbf{I})\mathbf{h}$$

Repeating the argument in part a. with the single decision rule d^* and accounting for the reversed direction of the inequality implies that

$$g\mathbf{e} \leq \liminf_{N\to\infty} \frac{1}{N} v_N^{(d^*)^\infty} \leq \sup_{\pi\in\Pi^{HR}} \left\{ \liminf_{N\to\infty} \frac{1}{N} v_N^{d^\infty} \right\} = \mathbf{g}_-^*$$

from which the result follows. Note that since $(d^*)^\infty$ is stationary, the limit in the first inequality exits and equals the "lim inf".

Part c. follows by combining parts a. and b. □

1.11.3 Convergence of value iteration: Proving a key intermediate result

In this appendix we describe the key steps in showing that in an aperiodic unichain model, the limit in (1.68) exists. Above we show that the existence of this limit implies convergence of value iteration. Define the quantity

$$e_n(s) := v^n(s) - ng^* - h^*(s)$$

for all $s \in S$.

The following series of steps lead to demonstrating the convergence of $e_n(s)$.

1. Establish upper and lower bounds on $e_n(s)$ for each $s \in S$. This justifies the use of the result that every bounded infinite series contains a convergent subsequence.

2. Show that $\lim_{n\to\infty} e_n(s)$ exists for s in the recurrent states of an optimal policy. The proof proceeds by showing the $\liminf_{n\to\infty} e_n(s) = \limsup_{n\to\infty} e_n(s)$.

3. Show that the above limit exists on transient states by using a similar proof to that on transient states and then appealing to the result on recurrent states.

Note that this result is valid in multi-chain models as well but requires a subtle intermediate step. Details appear in Section 9.4.1 in Puterman [1994].

1.12 Bibliographic Remarks

Our development closely follows Puterman [1994] but omits many technical issues specifically around multi-chain models.

Howard was the first to study average reward Markov decision problems although there are apparently antecedents in the game theory literature. The monograph Howard [1960], based on his MIT doctoral dissertation, formulates the model, introduces policy iteration and points out that distinct analyses were required for unichain and multi-chain models.

Howard's work motivated the seminal paper Blackwell [1962] which studies the average reward as the limit of the discounted reward as the discount rate approaches 1 and relies heavily on the partial Laurent expansion (see Theorem 1.4). He shows that in finite state and action models multi-chain policy iteration converges and moreover there exists a stationary policy that is optimal for all discount rates sufficiently close to 1, a concept that is now referred to as Blackwell optimality. The classic book Kemeny and Snell [1960] describes the concepts of limiting and fundamental matrices which underlie Blackwell's paper.

As noted above, Howard pointed out the need for distinguishing between unichain and multi-chain models. However Bather [1973] showed that classification excludes many applications. He introduces a class of models referred to as *communicating* which include a wide-range of applications and in which the optimal gain is constant. Platzman [1977] generalizes this concept to the class of weakly communicating models.

Our proof of existence of solutions to the unichain Bellman equation uses concepts in Blackwell [1962]. Derman [1970] and Sennott [1999] provide an alternative approach.

The use of value iteration in average rewards has been extensively studied. Howard [1960] conjectured that the limit in (1.68) exists. We show in Theorem 1.10 that when this limit exists, value iteration converges. However, it required many papers to establish the existence of this limit. Schweitzer and Federgruen [1977] provide a comprehensive analysis of this problem that applies also to countable state models. Our discussion of the aperiodicity transformation as proposed by Schweitzer [1971] follows Section 6.6 in Sennott [1999]. Relative value iteration is due to White [1963].

The use of linear programming in average reward models originates with DeGhelinick [1960] and Manne [1960] who focus on regular models. This work was extended to more general models by Denardo and Fox [1968]. Derman [1970] studies the use of linear programming in constrained models and Kallenberg [1983] provides a comprehensive overview and development in his thesis.

1.13 Exercises

1.13.1 Numerical and Algorithmic Exercises

1. Consider the three state model in exercise **??** of Chapter **??**.

   (a) Verify that the model is regular.

   (b) Give the Bellman equations for this model.

   (c) Find an average ϵ-optimal policy using value iteration and relative value iteration with $\epsilon = .0001$. Compare the number of iterations each require to achieve the same degree of precision.

   (d) Find an average optimal policy using policy iteration (starting with the decision rule that uses a_2 in each state and linear programming.

   (e) For what values of λ does the average optimal policy agree with the discounted optimal policy?

2. Consider the decision rule $d'(s_i) = a_2$ for $i = 1, 2, 3$ in the model in Problem 1. Investigate the quality of the approximation $\mathbf{v}_n \approx ng\mathbf{e}+\mathbf{h}$ and $\mathbf{v}_\lambda \approx (1-\lambda)^{-1}g\mathbf{e}+\mathbf{h}$ for the stationary policy $(d')^\infty$. Show your results graphically for each state.

3. Classify the model in Problem 1 when:

   (a) you add the action $a_{1,3}$ with transition probability $p(s_1|s_1, a_{1,3}) = 1$ and $r(s_1, a_{1,3}) = 2$.

   (b) you in addition add the action $a_{2,3}$ with transition probability $p(s_2|s_2, a_{1,3}) = 1$ and $r(s_1, a_{1,3}) = 4$.

   Use value iteration to find a .0001-optimal policy for the model with the two additional actions. What is the chain structure of the transition probability matrix corresponding to an optimal policy? What is the form of the ϵ-optimal average reward?

4. Show that the limit in (1.1) exists for all stationary policies in Example 1.1.

5. Find .001-optimal and optimal policies for the replacement model in Problem **??**. Compare your results to those in the discounted case.

6. Find .001-optimal and optimal policies for the inventory model in Problem **??**. Compare your results to those in the discounted case.

7. Find an average optimal policy for an infinite horizon reward variant of the grid world navigation problem of Section **??** in which after delivering the coffee (cell 1) or falling down the stairs (cell 7), the robot begin the subsequent decision epoch in the coffee room (cell 13). Assume the system starts in the coffee room. Choose appropriate values for the parameters. How does the optimal policy for this model compare with he optimal solution to the single pass problem.

8. Consider the following **deterministic** stationary model as given in table 1.7:

State	Action	Next state	Reward
s_1	$a_{1,1}$	s_3	4
s_1	$a_{1,2}$	s_2	2
s_2	$a_{2,1}$	s_3	3
s_3	$a_{3,1}$	s_3	0

Table 1.7: Description of transitions and rewards for problem 8.

   (a) Depict the model graphically and classify its states.

   (b) Find all policies that maximize the long run average reward. What does this imply about the long run average reward criterion?

   (c) Find all policies that maximize the expected total reward.

   (d) Find all policies that maximize the discounted reward for λ close to 1.

   (e) Suppose we replace the transition under action $a_{3,1}$ by a zero-reward transition to s_1. Find all policies that maximize the average reward for this modified problem. How is this policy related to those above?

   (f) How does this observation apply to the previous problem?

9. Consider an admission control queuing model (Section **??**) with finite buffer in which at the start of each period a job arrives with certainty and is processed in that period with probability p and not processed with probability $1-p$. Assume further that when the buffer is full, no jobs arrive.

   (a) Show that there exist stationary policies with two or more closed classes.

   (b) Show that the model is communicating.

10. Consider an inventory model with a capacity of 3 units in which the probability that the demand in any period is 1 unit equals p and the probability demand is 0 equals $1-p$. Assume unfilled demand is lost so that $S = \{0, 1, 2, 3\}$.

(a) Show that there exists a stationary deterministic policy which has a transition probability matrix with two closed classes so that the model is multichain.

(b) Show that the model is communicating.

(c) What algorithm would you use to find an optimal policy?

11. Provide an example of a weakly communicating Markov decision process and prove that the optimal reward in a weakly communicating model must be constant.

12. In Example 1.5, show that g and $\mathbf{h}$ as given in the example satisfy the optimality equation.

13. Verify all calculations in Example 1.7. Show that when $r(s_2, a_{2,1}, s_2) = 0$ that the gain and bias of both stationary policies satisfies the Bellman equation.

14. Consider a deterministic decision rule $d \in D^{MD}$ with

$$\mathbf{r}_d = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{and} \quad \mathbf{P}_d = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

(a) Verify that $\mathbf{P}_d$ is periodic. What is its period?

(b) Find g^{d^∞} and $\mathbf{h}^{d^\infty}$ by solving (1.39).

(c) Show through calculations that the result in Corollary 1.4 holds.

(d) Transform $\mathbf{P}_d$ to $\widetilde{\mathbf{P}_d}$ using aperiodicity transformation (1.63).

(e) Provide a graphical representation of the transformed model.

(f) Compare the eigenvalues of $\mathbf{P}_d$ to those of $\widetilde{\mathbf{P}_d}$ when $\tau = .05$. What difference do you observe?

15. Verify all calculations and statements in Example 1.9(ctd.).

1.13.2 Theoretical Exercises

1. Show by induction on n or direct expansion of the left hand side that for $n \geq 1$, $(\mathbf{P}_d - \mathbf{P}_d^*)^n = (\mathbf{P}_d - \mathbf{P}_d^*)^n$.

2. Show that in a communicating model that the there exists a Markovian deterministic decision rule which generates a recurrent chain.

3. Prove that $\mathbf{P}_d^* = \widetilde{\mathbf{P}_d^*}$ where $\widetilde{\mathbf{P}_d^*}$ is defined by (1.63).

4. Assuming that (1.68) is valid, show that all parts of Theorem 1.10 hold.

5. Write out one step of relative value iteration for a fixed decision rule in the form $\mathbf{w}' = \mathbf{Q}\mathbf{v}'$ where $\mathbf{Q}$ is a rank-one matrix that subtracts the first component from all other elements of $\mathbf{v}'$ where $\mathbf{v}' = \mathbf{r}_d + \mathbf{P}_d\mathbf{w}$.

   (a) Show by example that when $\mathbf{P}_d$ is regular, the largest eigenvalue of $\mathbf{Q}\mathbf{P}_d$ is strictly less than one.

   (b) Prove this result.

6. Develop, apply and demonstrate the convergence of Gauss-Seidel and modified policy iteration algorithms for unichain average reward models.

7. Prove that under the average reward criterion there exists an optimal control limit policy in the equipment replacement model in Section **??**.

Bibliography

J.A. Bather. Optimal decision procedures for finite Markov chains, ii. *Adv. Appl. Prob.*, 5:521–540, 1973.

D. Blackwell. Discrete dynamic programming. *Ann. Math. Stat.*, 33:719–726, 1962.

G.T. DeGhelinick. Les problèmes de décisions séquentielles. *Cahiers Centre d'Études Rec. Opér.*, 2:161–179, 1960.

E. Denardo and B. Fox. Multichain markov renewal programs. *SIAM J. Appl. Math.*, 16:468–487, 1968.

C. Derman. *Finite State Markovian Decision Processes*. Academic Press, Newyork, NY., 1970.

R. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

L.C.M. Kallenberg. *Linear Programming and Finite Markovian Control Problems*. Mathematisch Centrum, Amsterdam, 1983.

J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Van Nostrand-Reinhold, New York, 1960.

A. Manne. Linear programming and sequential decisions. *Man. Sci.*, 6:259–267, 1960.

L.K. Platzman. Negative dynamic programming. *Operations Research*, 25:529–533, 1977.

M. L. Puterman. *Markov Decision Processes:Discrete Stochastic Dynamic Programming*. John Wiley & Sons., 1994.

P. Schweitzer. Iterative solution of the functional equations of undiscounted markov renewal programming. *J. Math. Anal. Appl.*, 34:495–501, 1971.

P. Schweitzer and A. Federgruen. Asymptotic behavior of value iteration in markov decision problems. *Mat. Op. Res.*, 2:360–381, 1977.

L. Sennott. *Stochastic Dynamic Programming and the Control of Queueing Systems* . John Wiley and Sons, 1999.

D.J. White. Dynamic programming, markov chains and the method of sucessive approximations. *J. Math. Anal. Appl.*, 6:373–376, 1963.

Chapter 2

Index

1. xx