Chapter 6

Infinite Horizon Models: Expected Total Reward

Starting a long way off the true point, and proceeding by loops and zigzags, we now and then arrive just where we ought to be[1].

From the novel *Middlemarch*, by George Eliot, British novelist, 1819-1880.

6.1 Introduction

In the spirit of the above quote from *Middlemarch*, this chapter focuses on models in which there is a destination that the decision-maker seeks to reach as quickly as possible, or on the other hand avoid as long as possible. These models evaluate policies using their infinite horizon expected total reward defined below. Throughout this chapter such models will be referred to as *expected total reward models*.

[1] Eliot [1874].

Definition 6.1. For each $s \in S$ and $\pi \in \Pi^{\mathrm{HR}}$, define the *infinite horizon expected total reward*[a] by

$$v^{\pi}(s) := \lim_{N\to\infty} E^{\pi}\left[\sum_{n=1}^{N} r(X_n, Y_n) \,\middle|\, X_1 = s\right]. \tag{6.1}$$

[a]This is also referred to as the policy value function.

Upon examining this expression, it is natural to ask "How can such expectations be finite?". Finiteness is achieved by restricting models to those containing zero-reward absorbing states that can be reached in a finite number of transitions by some or all policies. Additional conditions, either explicitly or implicitly stated, ensure that non-terminating policies cannot be optimal.

Using (6.1) to evaluate policies raises a number of technical issues, namely, this quantity may be unbounded or not even exist for some policies π. While many of the challenges arise in countable state[2] models, finite state and action models still require subtle analyses.

Initially, as a result of their intrinsic complexity, expected total reward models were formulated and analyzed by mathematicians. But such models also provided a unified framework to study colourful and significant applications including:

1. optimal stopping,
2. Gridworld models,
3. sequential statistical hypothesis testing,
4. finding the shortest path in a network with random transitions, and
5. gambling strategies.

The recent trend towards using Markov decision process and reinforcement learning methods in *task-oriented episodic* applications such as game playing, gaming, autonomous vehicle guidance, and robotic control, has rekindled interest in such models, underscoring the need for this chapter.

6.1.1 Motivating examples

To appreciate some of the technical complexities faced when using the expected total reward criterion, consider three examples.

First, consider the two-state model in Section 2.5. In this model there are two stationary policies (those choosing action $a_{2,2}$ in s_2) with expected total reward $+\infty$ and two stationary policies (those choosing action $a_{2,1}$ in s_2) with expected total reward

[2]See Chapter 7 in Puterman [1994].

reward $-\infty$. Hence, the expected total reward criterion would not be effective for choosing among these policies regardless of whether the goal is to maximize rewards or minimize costs.

The next example illustrates further pitfalls that arise when using the expected total reward criterion, and how seemingly insignificant changes to the reward function can have significant changes on the optimal value.

Example 6.1. Let $S = \{s_1, s_2\}$, $A_{s_i} = \{a_{i,1}, a_{i,2}\}$ for $i = 1, 2$, $r(s_1, a_{1,1}) = -1$, $r(s_1, a_{1,2}) = 1$, $r(s_2, a_{2,1}) = 0$, $r(s_2, a_{2,2}) = -1$ and $p(s_1|s_1, a_{1,1}) = p(s_2|s_1, a_{1,2}) = p(s_2|s_2, a_{2,1}) = p(s_1|s_2, a_{2,2}) = 1$. See Figure 6.1.

In this model there are four stationary policies: $d_i^\infty, i = 1, 2, 3, 4$ where $d_1 = (a_{1,1}, a_{2,1}), d_2 = (a_{1,1}, a_{2,2}), d_3 = (a_{1,2}, a_{2,1})$ and $d_4 = (a_{1,2}, a_{2,2})$. For the first three

$$\mathbf{v}^{d_1^\infty} = \begin{bmatrix} -\infty \\ 0 \end{bmatrix}, \quad \mathbf{v}^{d_2^\infty} = \begin{bmatrix} -\infty \\ -\infty \end{bmatrix}, \quad \mathbf{v}^{d_3^\infty} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

For d_4^∞, when $X_1 = s_1$,[a]

$$\liminf_{N\to\infty} \sum_{n=1}^{N} r(X_n, d_4(X_n)) = 0 \quad \text{and} \quad \limsup_{N\to\infty} \sum_{n=1}^{N} r(X_n, d_4(X_n)) = 1,$$

and when $X_1 = s_2$,

$$\liminf_{N\to\infty} \sum_{n=1}^{N} r(X_n, d_4(X_n)) = -1 \quad \text{and} \quad \limsup_{N\to\infty} \sum_{n=1}^{N} r(X_n, d_4(X_n)) = 0.$$

If the goal is to maximize expected total reward the decision-maker would prefer d_3^∞. However this example shows that there are stationary policies for which the expected total reward is $-\infty$ and stationary policies for which the limit used to define the expected total reward of a policy does not exist.

Suppose now -1 is subtracted from all rewards so that $r(s, a) \leq 0$ for all states and actions. In the resulting model, all policies have expected total reward $-\infty$. Thus something is clearly special about the existence of policies with absorbing states with zero reward.

[a] Note that since all transitions are deterministic, expectations are not needed for a deterministic policy.

Finally, consider the Gridworld navigation model of Section 3.2. This is an example of what is often referred to as an *episodic* model, one in which the decision process terminates at the end of an episode and the episode length varies among policies. For another concrete example, in the game of golf, an elite player may usually require 4 shots on a par 4 hole but on some occasions may require many more.

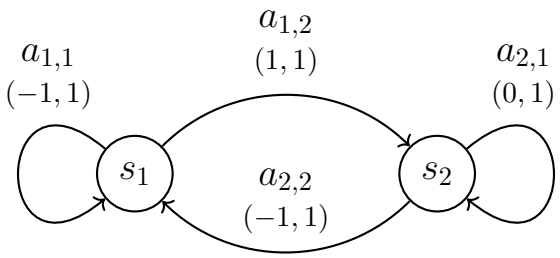

Figure 6.1: Symbolic representation of the model in Example 6.1. The labels (r, p) refer, respectively, to the reward and transition probabilities associated with using the specified action in the given state.

Example 6.2. Consider the Gridworld model of Section 3.2 in which the objective is to deliver coffee from the coffee room (cell 13) to the professor's office (cell 1). In this model, the decision process terminates when the robot successfully delivers coffee, or falls down the stairs (cell 7). The goal in this application is to choose a series of actions for the robot so as to maximize the expected total reward consisting of a positive reward for delivering the coffee, a negative reward (cost) if the robot falls down the stairs and a negative reward (cost) per transition.

In the presence of randomness in the robot's motion in all cells, under any policy, the decision process terminates at a random time so that a finite horizon model with a *fixed* horizon would not apply. Discounting makes little sense in this application since the time scale is so short and averaging is inappropriate because the process terminates when the task is completed so the long-run average reward of any policy would be zero.

On the other hand, when transitions are deterministic there exist polices that never terminate, so that their expected total reward equals $-\infty$. However, even in this case, there are also polices with finite total reward.

The examples above show that the expected total reward of (stationary) policies may be either finite, infinite, or non-existent.

6.2 Model classification

Model classification is based on either:

1. the chain structure of Markov chains generated by stationary policies, or
2. the sign of the rewards.

Historically, expected total reward model research focused on cases with either positive or negative rewards. It did so to ensure that the expected total reward of stationary

policies was well defined. Most of this research concerned establishing the existence of optimal policies in countable state models. A newer and distinct research stream focuses on finite state models with zero-reward absorbing states. Such models are more relevant for modern applications, possess some nice theoretical properties and will be the primary focus of this chapter.

6.2.1 Preliminaries

Throughout this chapter assume stationary rewards and transition probabilities, and without loss of generality, the rewards do not depend on the subsequent state[3].

For each $s \in S$ and $a \in A_s$, define

$$r_+(s,a) := \max\{r(s,a), 0\} \quad \text{and} \quad r_-(s,a) := \min\{r(s,a), 0\}. \tag{6.2}$$

and for each $\pi \in \Pi^{\text{HR}}$ define

$$v_+^\pi(s) := E^\pi \left[\sum_{n=1}^{\infty} r_+(X_n, Y_n) \middle| X_1 = s \right] \tag{6.3}$$

and

$$v_-^\pi(s) := E^\pi \left[\sum_{n=1}^{\infty} r_-(X_n, Y_n) \middle| X_1 = s \right]. \tag{6.4}$$

The limits implicit in defining $v_+^\pi(s)$ and $v_-^\pi(s)$ are always assumed to exist but may be infinite.

Whenever *at least one* of $v_+^\pi(s)$ and $v_-^\pi(s)$ is finite, then $v^\pi(s)$ can be written

$$v^\pi(s) = v_+^\pi(s) + v_-^\pi(s). \tag{6.5}$$

The following example shows that $v^\pi(s)$ may exist, yet need not equal $v_+^\pi(s) + v_-^\pi(s)$ when both quantities are infinite.

Example 6.3. Consider the policy π represented symbolically by the transitions in Figure 6.2. In it all transitions are deterministic with the exception of that from s_3 which goes to the right and left each with probability 1/2. Rewards are 0 except on the arcs from s_4 to s_5 and s_2 to s_1 where they are 2 and -1, respectively. Then its easy to see that $v_+^\pi(s_3) = +\infty$, $v_-^\pi(s_3) = -\infty$ but $v^\pi(s_3) = +\infty$ so that (6.5) does not hold.

Note that if there was another policy π' which had similar transitions as above but with the reward on the arc from s_4 to s_5 equal to 3, then $v^{\pi'}(s_3) = +\infty$ but clearly π' is preferable to π. Unfortunately, the expected total reward criterion

[3] If they do, replace $r(s,a,j)$ by $r(s,a) = \sum_{j \in S} r(s,a,j)p(j|s,a)$. Simulation models are best analyzed in terms of $r(s,a,j)$ since their purpose is to avoid computing this expectation.

does not distinguish these two policies. Either the discounted reward criterion (Chapter 5) or average reward criterion (Chapter 7) would be appropriate in such situations.

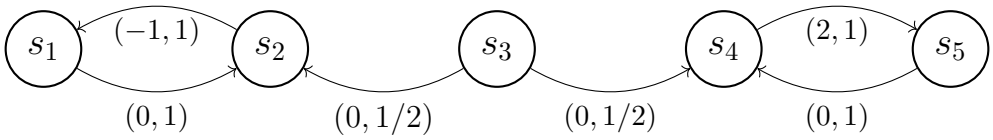

Figure 6.2: Symbolic representation of policy π analyzed in Example 6.3. The labels (r, p) refer, respectively, to the reward and transition probabilities associated with using the specified action in the given state.

To exclude such pathologies and simplify analyses the following assumption will be imposed throughout the remainder of this chapter and frequently be restated for emphasis:.

Assumption 6.1. For each $\pi \in \Pi^{\text{HR}}$ and $s \in S$, either:

1. $v_+^\pi(s)$ is finite,
2. $v_-^\pi(s)$ is finite, or
3. both $v_+^\pi(s)$ or $v_-^\pi(s)$ are finite.

The most significant consequence of this assumption is that when it holds, the limit in (6.1) exists and moreover is finite under condition 3. Furthermore, it excludes models such as those in Example 6.3.

The next four subsections describe the following expected total reward model classes:

- Transient
- Stochastic shortest path
- Positive
- Negative.

6.2.2 Transient models

First, distinguish the class of *transient*[4] models. Appendix B provides an overview of relevant Markov chain concepts and results.

Definition 6.2. A *transient model* is characterized by the existence of an absorbing state $\Delta \in S$ with $A_\Delta = \{\delta\}$ such that :

1. $r(\Delta, \delta) = 0$ and $p(\Delta|\Delta, \delta) = 1$;
2. for every stationary policy d^∞,

$$\lim_{n\to\infty} P^{d^\infty}[X_n = \Delta | X_1 = s] = 1 \tag{6.6}$$

for all $s \in S$.

The only assumptions on rewards in a transient model are that $r(\Delta, \delta) = 0$. Consequently, a transient model can have both positive and negative rewards. The condition that there is a single state Δ satisfying conditions 1 and 2 can be relaxed as follows:

1. The quantity Δ may represent a *set* of absorbing states with the property that for each $s' \in \Delta$, $A_{s'}$ contains a single action a' for which $r(s', a') = 0$ and $p(s'|s', a') = 1$.
2. The set of absorbing states may vary by stationary policy. This can be reduced to the above definition by adding a zero-reward transition from each of these absorbing states to a common absorbing state Δ.

The most important consequence of the transient model definition is that under every stationary policy the corresponding Markov chain is *unichain* with a *single* absorbing state Δ and a set of transient states $S \setminus \Delta$. This means that the transition probabilities and rewards corresponding to decision rule d can be written as

$$\mathbf{P}_d = \begin{bmatrix} \mathbf{Q}_d & \mathbf{R}_d \\ \mathbf{0} & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{r}_d = \begin{bmatrix} (\mathbf{r}_d)_T \\ 0 \end{bmatrix} \tag{6.7}$$

where $\mathbf{Q}_d$ is an $|S| - 1 \times |S| - 1$ matrix that represents the probability of transitions between transient states in $S \setminus \Delta$ and $\mathbf{R}_d$ represents the probability of transitions from $S \setminus \Delta$ to Δ. The last row of the matrix corresponds to transition from Δ. Since it

[4]The expression *transient model* originates with Veinott [1969a]. However, he defines this concept in terms of transition probability matrices of stationary policies and does not explicitly distinguish the set Δ of absorbing states. Derman [1970] referred to such models as *optimal first passage models*. Others have referred to them as *contracting models*.

is absorbing, states in $S \setminus \Delta$ cannot be reached from it. Instead the system remains in Δ in perpetuity, corresponding to the entry 1 in the lower right-hand corner of the matrix.

The reward vector can be partitioned into two parts, with $(\mathbf{r}_d)_T$ representing the reward in the transient states $S \setminus \Delta$ and a reward of 0 in state Δ. For convenience, let $(\mathbf{P}_d)_T := \mathbf{Q}_d$, the sub-matrix of $\mathbf{P}_d$ corresponding to the transient states.

Key properties of transient models

The following proposition accumulates key properties of $\mathbf{Q}_d$ in a transient model. Recall that the spectral radius of a finite square matrix $\mathbf{Q}$ denoted by $\sigma(\mathbf{Q})$ is its largest eigenvalue in absolute value.

Proposition 6.1. In a transient model for every $d \in D^{\text{MD}}$:

1. The *spectral radius* of $\mathbf{Q}_d$ satisfies
$$\sigma(\mathbf{Q}_d) < 1. \tag{6.8}$$
2. $(\mathbf{I} - \mathbf{Q}_d)^{-1}$ exists and satisfies
$$(\mathbf{I} - \mathbf{Q}_d)^{-1} = \sum_{n=0}^{\infty} \mathbf{Q}_d^n. \tag{6.9}$$
3. For $\mathbf{u} \geq \mathbf{0}$, $(\mathbf{I} - \mathbf{Q}_d)^{-1}\mathbf{u} \geq \mathbf{u}$.

Proof. The first result follows from Veinott [1969a] or p. 223 in Berman and Plemmons [1979].

The second part follows directly from the first part by Lemma B.2. Alternatively, in a transient model the system reaches state Δ with certainty under each stationary policy, so that[5] $\mathbf{Q}_d^n \to \mathbf{0}$. Consequently it follows from Lemma B.1 that $(\mathbf{I} - \mathbf{Q}_d)^{-1}$ exists.

To prove the third part, since $\mathbf{Q}_d \geq \mathbf{0}$, it follows from (6.9)
$$(\mathbf{I} - \mathbf{Q}_d)^{-1}\mathbf{u} = \mathbf{u} + \mathbf{Q}_d\mathbf{u} + \mathbf{Q}_d^2\mathbf{u} + \ldots \geq \mathbf{u}.$$
□

Consequently in a transient model, the value of a stationary policy can be expressed in closed form as:
$$\mathbf{v}^{d^\infty} = \sum_{n=1}^{\infty} \mathbf{P}_d^{n-1}\mathbf{r}_d = \sum_{n=1}^{\infty} \mathbf{Q}_d^{n-1}(\mathbf{r}_d)_T = (\mathbf{I} - \mathbf{Q}_d)^{-1}(\mathbf{r}_d)_T. \tag{6.10}$$

[5] Otherwise $S \setminus \Delta$ would be a closed class and not transient.

One might hypothesize that the condition that $\sigma(\mathbf{Q}_d) < 1$ for all $d \in D^{\text{MD}}$ could serve as a definition of a transient model, since this condition underpins many key results for transient models. The downside of doing this is that it admits models in which transition matrix row sums could exceed one, which challenges the probabilistic interpretation of a transition matrix.

Define the *absorption time* N_Δ as follows.

Definition 6.3. The random variable N_Δ defined by

$$N_\Delta := \inf\{n \geq 1 \,|\, X_n = \Delta\} \tag{6.11}$$

is referred to as an *absorption time* or *effective horizon*[a].

[a] Absorption times are often referred to as first passage times or stopping times in the probability literature. It is often referred to as the *effective horizon* in the reinforcement learning literature.

As a consequence of (6.6) or directly from (6.9) the following holds.

Proposition 6.2. In a transient model

$$E^{d^\infty}[N_\Delta \,|\, X_1 = s] < \infty \tag{6.12}$$

for all $s \in S \setminus \Delta$.

Applying this result gives the following:

Theorem 6.1. In a transient model, the expected total reward of every stationary policy exists and is finite.

Proof. Since

$$v^{d^\infty}(s) = E^{d^\infty}\left[\sum_{n=1}^{N_\Delta} r(X_n, d(X_n)) \middle| X_1 = s\right].$$

it follows from Proposition 6.2 that

$$-\infty < E^{d^\infty}[N_\Delta | X_1 = s] r_{\min} \leq v^{d^\infty}(s) \leq E^{d^\infty}[N_\Delta | X_1 = s] r_{\max} < \infty,$$

where

$$r_{\min} = \min_{a \in A_s, s \in S} r(s,a) \quad \text{and} \quad r_{\max} = \max_{a \in A_s, s \in S} r(s,a).$$

The finiteness of $r_{\min}$ and $r_{\max}$ are a consequence of the assumption of a finite state and action model. □

Note that Veinott [1969a] showed this result holds for all Markovian policies by establishing that when all stationary policies are transient, then all policies are transient.

As noted in Section 2.24, a discounted model may be regarded as a transient model by interpreting discounting as random termination at a geometrically distributed stopping time independent of the policy. Conversely, transient models inherit many of the properties of discounted reward models.

6.2.3 Stochastic shortest path models

Stochastic shortest path[6] (SSP) models generalize transient models by allowing the possibility that some stationary policies do not reach a zero-reward absorbing state with certainty. However they exclude the possibility that such a policy can be optimal by requiring that the value of any such policy be equal to $-\infty$ in some state.

Definition 6.4. A *stochastic shortest path model* is characterized by an absorbing state Δ with a single action $A_\Delta = \{\delta\}$ for which:

1. $r(\Delta, \delta) = 0$ and $p(\Delta|\Delta, \delta) = 1$,
2. there exists at least one stationary policy d^∞ for which
$$\lim_{N\to\infty} P^{d^\infty}[X_N = \Delta | X_1 = s] = 1$$
for all $s \in S$, and
3. if
$$\lim_{N\to\infty} P^{(d')^\infty}[X_N = \Delta | X_1 = s] < 1 \tag{6.13}$$
for some stationary policy d' and $s \in S$, then $v^{(d')^\infty}(s) = -\infty$.

A stationary policy is *proper*[7] if it satisfies

$$\lim_{N\to\infty} P^{d^\infty}[X_N = \Delta | X_1 = s] = 1.$$

If (6.13) holds, the policy is said to be *improper*. Thus, an SSP generalizes a transient model by allowing improper policies, but adds the condition that the expected total reward of an improper policy must exist and have value $-\infty$ for at least one state[8].

[6] The nomenclature stochastic shortest path model and the associated terminology originates with Bertsekas and Tsitsiklis [1991] in the context of cost minimization models. Since the models are formulated in terms of rewards, they may be thought of as stochastic longest path models. Nevertheless, they will continue to be referred to as stochastic shortest path models to be consistent with the literature.

[7] Following the terminology of Bertsekas and Tsitsiklis [1991].

[8] This excludes models with policies that can loop between two states and have rewards 1 and -1 respectively, as in Example 6.1 so that the limit defining $v^{(d')^\infty}$ does not exist.

Consequently, an improper policy cannot be optimal. Note that improper policies will contain closed classes of states in which there are negative rewards. On the other hand, an SSP model in which every stationary policy is proper is a transient model.

6.2.4 Positive models

A discussion of positive and negative models is included to relate results herein to the classical expected total reward literature. Moreover, derivations of results for these models provide additional insight into the structure of Markov decision process models.

Positive and negative models are based on the decomposition of $v^\pi(s)$ in (6.5). Positive models were first formulated as those in which $r(s,a) \geq 0$ for all $a \in A_s, s \in S$. The following definition is more general and isolates the critical components of the model.

Definition 6.5. A *positive model* satisfies:

1. $v_+^\pi(s) < \infty$ for all $s \in S$ and $\pi \in \Pi^{\text{HR}}$, and
2. for each $s \in S$, $r(s,a) \geq 0$ for some $a \in A_s$.

The first condition ensures that there are no policies with infinite expected total reward. Thus, in a positive model it is possible to distinguish among policies with non-negative expected total reward. The second condition ensures that there is at least one (stationary) policy with non-negative expected total reward. Note that in this model, some policies may have $v^\pi(s) = -\infty$ but it follows from (6.5) that the first condition in the definition of a positive model excludes the possibility that the limit defining the expected total reward does not exist.

6.2.5 Negative models

Negative models were first formulated as ones in which $r(s,a) \leq 0$ for all $a \in A_s, s \in S$. The following definition is more general.

Definition 6.6. A *negative model* satisfies:

1. $v_+^\pi(s) = 0$ for all $s \in S$ and $\pi \in \Pi^{\text{HR}}$, and
2. there exists a $\pi \in \Pi^{\text{HR}}$ for which $v_-^\pi(s) > -\infty$ for all $s \in S$.

The first condition ensures that $v^\pi(s) \leq 0$ for all policies and the second condition ensures there is at least one policy with a finite non-positive value. Observe also that in a negative model, the limit defining $v^\pi(s)$ exists for all $\pi \in \Pi^{\text{HR}}$. Note that a subclass

of models with non-negative rewards in some states can be transformed to a negative model. This point is expanded on below in the context of the Gridworld navigation model in Example 6.4.

6.2.6 Comparison of model classes

At first glance it might appear that by changing signs of rewards, positive and negative models are equivalent. This is not the case because with an objective of maximizing the expected total reward in positive models, the decision maker seeks a policy with its expected total reward as far above zero as possible while in a negative model the decision maker seeks a policy with its expected total reward as close to zero as possible. By regarding negative rewards as costs, maximizing the expected total reward in a negative model is equivalent to minimizing the expected total cost.

Note that the classifications positive and negative model are *not* mutually exclusive as shown by the model in Figure 6.3, which satisfies both model definitions.

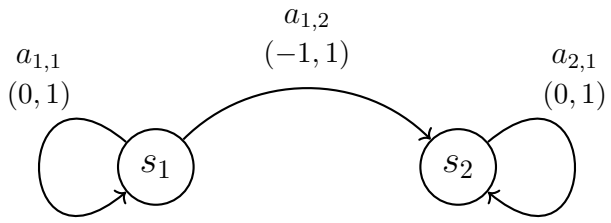

Figure 6.3: Symbolic representation of a model that is both negative and positive. The labels (r, p) refer, respectively, to the reward and transition probabilities associated with using the specified action in the given state.

Transient and SSP models allow more general cost structures than positive and negative models, however there are two special cases to consider:

1. When all rewards are less than or equal to zero, transient and SSP models are negative models.

2. An SSP model in which at least one stationary policy has a non-negative reward is a positive model.

Examples

The following two examples provide concrete illustrations of the above model classes.

Example 6.4. Recall that in the Gridworld model, there is a reward of $B > 0$ for successfully delivering the coffee cup, a cost of $X > 0$ (reward of $-X$) for falling down the stairs and a cost of c (reward of $-c$) for each transition. Since the (expected) positive reward for delivering the coffee can only be obtained from cells 2 and 4 (see Figure 3.3 in Section 3.2), the condition that for each $s \in S$ there exists an $a \in A_s$ for which $r(s,a) \geq 0$ does not hold. Moreover the condition that $v_+^\pi(s) = 0$ in the negative model definition does not hold. Hence this model is neither positive nor negative.

However, this model can be transformed into a negative model by subtracting B from all rewards (or just from the reward associated with delivering the coffee and adding it to the penalty for falling down the stairs). On the other hand, two special cases are noteworthy: 1) when $B = 0$, this is negative model; 2) when $X = 0$ and there is no transition cost, the model becomes a positive model. If $B = 1$ and all other costs are 0, maximizing the expected total reward is equivalent to maximizing the probability that the robot succeeds at its task of delivering the coffee. So, positive models include those that maximize the probability of reaching a particular set of states.

With regards to transient and SSP models, there are two zero-reward absorbing states, so the model does not satisfy the first condition in the definition of a transient model. However, by adding a zero-reward absorbing state Δ and deterministic zero-reward transitions from cells 1 and 7 to Δ, the first transient model condition is satisfied. When transitions are random, all stationary policies are proper so that the model is transient. When transitions are deterministic, infinite loops are possible so that there are improper stationary policies. However, since each transition has a reward of $-c$, when $c \neq 0$ such improper policies have expected total reward equal to $-\infty$ so that it is an SSP model. Thus, in this example, the SSP classification allows deterministic actions while the transient model classification does not.

Example 6.5. Classifying the model in Example 6.1 is more problematic. Because $v_+^{d_4^\infty}(s) = +\infty$ and $v_-^{d_4^\infty}(s) = -\infty$ the model is neither positive nor negative. Subtracting 1 from all rewards satisfies condition 1 for a negative model, but after this transformation, $v_-^\pi(s) = -\infty$ for each stationary policy so condition 2 does not hold. Moreover it is not an SSP model because the improper policy d_3^∞ does not have an expected total reward equal to $-\infty$.

6.3 Value functions, optimal policies and the Bellman equation

Suppose that $v^\pi(s)$ exists for all $\pi \in \Pi^{\text{HR}}$ and $s \in S$. As noted above, a sufficient condition for this to hold is given by Assumption 6.1, namely that either $v^\pi_+(s)$, $v^\pi_-(s)$ or both are finite. This assumption ensures that $v^\pi(s)$ is well-defined, although it may be infinite. It is important to emphasize that:

Assumption 6.1 is satisfied in transient, SSP, positive, and negative models.

More specifically:

1. In transient models $v^\pi(s)$ is finite for all policies.
2. In SSP, positive and negative models, $v^\pi(s)$ exists but may equal $-\infty$ for some policies. Other assumptions ensure that such a policy cannot be optimal.

Given a model for which this assumption is satisfied, define its optimal value function $v^*(s)$ by

$$v^*(s) := \sup_{\pi \in \Pi^{\text{HR}}} v^\pi(s). \tag{6.14}$$

for all $s \in S$. Equivalently, $v^*(s)$ is the "smallest" function that satisfies $v(s) \geq v^\pi(s)$ for all $\pi \in \Pi^{\text{HR}}$ and $s \in S$.

What this condition means is that given $\epsilon > 0$, for each $s \in S$ there exists a $\pi^* \in \Pi^{\text{HR}}$ (which can vary with s) for which

$$v^{\pi^*}(s) \geq v^*(s) - \epsilon.$$

Consequently it follows that:

Theorem 6.2. The optimal value function $v^*(s)$ is finite in transient, SSP, positive and negative models.

The following proposition establishes that for each state, the supremum over history-dependent randomized policies equals the supremum over Markovian randomized policies. This was proved in Chapter 5 using the key Lemma 5.9, which showed that for any history-dependent randomized policy, for each state there exists a Markovian randomized policy with the same transition probabilities. The same result applies here.

Proposition 6.3. For each $s \in S$,

$$v^*(s) = \sup_{\pi \in \Pi^{\text{MR}}} v^\pi(s). \tag{6.15}$$

Definition 6.7. A policy π^* is *optimal* if

$$v^{\pi^*}(s) \geq v^\pi(s) \tag{6.16}$$

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

This definition along with the previous proposition allows the decision maker to restrict the search for optimal policies to the class of Markovian randomized policies.

A technical aside*

When the limit defining $v^\pi(s)$ does not exist for some $\pi \in \Pi^{\text{HR}}$ the above definitions can be modified as follows. Define $v^*(s)$ to be the smallest function $v(s)$ that satisfies

$$v(s) \geq \limsup_{N\to\infty} E^\pi \left[\sum_{n=1}^{N} r(X_n, Y_n) \middle| X_1 = s \right] \tag{6.17}$$

for all $\pi \in \Pi^{\text{HR}}$. Using this definition, it follows that $v^*(s)$ is well defined in Example 6.1.

Moreover, policy $\pi^* \in \Pi^{\text{HR}}$ is *optimal* if

$$\liminf_{N\to\infty} E^{\pi^*} \left[\sum_{n=1}^{N} r(X_n, Y_n) \middle| X_1 = s \right] \geq \limsup_{N\to\infty} E^\pi \left[\sum_{n=1}^{N} r(X_n, Y_n) \middle| X_1 = s \right] \tag{6.18}$$

for all $\pi \in \Pi^{\text{HR}}$. This definition establishes the optimality of d_3^∞ in Example 6.1.

6.3.1 The Bellman equation in an expected total reward model

This section investigates properties of the Bellman equation for expected total reward models. In component notation, the Bellman equation can be expressed as:

$$v(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v(j) \right\} \quad \text{for all } s \in S. \tag{6.19}$$

In vector notation it is given by:

$$\mathbf{v} = \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\{\mathbf{r}_d + \mathbf{P}_d \mathbf{v}\}. \tag{6.20}$$

Observe that the Bellman equation in expected total reward models is identical to that in a discounted model with $\lambda = 1$.

In the special case that D^{MD} consists of a single decision rule d, (6.19) reduces to the (deterministic stationary) policy evaluation equation

$$v(s) = r(s, d(s)) + \sum_{j \in S} p(j|s, d(s))v(j). \tag{6.21}$$

In vector notation (6.21) reduces to the policy evaluation equation

$$\mathbf{v} = \mathbf{r}_d + \mathbf{P}_d \mathbf{v}. \tag{6.22}$$

In greater generality[9] this expression becomes:

$$v(s) = E^d \left[r(X, Y, X') + v(X') \middle| X = s \right]. \tag{6.23}$$

Recalling that V denotes the set of real valued functions on S, define the *Bellman operator*, $L : V \to V$ and the *policy evaluation* operator, $L_d : V \to V$ by

$$L\mathbf{v} := \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\{\mathbf{r}_d + \mathbf{P}_d \mathbf{v}\} \quad \text{and} \quad L_d \mathbf{v} := \mathbf{r}_d + \mathbf{P}_d \mathbf{v}. \tag{6.24}$$

Consequently, the Bellman equation and policy evaluation equations can be written in vector notation as:

$$\mathbf{v} = L\mathbf{v} \quad \text{and} \quad \mathbf{v} = L_d \mathbf{v}. \tag{6.25}$$

Noting that Lemma 5.4 holds when $\lambda = 1$ enables restricting attention to Markovian deterministic decision rules in the definition of operator L.

6.3.2 The transient Bellman equation

In transient models, the Bellman equation is only of interest on the set of transient states $S \setminus \Delta$, since the definition of a transient model implies $v^{d^\infty}(\Delta) = 0$ for any $d \in D^{\mathrm{MD}}$. Let $\mathbf{v}_T$ denote the vector of components of $\mathbf{v}$ restricted to transient states

[9] Allowing rewards to depend on the subsequent state and Markovian randomized decision rules. Such a representation will be useful in Chapters 10 and 11.

and let V_T denote the set of bounded functions on $S \setminus \Delta$. Then, using the matrix partitioning in (6.7), the Bellman equation can be expressed as

$$\begin{bmatrix} \mathbf{v}_T \\ v(\Delta) \end{bmatrix} = \underset{d \in D^{\mathrm{MD}}}{\text{c-max}} \left\{ \begin{bmatrix} (\mathbf{r}_d)_T \\ 0 \end{bmatrix} + \begin{bmatrix} \mathbf{Q}_d & \mathbf{R}_d \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{v}_T \\ v(\Delta) \end{bmatrix} \right\}. \tag{6.26}$$

Although (6.26) does not explicitly determine $v(\Delta)$, it is obvious in such models that $v(\Delta) = 0$. Substituting $v(\Delta) = 0$ into the above equation gives:

$$\mathbf{v}_T = \underset{d \in D^{MD}}{\text{c-max}} \{(\mathbf{r}_d)_T + \mathbf{Q}_d \mathbf{v}_T\}. \tag{6.27}$$

Refer to (6.27) as the *transient Bellman equation* and define the *transient Bellman operator* $L_T : V_T \to V_T$ by

$$L_T \mathbf{v} := \underset{d \in D^{\mathrm{MD}}}{\text{c-max}} \{(\mathbf{r}_d)_T + \mathbf{Q}_d \mathbf{v}_T\}. \tag{6.28}$$

Then the transient Bellman equation can be represented in operator form by

$$\mathbf{v}_T = L_T \mathbf{v}_T. \tag{6.29}$$

By replacing D^{MD} with a single decision rule d, the policy evaluation equation in a transient model, $(L_d)_T$, reduces to

$$\mathbf{v}_T = (\mathbf{r}_d)_T + \mathbf{Q}_d \mathbf{v}_T = (L_d)_T \mathbf{v}_T. \tag{6.30}$$

It will be established below that as a consequence of the observation that at least one row sum of $\mathbf{Q}_d$ is strictly less than 1 in a transient model, the transient Bellman operator is a contraction mapping so that it can be analyzed using a similar approach as for discounted models.

6.3.3 Solutions of the Bellman equation in expected total reward models

This section establishes properties of the Bellman operator that apply for any model class. It will show that:

1. the operator L is monotone,
2. the optimal value function satisfies the Bellman equation, and
3. the solution of the Bellman equation is not unique.

In contrast to the discounted model for which the Bellman operator was a contraction mapping because $\lambda < 1$, without further assumptions, the operator L is not a contraction mapping in expected total reward models. Therefore contraction mapping properties cannot be used to establish existence of solutions to the Bellman equation in all expected total reward models.

Proposition 6.4. Let $\mathbf{u}$ and $\mathbf{v}$ be elements of V. Then

1. For any scalar c, $L(\mathbf{v} + c\mathbf{e}) = L\mathbf{v} + c\mathbf{e}$.
2. If $\mathbf{u} \geq \mathbf{v}$, $L\mathbf{u} \geq L\mathbf{v}$.

Proof. To prove the first part,

$$L(\mathbf{v} + c\mathbf{e}) = \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\{\mathbf{r}_d + \mathbf{P}_d(\mathbf{v} + c\mathbf{e})\} = \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\{\mathbf{r}_d + \mathbf{P}_d\mathbf{v} + c\mathbf{e}\} = L\mathbf{v} + c\mathbf{e}.$$

The second part follows by noting that there exists a $d \in D^{\mathrm{MD}}$ for which

$$L\mathbf{v} = \mathbf{r}_d + \mathbf{P}_d\mathbf{v} \leq \mathbf{r}_d + \mathbf{P}_d\mathbf{u} \leq L\mathbf{u}.$$

□

The following result establishes that $v^*(s)$ is a solution of the Bellman equation. Its subtle proof also applies when $\lambda < 1$, however in that case a more general and obvious approach was used.

Theorem 6.3. Suppose $v^*(s)$ is finite for all $s \in S$, then $\mathbf{v}^*$ is a solution of the Bellman equation.

Proof. First show that $\mathbf{v}^* \geq L\mathbf{v}^*$. From the definition of supremum, given $\epsilon > 0$ for each $s \in S$, there exists a $\pi_s^* \in \Pi^{\mathrm{HR}}$ for which

$$v^{\pi_s^*}(s) \geq v^*(s) - \epsilon. \tag{6.31}$$

Since for all $a \in A_s$

$$v^*(s) \geq r(s,a) + \sum_{j \in S} p(j|s,a) v^{\pi_j^*}(j)$$

it follows from the monotonicity of L (part 2 of Proposition 6.4) that

$$v^*(s) \geq \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^{\pi_j^*}(j) \right\} \geq \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^*(j) \right\} - \epsilon.$$

Since the result holds for every $\epsilon > 0$, and s was arbitrary, $\mathbf{v}^* \geq L\mathbf{v}^*$.

Now show that $\mathbf{v}^* \leq L\mathbf{v}^*$. Write π_s^* in (6.31), as $\pi_s^* = (a_s, \pi_s')$. That is, in state s the policy chooses action a_s at the first decision epoch and then uses the history-dependent randomized policy π_s' thereafter. It follows that

$$v^*(s) - \epsilon \leq v^{\pi_s^*}(s) = r(s, a_s) + \sum_{j \in S} p(j|s, a_s) v^{\pi_j'}(j) \leq \max_{a \in A_S} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^*(j) \right\}.$$

Since ϵ was arbitrary, it follows that $\mathbf{v}^* \leq L\mathbf{v}^*$.

Thus, $\mathbf{v}^* = L\mathbf{v}^*$ and $\mathbf{v}^*$ is a solution of the Bellman equation. □

While $\mathbf{v}^*$ is a solution of the Bellman equation, as a consequence of part 1 of Proposition 6.4, the solution is not unique. Note that if $\mathbf{v}^*$ is infinite, it could also be determined by solving the Bellman equation[10].

The following is an immediate corollary of Theorem 6.3.

Corollary 6.1. For every $d \in D^{\mathrm{MR}}$, $\mathbf{v}^{d^\infty} = L_d \mathbf{v}^{d^\infty}$.

6.3.4 Existence of optimal stationary policies in expected total reward models*

This section shows that under mild assumptions, in a finite state and action Markov decision process, there exists a stationary policy that maximizes the expected total reward. Its proof is rather technical and uses the result (Theorem 5.9) that in a discounted model, for each non-negative $\lambda < 1$ there exists an optimal stationary policy.

Analysis relies on the following technical result, the proof of which is a direct application of Abel's Theorem[11], which is valid even if $v^\pi(s)$ is infinite as would be the case in a one-state one-action model with $r(s) = 1$ and $p(s|s) = 1$.

Lemma 6.1. Suppose for some $\pi \in \Pi^{\mathrm{HR}}$ and $s \in S$

$$\lim_{N \to \infty} E_s^\pi \left[\sum_{n=0}^{N} r(X_n, Y_n) \,\middle|\, X_1 = s \right] = v^\pi(s) \tag{6.32}$$

exists. Then

$$\lim_{\lambda \uparrow 1} v_\lambda^\pi(s) = v^\pi(s). \tag{6.33}$$

[10] Consider a one-state Markov chain resulting in the Bellman equation $v(1) = 1 + v(1)$. This equation has no finite solution, but $v(1) = \infty$ can be regarded to be its solution.

[11] Abel's Theorem (Theorem 8.2, Rudin [1964]) states that if $\sum_{n=0}^{\infty} x_n$ converges, and $f(\lambda) = \sum_{n=0}^{\infty} \lambda^n x_n$ converges for $|\lambda| < 1$, then $\lim_{\lambda \uparrow 1} f(\lambda) = \sum_{n=0}^{\infty} x_n$.

The following example shows that the assumption the limit in (6.32) exists cannot be easily relaxed.

Example 6.6. Consider the stationary policy d_4^∞ in Example 6.1. Letting $\lambda < 1$, gives

$$v_\lambda^{d_4^\infty}(s_1) = 1 - \lambda + \lambda^2 - \lambda^3 + ... = \frac{1}{1+\lambda}.$$

Hence $\lim_{\lambda\uparrow 1} v_\lambda^{d_4^\infty}(s_1) = 1/2$, but $v_\lambda^{d_4^\infty}(s_1)$ does not exist as shown previously in Example 6.1. Moreover

$$\liminf_{N\to\infty} \sum_{n=1}^{N} r(X_n, d_4(X_n)) = 0 < \lim_{\lambda\uparrow 1} v_\lambda^{d_4^\infty}(s_1) < \limsup_{N\to\infty} \sum_{n=1}^{N} r(X_n, d_4(X_n)) = 1,$$

so that the limit of $v_\lambda^{d_4^\infty}(s_1)$ provides no insight about the lim inf or lim sup of the partial sums of (expected) rewards.

Theorem 6.4. Suppose Assumption 6.1 holds. Then there exists an optimal stationary deterministic policy.

Proof. Let $\lambda_n, n = 1, 2, \ldots$ denote a sequence that converges monotonically to 1 from below. From Theorem 5.9 there exists an optimal deterministic stationary policy for all non-negative $\lambda < 1$ in a discounted model. Thus, for each λ_n, there exists an optimal deterministic stationary policy. Since there are only finitely many deterministic stationary policies, there must exist a deterministic stationary policy $(d^*)^\infty$ and a subsequence $\lambda_{n_k}, k = 1, 2, \ldots$ for which $(d^*)^\infty$ is optimal.

Therefore for all $\pi \in \Pi^{\text{HR}}$, $s \in S$, and $k = 1, 2, \ldots$,

$$v_{\lambda_{n_k}}^{\pi}(s) \le v_{\lambda_{n_k}}^{(d^*)^\infty}(s).$$

Hence from Assumption 6.1 and Lemma 6.1, both limits below exist so that

$$v^{\pi}(s) = \lim_{k\to\infty} v_{\lambda_{n_k}}^{\pi}(s) \le \lim_{k\to\infty} v_{\lambda_{n_k}}^{(d^*)^\infty}(s) = v^{(d^*)^\infty}(s). \tag{6.34}$$

Thus $(d^*)^\infty$ is an optimal policy. □

Since Assumption 6.1 holds for all four model classes defined above, the following important result follows.

Corollary 6.2. There exists an optimal deterministic stationary policy in transient, SSP, positive, and negative models.

6.3.5 Identification of optimal policies

This section describes how to identify an optimal policy. One would suspect that using the same approach as in a discounted model, namely that a stationary policy derived from any $\mathbf{v}^*$-greedy policy is optimal. The following simple example shows that this is not the case.

Example 6.7. Let $S = \{s_1, s_2\}, A_{s_1} = \{a_{1,1}, a_{1,2}\}, A_{s_2} = \{a_{2,1}\}, r(s_1, a_{1,1}) = 0, r(s_1, a_{1,2}) = 1, r(s_2, a_{2,1}) = 0$ and $p(s_1|s_1, a_{1,1}) = p(s_2|s_1, a_{1,2}) = p(s_2|s_2, a_{2,1}) = 1$.

This is an example of a positive model. If it is modified by adding a zero-reward absorbing state Δ and changing transition probabilities for actions $a_{1,1}$ and $a_{2,1}$ to $p(\Delta|s_1, a_{1,1}) = p(\Delta|s_2, a_{2,1}) = 1$, it is a transient model.

The Bellman equation for this model is given by:

$$v(s_1) = \max\{v(s_1), 1 + v(s_2)\} \quad \text{and} \quad v(s_2) = v(s_2).$$

Since $v^*(s_1) = 1$ and $v^*(s_2) = 0$, it follows that both $d_1 = (a_{1,1}, a_{2,1})$ and $d_2 = (a_{1,2}, a_{2,1})$ are in $\arg\,\text{c-max}_{d \in D^{\text{MD}}}\{\mathbf{r}_d + \mathbf{P}_d\mathbf{v}^*\}$ but only d_2^∞ is optimal.

The following general result addresses this complication by providing a condition under which stationary policies derived from $\mathbf{v}^*$-greedy policies are optimal.

Theorem 6.5. Suppose Assumption 6.1 holds and d^* satisfies

$$d^* \in \arg\,\text{c-max}_{d \in D^{\text{MD}}}\{\mathbf{r}_d + \mathbf{P}_d\mathbf{v}^*\}. \tag{6.35}$$

Then $\mathbf{v}^{(d^*)^\infty} = \mathbf{v}^*$ if

$$\limsup_{N\to\infty} \mathbf{P}_{d^*}^N \mathbf{v}^* \leq 0. \tag{6.36}$$

Proof. Iteratively applying Theorem 6.3 and (6.35), yields

$$\mathbf{v}^* = \mathbf{r}_{d^*} + \mathbf{P}_{d^*}\mathbf{v}^* = \mathbf{r}_{d^*} + \mathbf{P}_{d^*}\mathbf{r}_{d^*} + \mathbf{P}_{d^*}^2\mathbf{v}^* = \ldots = \sum_{n=1}^{N} \mathbf{P}_{d^*}^{n-1}\mathbf{r}_{d^*} + \mathbf{P}_{d^*}^N\mathbf{v}^*.$$

Thus

$$\mathbf{v}^* \leq \limsup_{N\to\infty} \left(\sum_{n=1}^{N} \mathbf{P}_{d^*}^{n-1}\mathbf{r}_{d^*} + \mathbf{P}_{d^*}^N\mathbf{v}^* \right) \leq \limsup_{N\to\infty} \sum_{n=1}^{N} \mathbf{P}_{d^*}^{n-1}\mathbf{r}_{d^*} + \limsup_{N\to\infty} \mathbf{P}_{d^*}^N\mathbf{v}^* \tag{6.37}$$

As a consequence of Assumption 6.1, $\lim_{N\to\infty}\sum_{n=1}^{N}\mathbf{P}_{d^*}^{n-1}\mathbf{r}_{d^*}$ exists[12] so (6.36) gives

$$\mathbf{v}^* \leq \sum_{n=1}^{\infty}\mathbf{P}_{d^*}^{n-1}\mathbf{r}_{d^*} = \mathbf{v}^{(d^*)^\infty} \leq \mathbf{v}^*,$$

From which the result follows. □

Observe that the form of the rewards and/or the structure of transition probabilities corresponding to d^* determines whether (6.36) holds. The following corollary summarizes these results.

Corollary 6.3. Suppose d^* satisfies (6.35). Then $(d^*)^\infty$ is optimal in:

1. a transient model,
2. an SSP model in which d^* is proper,
3. a positive model in which (6.36) holds,
4. a negative model[a].

[a]Equation (6.36) is automatically satisfied in a negative model since $\mathbf{v}^* \leq \mathbf{0}$.

Returning to Example 6.7, since the model can be classified as a positive model, it requires the addition of condition (6.36) to establish the optimality of d_2^∞ since $\mathbf{P}_{d_1}^N\mathbf{v}^*(s_1) = 1$ for all N. On the other hand, under the modification used to obtain a transient model described in the example, only d_2^∞ attains the maximum in the Bellman equation.

The following example explores the implications of Theorem 6.5 in the context of Example 6.1.

Example 6.8. Recall that the model in Example 6.1 fell into none of the model classes described above. Its Bellman equation may be written as:

$$v(s_1) = \max\{-1 + v(s_1), 1 + v(s_2)\} \quad \text{and} \quad v(s_2) = \max\{v(s_2), -1 + v(s_1)\}.$$

Clearly $v^*(s_1) = 1$ and $v^*(s_2) = 0$ (taking into consideration that the limit defining $v^{d_4^\infty}(s)$ does not exist). Observe that in state s_2 both actions achieve the maximum on the right hand side of the Bellman equation, but only d_3, which satisfies (6.36) is optimal. Thus, this condition can be used to identify optimal policies even in the case when the model does not fall into any of the above

[12]Recall that if the limit of a sequence exists, its lim inf and lim sup are equal to each other and to the limit.

classes or even when Assumption 6.1 doesn't hold.

6.4 State-action value functions

State-action value functions will not play a significant role in this chapter but will be fundamental for analysing episodic models using simulation in Chapters 10 and 11.

For an expected total reward model, define the state-action value function corresponding to π by

$$q^{\pi}(s,a) := r(s,a) + \sum_{j \in S} p(j|s,a) v^{\pi}(j) \tag{6.38}$$

for any $\pi \in \Pi^{\mathrm{HR}}$ and define the optimal state-action value function by

$$q^{*}(s,a) := r(s,a) + \sum_{j|in S} p(j|s,a) v^{*}(j). \tag{6.39}$$

For a deterministic stationary policy d^{∞}, $q^{d^{\infty}}(s,a)$ is a solution of

$$q(s,a) = r(s,d(s)) + \sum_{j \in S} p(j|s,d(s)) q(j,d(j)) \tag{6.40}$$

and for a randomized stationary policy d^{∞} it is a solution of

$$q(s,a) = \sum_{a \in A_s} w_d(a|s) \left(r(s,a) + \sum_{j \in S} p(j|s,a) q(j,a) \right). \tag{6.41}$$

The optimal state-action value function satisfies the Bellman equation:

$$q(s,a) = r(s,a) + \sum_{j \in S} p(j|s,a) \left(\max_{a' \in A_s} q(j,a') \right). \tag{6.42}$$

6.5 Value iteration

This chapter now turns to computation in expected total reward models. It begins with an analysis of value iteration, which can be expressed in vector notation as follows.

Algorithm 6.1. Value iteration for an expected total reward model

1. **Initialize:** Specify $\mathbf{v}'$, $\epsilon > 0$ and $\sigma > \epsilon$.
2. **Iterate:** While $\sigma \geq \epsilon$:
   (a) $\mathbf{v} \leftarrow \mathbf{v}'$.

(b)

$$\mathbf{v}' \leftarrow \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\{\mathbf{r}_d + \mathbf{P}_d \mathbf{v}\} = L\mathbf{v}. \tag{6.43}$$

(c) $\sigma \leftarrow \|\mathbf{v}' - \mathbf{v}\|$.

3. **Terminate:** Return $\mathbf{v}^\epsilon = \mathbf{v}'$ and

$$d_\epsilon \in \underset{d \in D^{\mathrm{MD}}}{\arg\,\text{c-max}}\{\mathbf{r}_d + \mathbf{P}_d \mathbf{v}\}. \tag{6.44}$$

The above algorithm includes a norm-based stopping criterion. This is because we observed that when solving many examples numerically, the norm-based stopping criterion gave the same results as a span-based criterion (see Section 5.5.1)[13]. Note also that when applying value iteration to transient models, recursions use the transient Bellman operator L_T.

This section presents the analysis of the convergence of value iteration in transient and SSP models, and positive and negative models separately. Convergence in transient and SSP models relies on contraction properties of the transient Bellman operator while convergence in positive and negative models uses monotonicity. Note that in the latter case error bounds are not available.

6.5.1 Examples

Before analyzing convergence properties of value iteration, consider the following examples. The first two illustrate some pathological behavior in small deterministic examples. The third one illustrates more typical numerical results and the recursions underlying value iteration.

Example 6.9. (Example 6.7 revisited.) The value iteration recursion for this positive model may be written in component form as

$$v'(s_1) = \max\{v(s_1), 1 + v(s_2)\} \quad \text{and} \quad v'(s_2) = v(s_2).$$

Starting with $v(s_1) = v(s_2) = 0$, Algorithm 6.1 terminates in two iterations with the optimal value function $v'(s_1) = 1$ and $v'(s_2) = 0$ and $d_\epsilon = (a_{1,2}, a_{2,1})$. However, suppose the algorithm started from $v(s_1) = v(s_2) = 1$ then it would terminate with $v'(s_1) = 2$ and $v'(s_2) = 1$ and $d_\epsilon = (a_{1,2}, a_{2,1})$.

Thus, in this example the choice of the initial value affects the solution of the Bellman equation found by the algorithm but not the policy. This is a

[13] We observed that the iterates remain unchanged on states that lead to absorbing states. For example, if in state $s \in S$ there is a single action leading to zero-reward absorbing state Δ with reward c, then value iteration will eventually set $v(s) = c$ so that the quantity involved in defining the span, $\min_{s \in S}\{v'(s) - v(s)\} = 0$ resulting in $\mathrm{sp}(\mathbf{v}' - \mathbf{v}) = \|\mathbf{v}' - \mathbf{v}\|$.

consequence of the non-uniqueness of the solution of the Bellman equation in expected total reward models.

Note that the optimal value function of the discounted version of the problem, $v^*(s_1) = 1, v^*(s_2) = 0$, is in agreement with the first solution above.

Example 6.10. (Example 6.1 revisited.) Recall that the model in Example 6.1 did not fit into any of the specified model classes. The value iteration recursion (6.43) becomes

$$v'(s_1) = \max\{-1 + v(s_1), 1 + v(s_2)\} \quad \text{and} \quad v'(s_2) = \max\{v(s_2), -1 + v(s_1)\}.$$

Starting value iteration at $v(s_1) = v(s_2) = 0$, generates the value $v'(s_1) = 1$ and $v'(s_2) = 0$ at the next and all subsequent iterations. Moreover, (6.44), selects $d_\epsilon(s_1) = a_{1,2}$ and $d_\epsilon(s_2) \in \{a_{2,1}, a_{2,2}\}$.

Starting value iteration at $v(s_1) = v(s_2) = 1$ generates the value $v'(s_1) = 2$ and $v'(s_2) = 1$ at the next and all subsequent iterations, as well as the set of decision rules.

Note that in a discounted version of the problem, value iteration identifies the same optimal value function as when started at $v(s_1) = v(s_2) = 0$ but only the decision rule $d_\epsilon(s_1) = a_{1,2}$ and $d_\epsilon(s_2) = a_{2,1}$.

The next example applies value iteration to the Gridworld model.

Example 6.11. (Example 6.2 revisited.) This example will focus primarily on computational issues; the structure of policies will be explored in Section 6.10.1. Recall that B denotes the reward received when the robot successfully delivers the coffee, X denotes the penalty received if the robot falls down the stairs and c represents the cost of a single transition.

Consider three instances:

Instance 1: $B = 50$, $X = 200$ and $c = 1$,

Instance 2: $B = 0$, $X = 150$ and $c = 1$, and

Instance 3: $B = 1, X = 0$ and $c = 0$.

As noted previously, Instance 1 can be transformed into a negative model, Instance 2 is a negative model and Instance 3 is a positive model. On the other hand, when the probability the robot moves in the intended direction, p is strictly less than 1, the model is transient. When this probability equals 1, it is an SSP model.

The following relationships describe the value iteration recursion for selected states. Let $q = 1 - p$ and assume the only actions in each state are "left", "up" and "right". Note also that the recursions are easier to express when using $r(s, a, j)$ instead of $r(s, a)$ because reward and penalties are received after the transition occurs. It is left to the reader as an exercise to provide recursions for the remaining states.

$$
\begin{aligned}
v'(1) &= v(1) \\
v'(2) &= -c + \max\left\{ p\big(B + v(1)\big) + \frac{1}{2}q\big(v(3) + v(5)\big), pv(3) + \frac{1}{2}q\big(B + v(1) + v(5)\big) \right\} \\
v'(3) &= -c + pv(2) + qv(6), \\
v'(8) &= -c + \max\Big\{ p\big(-X + v(7)\big) + \frac{1}{3}q(v(5) + v(9) + v(11)), \\
&\qquad pv(5) + \frac{1}{3}q\big(-X + v(7) + v(9) + v(11)\big), \\
&\qquad pv(9) + \frac{1}{3}q\big(-X + v(5) + v(7) + v(11)\big) \Big\}.
\end{aligned}
$$

Observe that these recursions account for the consequences of choosing one of the available actions in a state. For example, in cell 2 there are two possible actions "left" and " right". Choosing either results in a transition in the intended direction with probability p, and transitions in the other direction and "down" with probability $q/2$. Other expressions are derived similarly.

Value iteration starts from $\mathbf{v} = \mathbf{0}$ and uses a stopping criterion $\|\mathbf{v}' - \mathbf{v}\| < 0.0001$. In all instances value iteration converges; the number of iterations to achieve the stopping criterion varies with p and instance as shown in Table 6.1. It shows that:

1. For each p the number iterations varied by instance with Instance 1 and 2 the most similar. Moreover, convergence was faster in Instance 3 for all values of p.

2. In each instance the number of iterations to achieve convergence was decreasing with p.

Inspection of calculations revealed that in Instance 1 the iterates of $v(13)$ were non-monotone[a] when $p \geq 0.6$, in Instance 2 the iterates of $v(13)$ were monotone non-increasing, and in Instance 3 the iterates of $v(13)$ were monotone non-decreasing for all values of p.

[a] With respect to iteration number.

p	Instance 1	Instance 2	Instance 3
0	480	481	305
0.25	199	201	101
0.50	114	124	81
0.75	51	44	38
0.95	21	19	17

Table 6.1: Number of iterations to achieve stopping criterion $\|\mathbf{v}' - \mathbf{v}\| < 0.0001$ in Gridworld model categorized by p and instance.

6.5.2 Transient models

Value iteration calculations in the Gridworld model in Example 6.11 showed that as p – the probability the robot moves in the intended direction – increased, the number of iterates required to achieve a specified level of precision decreased (Table 6.1). This makes sense since the larger the value of p, the more quickly the system under an optimal policy will reach an absorbing state, suggesting an underlying contraction behavior. The following analysis of the convergence of value iteration in transient models provides insight into why this is the case.

Recall that in a transient model all deterministic stationary policies have the same reward-free absorbing state Δ. Let $\mathbf{v}_T$ and $\mathbf{v}'_T$ denote the restriction of $\mathbf{v}$ and $\mathbf{v}'$ to the set of transient states, $S \setminus \Delta$. Then the value iteration recursion in a transient model may be written as:

$$\mathbf{v}'_T \leftarrow L_T \mathbf{v}_T = \underset{d \in D^{\mathrm{MD}}}{\text{c-max}} \{(\mathbf{r}_d)_T + \mathbf{Q}_d \mathbf{v}_T\}, \tag{6.45}$$

so that its convergence depends on properties of $\mathbf{Q}_d$.

Solutions of the Bellman equation in a transient model

Theorem 6.3 showed that $\mathbf{v}^*$ was a solution of the Bellman equation in an expected total reward model. The following stronger result follows for transient models.

Theorem 6.6. In a transient model, the transient Bellman equation

$$L_T \mathbf{v} = \mathbf{v}$$

has a unique solution $\mathbf{v}^*_T \in V_T$.

Proof. The result is trivially true if $\mathbf{Q}_d = \mathbf{0}$ for all $d \in D^{\mathrm{MD}}$.

Assume now that $\mathbf{Q}_d$ has at least one positive row sum for each $d \in D^{\mathrm{MD}}$. As a consequence of Part 1 of Proposition 6.4 applied to a transient model, $\mathbf{v}^* + c\mathbf{e}$ is a solution of the transient Bellman equation for any scalar c. Hence it follows for its restriction to transient states, $\mathbf{v}_T^*$, that

$$\mathbf{v}_T^* + c\mathbf{e} = L_T(\mathbf{v}_T^* + c\mathbf{e}) \leq L_T\mathbf{v}_T^* + \underset{d\in D^{\mathrm{MD}}}{\text{c-max}}\, c\mathbf{Q}_d\mathbf{e} = \mathbf{v}_T^* + c\mathbf{Q}_{d'}\mathbf{e}$$

for some $d' \in D^{\mathrm{MD}}$. Hence $c \leq c\mathbf{Q}_{d'}\mathbf{e}(s)$ for all $s \in S \setminus \Delta$. That $c = 0$ follows from the above assumption regarding the positivity of at least one row sum of $\mathbf{Q}_{d'}$. □

Theorem 6.6 establishes the uniqueness of the solution of the Bellman equation on transient states $S \setminus \Delta$. Since the solution on all of S is not unique (Proposition 6.4), adding the condition that $v(\Delta) = 0$ specifies a unique solution as follows.

Corollary 6.4. In a transient model, the Bellman equation

$$L\mathbf{v} = \mathbf{v}$$

subject to $v(\Delta) = 0$ has a unique solution $\mathbf{v}^* \in V$.

Restricting attention to a single decision rule leads to the following result concerning policy evaluation.

Corollary 6.5. In a transient model:

1. The transient policy evaluation equation

$$\mathbf{v} = (\mathbf{r}_d)_T + \mathbf{Q}_d\mathbf{v} = (L_d)_T\mathbf{v} \tag{6.46}$$

has the unique solution $\mathbf{v}_T^{d^\infty} \in V_T$.

2. The policy evaluation equation

$$\mathbf{v} = \mathbf{r}_d + \mathbf{P}_d\mathbf{v} = (L_d)\mathbf{v} \tag{6.47}$$

subject to $v(\Delta) = 0$ has the unique solution $\mathbf{v}^{d^\infty} = (\mathbf{v}_T^{d^\infty}, 0) \in V$.

Convergence of value iteration in transient models

Proving convergence of value iteration in a transient model is based on establishing one of the following results:

1. L_T^N is a contraction mapping with respect to the sup-norm for some $N > 0$,
2. L_T is a contraction mapping with respect to the Euclidean norm,

3. there is an equivalent Markov decision process in which L_T is a contraction mapping with respect to the sup-norm, or

4. L_T is a contraction mapping with respect to a specific weighted sup-norm,

where the concepts of equivalent Markov decision process and weighted sup-norm are defined below.

Under each of these results, it follows from the Banach fixed-point Theorem (Theorem 5.5) that:

Theorem 6.7. In a transient model, value iteration converges geometrically to the unique fixed point of L_T.

The following sections establish this result in different ways. Recall that a sequence $\mathbf{v}^n$, $n = 0, 1, \ldots$ converges geometrically to $\mathbf{v}^*$ if there exists constants $K > 0$ and $0 < \alpha < 1$ for which:

$$\|\mathbf{v}^n - \mathbf{v}^*\| \leq K\alpha^n. \tag{6.48}$$

This is sometimes written as $\|\mathbf{v}^n - \mathbf{v}^*\| = O(\alpha^n)$.

N-stage contractions

By the defining property of a transient model, for each $d \in D^{\text{MD}}$ there exists a state in $S \setminus \Delta$ in which the probability of reaching Δ in one transition is positive. Hence for all states in $S \setminus \Delta$, there is positive probability of reaching Δ in *at most* $N = |S \setminus \Delta|$ transitions. Consequently, there exists a positive $\alpha < 1$ for which $\|\mathbf{Q}_d^N\| \leq \alpha < 1$ for all $d \in D^{\text{MD}}$ so that L_T satisfies

$$\|L_T^N \mathbf{v}_T - L_T^N \mathbf{v}'_T\| \leq \alpha \|\mathbf{v}_T - \mathbf{v}'_T\|$$

for any $\mathbf{v}_T$ and $\mathbf{v}'_T$ in V_T. Thus L_T^N is a contraction mapping so value iteration converges.

Euclidean norm convergence

This approach[14] is based on some convenient properties of the Euclidean norm[15]. Recall that the Euclidean norm of vector $\mathbf{v} \in V_T$ is defined by

$$\|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^{|S\setminus\Delta|} v(s_i)^2}.$$

Given a non-negative matrix $\mathbf{A} : V_T \to V_T$, its subordinate matrix norm with respect to the Euclidean norm is

$$\|\mathbf{A}\|_2 = \sigma(\mathbf{A}),$$

[14] In most settings Markov decision processes are analyzed with respect to the sup-norm.

[15] See Faddeeva [1959].

where $\sigma(A)$ denotes the spectral radius (largest eigenvalue of $\mathbf{A}$).

Modifying the proof of Theorem 5.1 in the discounted case[16], it follows that for $\mathbf{u}$ and $\mathbf{v}$ in V_T, L_T is a contraction mapping. In other words,

$$\|L_T\mathbf{v} - L_T\mathbf{u}\|_2 \leq \alpha\|\mathbf{v} - \mathbf{u}\|_2, \tag{6.49}$$

where

$$\alpha = \max_{d \in D^{\text{MD}}} \sigma(\mathbf{Q}_d) < 1. \tag{6.50}$$

That this quantity is less than 1 follows from the assumptions of transience and the finiteness of the sets of states and actions.

Thus it follows from Theorem 6.7 that value iteration converges geometrically in the Euclidean norm with

$$\|\mathbf{v}^n - \mathbf{v}^*\|_2 \leq K\alpha^n.$$

Since $\|\mathbf{v}\| \leq \|\mathbf{v}\|_2$, it follows that this inequality holds as well for the sup-norm.

Positively similar Markov decision processes*

This section describes Veinott's elegant analysis of the transient model[17]. The rationale for this analysis is that the condition $\|\mathbf{Q}_d^N\| < 1$ for some $N > 1$ does not imply that $\|\mathbf{Q}_d\| < 1$. This means that L_T^N being a contraction does not imply that the "Bellman operator" L_T is a contraction. Hence, the analysis of contraction mappings in Chapter 5 *does not* extend directly to this model.

Example 6.12. To illustrate the above dilemma, consider a model with $S = \{s_1, s_2, \Delta\}$ in which

$$\mathbf{P}_d = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/4 & 0 & 3/4 \\ 0 & 0 & 1 \end{bmatrix}$$

so that

$$\mathbf{Q}_d = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_d^2 = \begin{bmatrix} 3/8 & 1/4 \\ 1/8 & 1/8 \end{bmatrix}.$$

Then $\|\mathbf{Q}_d^2\| = 5/8$ but $\|\mathbf{Q}_d\| = 1$, so $(L_d)_T^2$ is a contraction while $(L_d)_T$ is not. Note further that $\sigma(\mathbf{Q}_d) = 0.6830$.

What Veinott [1969a] showed is that there exists a related Markov decision process model that shares many key properties with the original Markov decision process. This related model is constructed in a way that ensures that the Bellman operator is a contraction mapping, which is sufficient to establish that value iteration converges

[16] A proof is outlined in Problem 3 at the end of this chapter.

[17] See Veinott [1969a].

geometrically. Since this is a property shared by both models, it follows that value iteration converges geometrically in the original model. Details of this analysis follow.

Definition 6.8. Two Markov decision processes with the same state and action spaces, and rewards and transition probabilities given by $(\mathbf{r}_d, \mathbf{P}_d)$ and $(\tilde{\mathbf{r}}_d, \tilde{\mathbf{P}}_d)$ are said to be *positively similar* if there exists a positive diagonal matrix $\mathbf{B}$ for which:

$$\tilde{\mathbf{r}}_d = \mathbf{B}\mathbf{r}_d \quad \text{and} \quad \tilde{\mathbf{P}}_d = \mathbf{B}\mathbf{P}_d\mathbf{B}^{-1}. \tag{6.51}$$

Equivalently, in component form, Markov decision processes with the same state and action spaces, and rewards and transition probabilities given by $\big(r(s,a), p(j|s,a)\big)$ and $\big(\tilde{r}(s,a), \tilde{p}(j|s,a)\big)$ are positively similar if there exists a vector $b(s) > 0$ on $s \in S$ for which:

$$\tilde{r}(s,a) = b(s)r(s,a) \quad \text{and} \quad \tilde{p}(j|s,a) = \frac{1}{b(s)}p(j|s,a)b(j). \tag{6.52}$$

Let $\mathbf{B}_T$ be the restriction of the $\mathbf{B}$ matrix to the transient states of $\mathbf{P}_d$. Observe that if

$$\mathbf{v} = (\mathbf{r}_d)_T + \mathbf{Q}_d\mathbf{v},$$

then

$$\mathbf{B}_T\mathbf{v} = \mathbf{B}_T(\mathbf{r}_d)_T + \mathbf{B}_T\mathbf{Q}_d\mathbf{v} = \mathbf{B}_T(\mathbf{r}_d)_T + \mathbf{B}_T\mathbf{Q}_d(\mathbf{B}_T)^{-1}\mathbf{B}\mathbf{v} = (\tilde{\mathbf{r}}_d)_T + \tilde{\mathbf{Q}}_d(\mathbf{B}_T\mathbf{v}),$$

where $\tilde{\mathbf{Q}}_d := \mathbf{B}_T\mathbf{Q}_d(\mathbf{B}_T)^{-1}$ is the restriction of $\tilde{\mathbf{P}}_d$ to the transient states. Hence, $\mathbf{B}_T\mathbf{v}$ is a solution of

$$\mathbf{v} = (\tilde{\mathbf{r}}_d)_T + \tilde{\mathbf{Q}}_d\mathbf{v}.$$

What this shows that if $\mathbf{v}$ is a solution of the policy evaluation equation in the model with $(\mathbf{r}_d, \mathbf{P}_d)$, then $\mathbf{B}\mathbf{v}$ is a solution in the positively similar model with $(\tilde{\mathbf{r}}_d, \tilde{\mathbf{P}}_d)$. Because the Bellman operator L_T applies the maximum component-wise, it follows that if $\mathbf{v}^*$ is the optimal value in $(\mathbf{r}_d, \mathbf{P}_d)$, then $\mathbf{B}\mathbf{v}^*$ is the optimal value in the positively similar model $(\tilde{\mathbf{r}}_d, \tilde{\mathbf{P}}_d)$. Moreover for any $\mathbf{v} \in V_T$, the same decision rule obtains the arg max in both models.

It is left as an exercise to show that positively similar Markov decision processes:

1. have the same sets of transient policies,
2. have the same optimal policy,
3. have the same spectral radii, and
4. have value iteration iterates that converge to the unique fixed point of the corresponding Bellman operator when the models are transient.

Given a Markov decision process model, the following argument shows how to construct a positively similar Markov decision process with $\|\tilde{\mathbf{Q}}_d\| < 1$, ensuring that value iteration converges geometrically.

Consider an expected total reward Markov decision process with the objective to maximize the number of transitions to reach Δ from each $s \neq \Delta$. This maximum can be found by solving the transient Bellman equation[18]:

$$\mathbf{v}_T = \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\{\mathbf{e} + \mathbf{Q}_d \mathbf{v}_T\}. \tag{6.53}$$

Note that for any $d \in D^{\mathrm{MD}}$, $\mathbf{v}_T^{d^\infty} \geq \mathbf{e}$ so that $\mathbf{v}_T \geq \mathbf{e}$. When $\mathbf{Q}_d \neq \mathbf{0}$, this inequality is strict[19].

The following more general result will be fundamental to what follows. The key assumption is that $\sigma(\mathbf{Q}_d) < 1$, so it also applies to matrices with row sums exceeding 1.

Proposition 6.5. Suppose for all $d \in D^{\mathrm{MD}}$

1. $\mathbf{Q}_d \geq \mathbf{0}$, and
2. $\sigma(\mathbf{Q}_d) < 1$.

Then if $\mathbf{b} \in V_T$ is the solution of

$$\mathbf{v} = \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\{\mathbf{e} + \mathbf{Q}_d \mathbf{v}\} \tag{6.54}$$

and $\mathbf{B}$ is an $|S \setminus \Delta| \times |S \setminus \Delta|$ diagonal matrix with entries $1/b(s)$,

$$\underset{d \in D^{\mathrm{MD}}}{\text{c-max}} \|\mathbf{B}\mathbf{Q}_d\mathbf{B}^{-1}\| < 1. \tag{6.55}$$

Proof. Let $\mathbf{b}$ denote the unique solution of (6.54), which exists by Theorem 6.6. Then

$$\mathbf{b} = \mathbf{e} + \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\, \mathbf{Q}_d \mathbf{b}. \tag{6.56}$$

Let $d^* \in \arg\text{c-max}_{d \in D^{\mathrm{MD}}} \mathbf{Q}_d\mathbf{b}$. Then from Proposition 6.1, $\mathbf{b} = (\mathbf{I} - \mathbf{Q}_{d^*})^{-1}\mathbf{e} \geq \mathbf{e}$ so that $b(s) > 0$ for all $s \in S \setminus \Delta$. Thus, $\mathbf{B}$ is well defined and $\mathbf{B}\mathbf{e}$ is positive.

Noting that $\mathbf{b} = \mathbf{B}^{-1}\mathbf{e}$ and multiplying both sides of (6.56) on the left by $\mathbf{B}$ gives:

$$\mathbf{e} = \mathbf{B}\mathbf{b} = \mathbf{B}\mathbf{e} + \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\, \mathbf{B}\mathbf{Q}_d\mathbf{b} = \mathbf{B}\mathbf{e} + \underset{d \in D^{\mathrm{MD}}}{\text{c-max}}\, \mathbf{B}\mathbf{Q}_d\mathbf{B}^{-1}\mathbf{e}.$$

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

[19] To illustrate the need for this extra condition consider an example with $|S \setminus \Delta| = 1$ and $\mathbf{Q}_d = \mathbf{0}$. In this case, then $\mathbf{v}^{d^\infty} = 1$.

Since $\mathbf{Be}$ is positive,

$$\mathbf{e} > \underset{d\in D^{\mathrm{MD}}}{\text{c-max}} \|\mathbf{B}\mathbf{Q}_d\mathbf{B}^{-1}\|\mathbf{e},$$

which is the desired result.

□

The following result shows that a positively similar model can be constructed that has its norm arbitrarily close to its spectral radius. Its proof below is rather subtle and appeared in Veinott [1969b]. Note this result is trivially true and often overlooked in discounted models because $\sigma(\lambda\mathbf{P}) = \|\lambda\mathbf{P}\| = |\lambda|$ for any probability matrix $\mathbf{P}$.

Theorem 6.8. Given a transient Markov decision process with transition probability matrices $\mathbf{Q}_d$, $d \in D^{\mathrm{MD}}$ on the transient states of S. Then for any $\epsilon > 0$ there exists a positively similar Markov decision process with corresponding transition probability matrix $\tilde{\mathbf{Q}}_d, d \in D^{\mathrm{MD}}$ for which

$$\max_{d\in D^{\mathrm{MD}}} \sigma(\mathbf{Q}_d) \le \max_{d\in D^{\mathrm{MD}}} \|\tilde{\mathbf{Q}}_d\| \le \max_{d\in D^{\mathrm{MD}}} \sigma(\mathbf{Q}_d) + \epsilon. \tag{6.57}$$

Proof. Since $\sigma(\mathbf{Q}) \le \|\mathbf{Q}\|$ for any square matrix $\mathbf{Q}$ the first inequality in (6.57) follows by noting that the spectral radii of matrices in positively similar Markov decision processes are equal.

To obtain the second inequality choose ϵ' such that $\epsilon' < \epsilon$ and $0 < \sigma(\mathbf{Q}_d) + \epsilon' < 1$ for all $d \in D^{\mathrm{MD}}$. Therefore there exists an $\alpha > 1$ for which

$$1 = \alpha \max_{d\in D^{\mathrm{MD}}} \sigma(\mathbf{Q}_d) + \epsilon' = \max_{d\in D^{\mathrm{MD}}} \sigma(\alpha\mathbf{Q}_d) + \epsilon'. \tag{6.58}$$

Since $\sigma(\alpha\mathbf{Q}_d) < 1$, it follows from Proposition 6.5 that there exists a matrix $\mathbf{B}$ for which

$$\max_{d\in D^{\mathrm{MD}}} \|\alpha\mathbf{B}\mathbf{Q}_d\mathbf{B}^{-1}\| < 1.$$

Hence from (6.58)

$$\alpha \max_{d\in D^{\mathrm{MD}}} \|\tilde{\mathbf{Q}}_d\| = \alpha \max_{d\in D^{\mathrm{MD}}} \|\mathbf{B}\mathbf{Q}_d\mathbf{B}^{-1}\| < 1 = \alpha \max_{d\in D^{\mathrm{MD}}} \sigma(\mathbf{Q}_d) + \epsilon'.$$

Dividing through by α, noting that $\alpha > 1$ and $\epsilon > \epsilon'$, gives the result.

□

Example 6.13. Returning to Example 6.12, choosing $\epsilon = 0.01$ and following the argument in the proof of Theorem 6.8 gives $\alpha = \epsilon/\sigma(\mathbf{Q}_d) = 1.45$. Solving (6.54) with $\mathbf{Q}_d$ replaced by $\alpha\mathbf{Q}_d$ gives $\|\tilde{\mathbf{Q}}_d\| = 0.6833$, which is within 0.0003 of $\sigma(\mathbf{Q}_d)$.

Since as a result of Proposition 6.1, $\sigma(\mathbf{Q}_d) < 1$ for all $d \in D^{\mathrm{MD}}$, it follows from Theorem 6.8 and the finiteness of D^{MD} that

$$\max_{d \in D^{\mathrm{MD}}} \|\tilde{\mathbf{Q}}_d\| := \alpha < 1.$$

Hence the same argument as in the discounted case[20] establishes that

$$\tilde{L}_T \mathbf{v} := \underset{d \in D^{\mathrm{MD}}}{\text{c-max}} \{(\tilde{\mathbf{r}}_d)_T + \tilde{\mathbf{Q}}_d \mathbf{v}\} \tag{6.59}$$

is a contraction mapping with modulus α on V_T in the positively similar Markov decision process. Applying the Banach fixed-point theorem (Theorem 5.5) results in:

Proposition 6.6. $\tilde{L}_T$ has a unique fixed point $\tilde{\mathbf{v}}^* \in V_T$ and for any $\mathbf{v} \in V_T$, $\tilde{L}_T^n \mathbf{v}$ converges geometrically to $\tilde{\mathbf{v}}^*$.

Given the relationship between the iterates of value iteration in positively similar Markov decision processes gives:

Theorem 6.9. L_T has a unique fixed point $\mathbf{v}^* \in V_T$ and for any $\mathbf{v} \in V_T$, $L_T^n \mathbf{v}$ converges geometrically to $\mathbf{v}^*$.

In the case of policy evaluation, this result reduces to:

Corollary 6.6. For any $d \in D^{\mathrm{MD}}$, $(L_d)_T$ has a unique fixed point $\mathbf{v}^{d^\infty} \in V_T$ and for any $\mathbf{v} \in V_T$, $(L_d)_T^n \mathbf{v}$ converges geometrically to $\mathbf{v}^{d^\infty}$.

A weighted norm approach*

A third approach to analyzing convergence of value iteration in transient models is based on using weighted norms. by the following.

Definition 6.9. For $\mathbf{v} \in V$, define the *weighted sup-norm* of $\mathbf{v}$ with respect to $\mathbf{b}$ by[a]

$$\|\mathbf{v}\|_{\mathbf{b}} := \max_{s \in S} \left| \frac{v(s)}{b(s)} \right|, \tag{6.60}$$

where $b(s) > 0$ for all $s \in S$.

[a] As in Chapter 5, define the sup-norm using max, since the max is achieved on a finite state space.

[20] See the proof of Proposition 5.1.

Let $V_{\mathbf{b}}$ denote the set of real valued functions on $|S|$ that are bounded in the weighted sup-norm with respect to $\mathbf{b}$. In other words, for each $\mathbf{v} \in V_{\mathbf{b}}$ there exists a $M > 0$ such that $\|\mathbf{v}\|_{\mathbf{b}} \leq M$. Consequently for all $s \in S$,

$$|v(s)| \leq Mb(s).$$

If S is partially ordered, this corresponds to a bounded growth rate. The corresponding *subordinate* matrix norm is defined by:

Definition 6.10. For $\mathbf{A} \in V \times V$, define the *weighted matrix norm* of $\mathbf{A}$ with respect to $\mathbf{b}$ by

$$\|\mathbf{A}\|_{\mathbf{b}} := \max_{s \in S} \frac{\sum_{j \in S} b(j)|a(s,j)|}{b(s)} \tag{6.61}$$

where $b(s) > 0$ for all $s \in S$.

When $\|\mathbf{A}\|_{\mathbf{b}} \leq M$,

$$\sum_{j \in S} b(j)|a(s,j)| \leq Mb(s)$$

for all $s \in S$.

Hence, boundedness in the weighted sup-norm imposes growth conditions on values and structure on matrices. Such norms play an important role in the analysis of countable state models and approximate dynamic programs.

Surprisingly this approach is very closely related to the analysis using positively similar Markov decision process. When $b(s)$ is determined by the methods in that section, it is easy to see that

$$\|\mathbf{Q}_d\|_{\mathbf{b}} = \|\tilde{\mathbf{Q}}_d\|, \tag{6.62}$$

where the norm on the left is the weighted sup-norm on the original Markov decision process while the norm on the right is the *unweighted* sup-norm applied to the constructed positively similar Markov decision process[21]. Therefore it follows that:

Proposition 6.7. L_T is a contraction mapping with respect to the weighted sup-norm $\|\cdot\|_{\mathbf{b}}$, where $\mathbf{b}$ is the unique solution to (6.54).

As a result of Proposition 6.7 value iteration converges geometrically in the weighted sup-norm in the original model. In other words,

$$\|L_T^n \mathbf{v}^0 - \mathbf{v}^*\|_{\mathbf{b}} \leq \alpha^n \|\mathbf{v}^0 - \mathbf{v}^*\|_{\mathbf{b}} \tag{6.63}$$

for some $0 < \alpha < 1$.

[21] Recall the definition of the matrix norm from equation (5.8): $\|\mathbf{A}\| := \max_{s \in S} \sum_{j \in S} |a(s,j)|$.

Error bounds

An error bound in a transient model is derived below. Consider the relationship

$$\|\mathbf{v} - \mathbf{v}^*\| \leq \|\mathbf{v} - \mathbf{v}'\| + \|\mathbf{v}' - \mathbf{v}^*\|.$$

Viewing L_T as an N-stage contraction and setting $\mathbf{v}' = L_T^N \mathbf{v}$ where $N = |S \setminus \Delta|$, it follows that

$$\|\mathbf{v}' - \mathbf{v}^*\| = \|L_T^N \mathbf{v} - L_T^N \mathbf{v}^*\| \leq \alpha \|\mathbf{v} - \mathbf{v}^*\|$$

with[22]

$$\alpha = \max_{d \in D^{\mathrm{MD}}} \|\mathbf{Q}_d^N\| < 1.$$

That this quantity is less than 1 follows from the finiteness of the model. Hence, this gives the error bound

$$\|\mathbf{v}^* - \mathbf{v}\| \leq \left(\frac{1}{1-\alpha}\right) \|L_T^N \mathbf{v} - \mathbf{v}\|. \tag{6.64}$$

In Example 6.12, in which there is a single decision rule and $N = 2$, $\alpha = 5/8$.

A second approach is to derive an error bound in terms of the *Euclidean* norm $\|\cdot\|_2$. Following the above argument shows that

$$\|\mathbf{v}^* - \mathbf{v}\|_2 \leq \left(\frac{1}{1-\alpha}\right) \|L_T \mathbf{v} - \mathbf{v}\|_2, \tag{6.65}$$

with

$$\alpha = \max_{d \in D^{\mathrm{MD}}} \sigma(\mathbf{Q}_d) < 1.$$

Since $\sigma(\mathbf{Q}_d) \leq \|\mathbf{Q}_d\|$ for any norm, the error bound in the Euclidean norm is *tighter* than that using the sup-norm and is consistent with our computations. Moreover since

$$\|\mathbf{v}\| \leq \|\mathbf{v}\|_2 \leq \sqrt{M}\|\mathbf{v}\|,$$

where $M = |S \setminus \Delta|$, it follows from (6.65) that

$$\|\mathbf{v}^* - \mathbf{v}\| \leq \left(\frac{\sqrt{M}}{1-\alpha}\right) \|L_T \mathbf{v} - \mathbf{v}\|. \tag{6.66}$$

In practice this error bound might not be useful since M may be large.

Other approaches for deriving error bounds include constructing a positively similar Markov decision process or using the weighted norm approach.

While these bounds exist in theory, computing them in practice is problematic due to the computation of α, whicch requires a priori knowledge of the optimal policy, i.e., the decision rule that achieves the max in the definition of α.

[22] In component notation,

$$\alpha = \max_{s \in S \setminus \Delta, a \in A_s} \sum_{j \in S \setminus \Delta} q^N(j|s,a)$$

where $q^N(j|s,a)$ denotes components of $p^N(j|s,a)$ restricted to $s \in S \setminus \Delta$ and $j \in S \setminus \Delta$.

6.5.3 Stochastic shortest path models

Recall that in an SSP model, an improper policy cannot be optimal.

Theorem 6.10. In an SSP model:

1. The transient Bellman equation
$$L_T\mathbf{v} = \mathbf{v}$$
has a unique solution $\mathbf{v}_T^* \in V_T$.
2. The Bellman equation
$$L\mathbf{v} = \mathbf{v}$$
subject to $v(\Delta) = 0$ has a unique solution $\mathbf{v}^* \in V$.
3. For any proper stationary policy d^∞ with $d \in D^{\text{MR}}$, the transient policy evaluation equation
$$\mathbf{v} = (\mathbf{r}_d)_T + \mathbf{Q}_d\mathbf{v}$$
has a unique solution $\mathbf{v}_T^{d^\infty} \in V_T$.
4. For any proper stationary policy d^∞ with $d \in D^{\text{MR}}$, the policy evaluation equation
$$\mathbf{v} = \mathbf{r}_d + \mathbf{P}_d\mathbf{v}$$
has a unique solution $\mathbf{v}^{d^\infty} \in V$.

Analyses in the previous section showed that value iteration converges in a transient model or equivalently in an SSP model in which every policy is proper[23].

6.5.4 Positive models

The following sections provide analyses of the convergence of value iteration in positive and negative models. The following result holds for value iteration in any expected total reward model.

Lemma 6.2. Suppose the sequence $\mathbf{v}^n, n = 0, 1, 2, \ldots$ is generated by value

[23] Bertsekas and Tsitsiklis [1991] establish this result in SSP models with improper policies, however convergence may not be at a geometric rate.

iteration with $\mathbf{v}^0 = \mathbf{0}$. Then for each $N \geq 1$,

$$\mathbf{v}^n = \mathbf{v}_N^\pi = \sum_{n=1}^{N} \mathbf{P}_{d_1} \cdots \mathbf{P}_{d_n} \mathbf{r}_{d_n}, \tag{6.67}$$

where $\pi = (d_1, d_2, ...) \in \Pi^{\text{MD}}$. Moreover $(d_1, \ldots, d_N)$ is an optimal policy in an $(N+1)$-period model with terminal reward 0.

The Bellman equation in positive models

The following technical result summarizes useful properties of the Bellman operator L in positive models. It is convenient to denote the set of non-negative values, $V^+ := \{\mathbf{v} \in V \,|\, \mathbf{v} \geq \mathbf{0}\}$.

Proposition 6.8. In a positive model:

1. There exists a $d \in D^{\text{MD}}$ for which $\mathbf{v}^{d^\infty} \geq \mathbf{0}$,
2. $L\mathbf{0} \geq \mathbf{0}$,
3. $\mathbf{v}^* \geq \mathbf{0}$,
4. $L : V^+ \to V^+$, and
5. for any $d \in D^{\text{MD}}$, $r_d(s) \leq 0$ for all s in each recurrent set of $\mathbf{P}_d$.

Proof. Parts 1-3 are immediate consequences of the second condition defining a positive model. Part 4 is an immediate consequence of Proposition 6.4 and Part 5is an immediate consequence of the first condition defining a positive model. □

The model in Example 6.1 with action $a_{2,2}$ deleted is a positive model and exhibits all the results in the above proposition. Note that under action $a_{1,1}$, s_1 is recurrent but its expected total reward is $-\infty$. Note that this is an SSP model since the policy using action $a_{1,1}$ every period is improper.

The following result identifies $\mathbf{v}^*$ among the solutions of the Bellman equation in a positive model.

Theorem 6.11. In a positive model:

1. If $\mathbf{v} \in V^+$ satisfies $\mathbf{v} \geq L\mathbf{v}$, then $\mathbf{v} \geq \mathbf{v}^*$.
2. The optimal value function $\mathbf{v}^*$ is the *minimal* solution in V^+ of
$$L\mathbf{v} = \mathbf{v}.$$
3. For any $d \in D^{\mathrm{MR}}$ the policy value function $\mathbf{v}^{d^\infty} \in V^+$ is the minimal solution of
$$L_d\mathbf{v} = \mathbf{v}.$$

Proof. To prove Part 1 choose $\pi = (d_1, d_2, \ldots) \in \Pi^{\mathrm{MR}}$. By Lemma 5.4 and the assumption that $\mathbf{v} \geq L\mathbf{v}$,

$$\begin{aligned}\mathbf{v} &\geq \mathbf{r}_{d_1} + \mathbf{P}_{d_1}\mathbf{v} \geq \mathbf{r}_{d_1} + \mathbf{P}_{d_1}\mathbf{r}_{d_2} + \mathbf{P}_{d_1}\mathbf{P}_{d_2}\mathbf{v} \geq \cdots \\ &\geq \sum_{n=1}^{N} \mathbf{P}_{d_1}\cdots\mathbf{P}_{d_n}\mathbf{r}_{d_n} + \mathbf{P}_{d_1}\cdots\mathbf{P}_{d_N}\mathbf{v} \geq \sum_{n=1}^{N} \mathbf{P}_{d_1}\cdots\mathbf{P}_{d_n}\mathbf{r}_{d_n},\end{aligned}$$

where the last inequality follows by noting that $\mathbf{v} \geq \mathbf{0}$. Hence, passing to the limit on the right hand side of the above expression establishes that $\mathbf{v} \geq \mathbf{v}^*$.

Part 2 follows immediately from Part 1 applied to $L\mathbf{v} = \mathbf{v}$. Part 3 follows by restricting Part 2 to a single decision rule and Lemma 5.4. □

Example 6.14. The deterministic positive model represented by Figure 6.4 illustrates how to use this result. Decision rules are specified by action choice in state s_1. The Bellman equation for this model can be written as:

$$\begin{aligned}v(s_1) &= \max\{v(s_1), 1 + v(s_2)\} \\ v(s_2) &= v(s_2).\end{aligned}$$

Any solution of the Bellman equation can be expressed as $v(s_1) = b$ and $v(s_2) = c$ where $b \geq c + 1$. Thus the minimal solution is $v^*(s_1) = 1$ and $v(s_2) = 0$, and the corresponding greedy decision rule is $d(s_1) = a_{1,2}$, $d(s_2) = a_{2,1}$.

Observe that s_2 is recurrent (absorbing) under the optimal policy so that the requirement that it be the minimal value in V^+ means that $v^*(s_2)$ must be 0.

If instead if $r(s_1, a_{1,2}) = -1$, surprisingly this model would still be a positive model even though one of the rewards is negative. The minimal solution of the Bellman equation for this modification would be $v(s_1) = v(s_2) = 0$.

Thus in both of these models, the minimal solution sets $v(s) = 0$ when s is in a recurrent state under the optimal policy.

Note that when $r(s_1, a_{1,2}) = -1$ it would also be a negative model (see Figure 6.3). Moreover either model could be transformed to a transient model,

by adding a zero-reward absorbing state Δ and replacing the zero-reward self transitions by ones from states s_1 and s_2 to Δ.

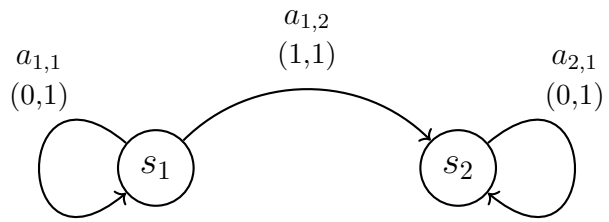

Figure 6.4: Symbolic representation of positive model in Example 6.14. Recall that the labels (r, p) refer to the reward and transition probabilities, respectively, associated with using the specified action in the given state.

Convergence of value iteration in positive models

Analysis of the Gridworld model in Example 6.11 showed that value iteration converged monotonically in Instance 3, which was a positive model. Such monotonicity holds for any positive model and provides the basis for a proof of convergence of value iteration.

Theorem 6.12. Suppose in a positive model the sequence $\mathbf{v}^n, n = 0, 1, 2, \ldots$ is generated by value iteration with $\mathbf{v}^0 = \mathbf{0}$. Then $\mathbf{v}^n$ converges monotonically to $\mathbf{v}^*$ and $\|\mathbf{v}^n - \mathbf{v}^*\| \to 0$.

Proof. Since $\mathbf{v}^1 = L\mathbf{0} \geq \mathbf{0} = \mathbf{v}^0$, it follows from part 2 of Proposition 6.4 that $\mathbf{v}^n$ is monotonically non-decreasing. Since $\mathbf{v}^n$ is the value of a policy, condition 1 in the definition of positive models implies that it is bounded above. Hence, by a standard result in analysis,[24] $v^n(s)$ converges for each $s \in S$. Call the limit $v'(s)$. Since S is finite, $\|\mathbf{v}^n - \mathbf{v}'\| \to 0$.

Since $\mathbf{v}^n$ is monotonically non-decreasing $L\mathbf{v}' \geq \mathbf{v}^n$ for all $n \geq 0$ so that $L\mathbf{v}' \geq \mathbf{v}'$. On the other hand, for any $d \in D^{\mathrm{MD}}$

$$L_d\mathbf{v}^n \leq L\mathbf{v}^n = \mathbf{v}^{n+1} \leq \mathbf{v}'.$$

Since $L_d\mathbf{v}$ is linear in $\mathbf{v}$, $L_d\mathbf{v}^n$ converges to $L_d\mathbf{v}'$ so that $L_d\mathbf{v}' \leq \mathbf{v}'$. Hence

$$L\mathbf{v}' = \operatorname*{c\text{-}max}_{d \in D^{\mathrm{MD}}} L_d\mathbf{v}' \leq \mathbf{v}'.$$

[24] Theorem 3.14 in Rudin [1964].

Therefore, $L\mathbf{v}' = \mathbf{v}'$.

Since for every $n \geq 0, \mathbf{v}^n$ is the value of a policy, it follows from Theorem 6.3 that $\mathbf{v}' = \mathbf{v}^*$. □

6.5.5 Negative models

Instance 2 in the Gridworld model in Example 6.11 is an example of a negative model. For this instance, value iteration generated a convergent monotonically non-increasing sequence of iterates. This section establishes that result in greater generality.

The Bellman equation in negative models

The following technical result summarizes useful properties of the Bellman operator L in negative models. As above, it is convenient to denote the set of non-positive values, $V^- := \{\mathbf{v} \in V \mid \mathbf{v} \leq \mathbf{0}\}$.

Proposition 6.9. In a negative model:

1. For all $d \in D^{\mathrm{MD}}$, $\mathbf{0} \geq \mathbf{v}^{d^\infty}$,
2. $L\mathbf{0} \leq \mathbf{0}$,
3. $\mathbf{v}^* \leq \mathbf{0}$,
4. $L : V^- \to V^-$, and
5. if $\mathbf{v}^{d^\infty}(s) > -\infty$ for some $s \in S$ and $d \in D^{\mathrm{MD}}$, then $r_d(s') = 0$ for all s' in the recurrent set of $\mathbf{P}_d$ containing s.

Proof. Condition 1 in the definition of a negative model implies that $r(s,a) \leq 0$ for $a \in A_s, s \in S$. From this observation, results 1-4 follow immediately. Part 5 follows by noting that if $r_d(s') < 0$ for some s' in a recurrent set of $\mathbf{P}_d$, then $v_d(s) = -\infty$ for all s in that recurrent set of $\mathbf{P}_d$, which would be a contradiction. □

The following result is the equivalent of Theorem 6.11 for negative models. The first part is also used to prove that value iteration converges below.

Theorem 6.13. In a negative model:

1. If $\mathbf{v} \in V^-$ satisfies $\mathbf{v} \leq L\mathbf{v}$, then $\mathbf{v} \leq \mathbf{v}^*$.
2. The optimal value function $\mathbf{v}^*$ is the *maximal* solution in V^- of
$$L\mathbf{v} = \mathbf{v}.$$
3. For any $d \in D^{\mathrm{MR}}$ the policy value function $\mathbf{v}^{d^\infty} \in V^-$ is the maximal solution of
$$L_d\mathbf{v} = \mathbf{v}.$$

Proof. A proof of Part 1 follows. Since $L\mathbf{v} \geq \mathbf{v}$, there exists a $d_1 \in D^{\mathrm{MD}}$ for which:

$$\mathbf{r}_{d_1} + \mathbf{P}_{d_1}\mathbf{v} \geq \mathbf{v}.$$

But since $\mathbf{v} \in V^-$, $\mathbf{P}_{d_1}\mathbf{v} \leq \mathbf{0}$ so that $\mathbf{r}_{d_1} \geq \mathbf{v}$. Applying L to both sides of this inequality implies that there exists a $d_2 \in D^{\mathrm{MD}}$ for which

$$\mathbf{r}_{d_2} + \mathbf{P}_{d_2}\mathbf{r}_{d_1} = L\mathbf{r}_{d_1} \geq L\mathbf{v} \geq \mathbf{v}.$$

By repeated application of this argument it follows that there exists a policy $\pi = (d_1, d_2, \ldots) \in \Pi^{\mathrm{MD}}$ for which $\mathbf{v}^\pi \geq \mathbf{v}$. Hence $\mathbf{v}^* \geq \mathbf{v}$. Part 2 follows immediately from Part 1 applied to $L\mathbf{v} = \mathbf{v}$. Part 3 follows by restricting Part 2 to a single decision rule and Lemma 5.4. □

Example 6.14 showed that when $r(s_1, a_{1,2}) = -1$, the model was also negative (see Figure 6.3). For this model the Bellman equation is:

$$\begin{aligned} v(s_1) &= \max\{v(s_1), -1 + v(s_2)\} \\ v(s_2) &= v(s_2). \end{aligned}$$

Therefore, any solution has the form $v(s_1) = b$, $v(s_2) = c$ where $b \geq c - 1$. Hence the maximal solution is $v^*(s_1) = v^*(s_2) = 0$, which is zero on the recurrent states of the optimal policy.

Convergence of value iteration in negative models

The following result establishes convergence of value iteration. Its proof is a bit more complicated than that for positive models.

Theorem 6.14. Suppose in a negative model the sequence $\mathbf{v}^n, n = 0, 1, 2, \ldots$ is generated by value iteration with $\mathbf{v}^0 = \mathbf{0}$. Then $\mathbf{v}^n$ converges monotonically to $\mathbf{v}^*$ and $\|\mathbf{v}^n - \mathbf{v}^*\| \to 0$.

Proof. Since $\mathbf{v}^1 = L\mathbf{0} \leq \mathbf{0} = \mathbf{v}^0$, it follows from part 2 of Proposition 6.4 that $\mathbf{v}^n$ is monotonically non-increasing. Since $\mathbf{v}^* \leq \mathbf{0}$ and satisfies the optimality equation, it follows from the monotonicity of L that for any $n \geq 1$

$$\mathbf{v}^* = L^n\mathbf{v}^* \leq L^n\mathbf{0} = \mathbf{v}^{n+1}.$$

Hence by the same argument as in the positive case, $\mathbf{v}^n$ converges to a limit $\mathbf{v}'$ which satisfies $L\mathbf{v}' \leq \mathbf{v}'$ and $\mathbf{v}' \geq \mathbf{v}^*$. Since S is finite, $\|\mathbf{v}^n - \mathbf{v}'\| \to 0$.

The following arguments establish the reverse inequality. Let $\epsilon > 0$.

$$\begin{aligned}
\mathbf{v}' = \lim_{n\to\infty} \mathbf{v}^n = \lim_{n\to\infty} L\mathbf{v}^n &= \lim_{n\to\infty} \text{c-max}_{d\in D^{\text{MD}}}\{\mathbf{r}_d + \mathbf{P}_d\mathbf{v}^n\} \\
&= \lim_{n\to\infty} \text{c-max}_{d\in D^{\text{MD}}}\{\mathbf{r}_d + \mathbf{P}_d\mathbf{v}' + \mathbf{P}_d(\mathbf{v}^n - \mathbf{v}')\} \\
&\leq \lim_{n\to\infty} \text{c-max}_{d\in D^{\text{MD}}}\{\mathbf{r}_d + \mathbf{P}_d\mathbf{v}' + \|\mathbf{v}^n - \mathbf{v}'\|\mathbf{e}\} \\
&\leq L\mathbf{v}' + \epsilon\mathbf{e},
\end{aligned}$$

where the last inequality follows by noting that for any $\epsilon > 0$ and n sufficiently large $\|\mathbf{v}^n - \mathbf{v}'\| \leq \epsilon$. Since ϵ was arbitrary, $\mathbf{v}' \leq L\mathbf{v}'$ so that it follows from Part 1 of Theorem 6.13 that $\mathbf{v}' \leq \mathbf{v}^*$.

That $\mathbf{v}' = \mathbf{v}^*$ follows by combining both parts of the proof. □

6.6 Policy iteration

This section concerns the convergence of policy iteration in transient and SSP models. Analysis of positive and negative models is more subtle[25] and not described here. In these models, policy iteration solves the transient Bellman equation by iterating between evaluation and improvement steps.

A statement of the algorithm follows.

Algorithm 6.2. Policy iteration for a transient model: vector form

1. **Initialize:** Specify $d \in D^{\text{MD}}$ and set $\Gamma \leftarrow S \setminus \Delta$.
2. **Iterate:** While $\Gamma \geq \emptyset$:
   (a) **Evaluate:** Find $\mathbf{v}'_T$ by solving
   $$\mathbf{v}_T = (\mathbf{r}_d)_T + \mathbf{Q}_d\mathbf{v}_T. \tag{6.68}$$
   (b) **Improve:** Choose
   $$d' \in \arg\text{c-max}_{\delta\in D^{\text{MD}}}\{(\mathbf{r}_\delta)_T + \mathbf{Q}_\delta\mathbf{v}'_T\}, \tag{6.69}$$

[25] See Sections 7.2.5 and 7.3.4 in Puterman [1994].

setting $d' = d$ if possible.

(c) $\Gamma \leftarrow \{s \in S \setminus \Delta \,|\, d'(s) \neq d(s)\}$.

(d) **Update:** $d \leftarrow d'$.

3. **Terminate:** Return $d^* = d$ and $\mathbf{v}^* = \mathbf{v}'_T$.

Note that the improvement step is executed component-wise as follows. For each $s \in S \setminus \Delta$ it chooses

$$a'_s \in \arg\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v(j) \right\}, \tag{6.70}$$

setting $a'_s = d(s)$ if possible.

The conditions "setting $d' = d$ if possible" or "setting $a'_s = d(s)$ if possible" are referred to as *anti-cycling rules*. They apply to the case when there is more than one decision rule or action that attains the max in the improvement step. The inclusion of this stipulation ensures that the stopping criterion is eventually satisfied.

Moreover, the evaluation step can be implemented recursively by iterating

$$\begin{aligned} \mathbf{v}'_T &\leftarrow (\mathbf{r}_d)_T + \mathbf{Q}_d \mathbf{v}_T \\ \mathbf{v}_T &\leftarrow \mathbf{v}'_T \end{aligned}$$

to achieve a high degree of precision. By the argument in the previous section, this iterative scheme converges to $\mathbf{v}^{d^\infty}$ in a transient model. Truncating this iterative scheme corresponds to modified policy iteration below.

6.6.1 Transient models

This section analyzes convergence of policy iteration in transient models. The development parallels that in the discounted case (Section 5.7) by showing that the iterates of policy iteration have a "Newton" representation. To do so, define the operator $B_T : V_T \to V_T$ by $B_T \mathbf{v}_T := L_T \mathbf{v}_T - \mathbf{v}_T$.

Proposition 6.10. Let $\mathbf{v}_T$ and $\mathbf{v}'_T$ be successive iterates of transient policy iteration (restricted to the set of transient states). Then

$$\mathbf{v}'_T = \mathbf{v}_T + (\mathbf{I} - \mathbf{Q}_{d_\mathbf{v}})^{-1} B_T \mathbf{v}_T \tag{6.71}$$

where $d_\mathbf{v} \in \arg\,\text{c-max}_{\delta \in D^{\text{MD}}} \{(\mathbf{r}_\delta)_T + \mathbf{Q}_\delta \mathbf{v}_T\}$.

Proof. The proof is identical to that in the discounted case (Proposition 5.10) after taking into account that iterates are restricted to transient states. It follows that:

$$\begin{aligned}\mathbf{v}'_T &= (\mathbf{I} - \mathbf{Q}_{d_\mathbf{v}})^{-1}(\mathbf{r}_{d_\mathbf{v}})_T \\ &= (\mathbf{I} - \mathbf{Q}_{d_\mathbf{v}})^{-1}(\mathbf{r}_{d_\mathbf{v}})_T - \mathbf{v}_T + \mathbf{v}_T \\ &= (\mathbf{I} - \mathbf{Q}_{d_\mathbf{v}})^{-1}\left((\mathbf{r}_{d_\mathbf{v}})_T + (\mathbf{Q}_{d_\mathbf{v}} - \mathbf{I})\mathbf{v}_T\right) + \mathbf{v}_T \\ &= \mathbf{v}_T + (\mathbf{I} - \mathbf{Q}_{d_\mathbf{v}})^{-1} B_T \mathbf{v}_T.\end{aligned}$$

□

This result leads to an easy proof of the convergence of policy iteration in transient models.

Theorem 6.15. In a transient model, policy iteration converges monotonically and in a finite number of iterations to the optimal value function $\mathbf{v}_T^*$ and an optimal stationary policy.

Proof. By construction

$$\mathbf{v}_T = (\mathbf{r}_d)_T + \mathbf{Q}_d \mathbf{v}_T \leq \underset{\delta \in D^{\mathrm{MD}}}{\text{c-max}}\{(\mathbf{r}_\delta)_T + \mathbf{Q}_\delta \mathbf{v}_T\} = L_T \mathbf{v}_T$$

with strict inequality whenever $d' \neq d$ where d' satisfies (6.69). Therefore, either $B_T \mathbf{v}_T = L_T \mathbf{v}_T - \mathbf{v}_T = \mathbf{0}$ or $B_T \mathbf{v}_T(s) > 0$ for some $s \in S \setminus \Delta$. In the former case $\mathbf{v}_T$ solves the transient Bellman equation and in the latter case it follows immediately from (6.71) and the observation that $(\mathbf{I} - \mathbf{Q}_d)^{-1} \geq \mathbf{I}$ that $v'_T(s) > v_T(s)$ for some $s \in S - \Delta$. Since there are only finitely many policies, the stopping criterion must be satisfied at some finite N at which point $\mathbf{v}_T$ satisfies the Bellman equation. From Corollary (6.3), d^∞ is an optimal stationary policy. □

Example 6.15. This example solves Instance 1 of the Gridworld model in Example 6.2 using policy iteration. Choose the initial decision rule as "go up" when possible, otherwise "go right" in cell 2 and "go left" in cell 3[a]. To avoid explicitly writing our $\mathbf{Q}_d$ and thus simplify coding, the evaluation step was carried out iteratively.

The number of iterations (improvement steps) required for convergence varied with p as shown in Table 6.2.

p	0.01	0.1	0.5	0.9	0.99
iterations	5	4	3	3	6

Table 6.2: Number of iterations for policy iteration to converge in Gridworld model as a function of p.

Note that when $p = 0$ or $p = 1$ the stationary policy corresponding to the initial decision rule was improper because it cycled between cells 2 and 3 forming a closed class consisting of cells 2 and 3. Since the cost associated with these states is negative infinity the model in this case is an SSP model. Hence, as noted below, in such a case, starting the algorithm with a proper policy ensure convergences.

[a] To simplify coding one could allow all actions in all cells but set $q(s, a)$ to a large negative value when action a is impossible in state s. If this adjustment is made, policy iteration could start with the decision rule "go up" in all states.

6.6.2 Stochastic shortest path models

The above result extends directly to SSP models in which policy iteration starts with a proper policy. The argument used to prove Theorem 6.15 ensures that an improper policy cannot be identified at a subsequent iterate of policy iteration.

Corollary 6.7. In a stochastic shortest path model, suppose transient policy iteration begins with a *proper* policy. Then it converges monotonically to $\mathbf{v}_T^*$ in a finite number of iterations to the optimal value function and an optimal stationary policy.

Numerical examples such as that in Example 6.18 below, show that when starting with an improper policy, (6.68) will not have a solution.

6.7 Modified policy iteration

Since transient models behave like discounted models, a modified policy iteration (MPI) algorithm can be analyzed by previously developed methods. Because $v^*(\Delta) = 0$ and all policies stop when reaching Δ, MPI can be restricted to transient states. A vector version of the algorithm follows.

6.7.1 Transient models

Algorithm 6.3. Modified policy iteration for a transient model: vector form

1. **Initialize:**
   (a) Specify $d' \in D^{\text{MD}}$.
   (b) Specify $\epsilon > 0$ and $\sigma > \epsilon$.

(c) Specify $\mathbf{v}_T \in V_T$.

(d) Specify a sequence of non-negative integers m_n, $n = 1, 2, \ldots$.

(e) $n \leftarrow 1$.

2. **Iterate:** While $\sigma \geq \epsilon$:

(a) **Update:** $d \leftarrow d'$.

(b) **Evaluate:**

i. $m \leftarrow 1$ and $\mathbf{u}_T \leftarrow \mathbf{v}_T$.

ii. While $m \leq m_n$:

A. $\mathbf{u}_T \leftarrow (\mathbf{r}_d)_T + \mathbf{Q}_d \mathbf{u}_T$.

B. $m \leftarrow m + 1$.

(c) **Improve:**

i. Choose

$$d' \in \arg\operatorname*{c\text{-}max}_{\delta \in D^{\mathrm{MD}}} \{(\mathbf{r}_\delta)_T + \mathbf{Q}_\delta \mathbf{u}_T\}. \tag{6.72}$$

ii. $\mathbf{v}_T \leftarrow L_T \mathbf{u}_T$.

iii. $\sigma \leftarrow \|\mathbf{v}_T - \mathbf{u}_T\|$.

iv. $n \leftarrow n + 1$.

3. **Terminate:** Return $d_\epsilon = d'$ and $\mathbf{v}_T^\epsilon = \mathbf{v}_T$.

Some comments about the algorithm follow:

1. The algorithm could start directly from the improvement step, avoiding the need to specify an initial policy.

2. The index n is necessary to determine m_n.

3. Setting $m_n = 0$ (so the "Evaluate" step above is skipped) for all n is identical to value iteration and $m_n = \infty$ corresponds to policy iteration.

4. The stopping criterion is based on the value function, while that of policy iteration is based on the decision rule.

Let $\mathbf{v}_T^n$ and d_n for $n = 1, 2, \ldots$ denote the sequence of values and decision rules, respectively, generated by modified policy iteration[26]. Then with the exception of the first iterate, which can be written as

$$\mathbf{v}_T^2 = L_{d^1}^{m_1} \mathbf{v}_T^1,$$

[26] This notation assumes the initial value is $\mathbf{v}_T^0$ and the initial decision rule is d_0.

iterates can be written as:

$$\mathbf{v}_T^{n+1} = L_{d'}^{m_n+1}\mathbf{v}_T^n. \tag{6.73}$$

Therefore

$$\mathbf{v}_T^{n+1} = L_{d_n}^{m_n+1}\dots L_{d_1}^{m_1}\mathbf{v}_T^1. \tag{6.74}$$

This quantity can be interpreted as the expected total reward obtained using the policy implied by the above sequence of operators with terminal reward $\mathbf{v}_T^1$. If instead the algorithm began with the improvement step, $\mathbf{v}_T^2 = L^{m_1+1}\mathbf{v}_T^1$ consistent with the remaining terms. Consequently, the assumption of a transient model, causes the impact of the terminal reward to decrease as n increases.

6.7.2 Convergence in transient models

An outline of a proof of the convergence of MPI under the assumption that the algorithm starts from a value that satisfies $L\mathbf{v}_T \geq \mathbf{v}_T$ follows. All the details appear in the analysis of the discounted case in Section 5.8.3.

Theorem 6.16. Let $m_n, n = 1, 2, \dots$ denote a sequence of non-negative integers and let $\mathbf{v}_T^n, n = 1, 2, \dots$ denote a sequence of iterates of modified policy iteration. Then if $L\mathbf{v}_T^1 \geq \mathbf{v}_T^1$, $\mathbf{v}_T^n$ converges monotonically to $\mathbf{v}_T^*$ and $\|\mathbf{v}_T^n - \mathbf{v}_T^*\| \to 0$.

Proof. Let $\mathbf{u}_T^n$ denote the iterates of value iteration starting with $\mathbf{u}_T^1 = \mathbf{v}_T^1$ for $n = 1, 2, \dots$ Then by induction it follows that

$$\mathbf{u}_T^n \leq \mathbf{v}_T^n \leq \mathbf{v}_T^*.$$

Since $\mathbf{u}_T^n$ converges monotonically to the optimal value function $\mathbf{v}_T^*$ when $L\mathbf{u}_T^1 \geq \mathbf{u}_T^1$, the above inequalities imply the result. □

This result also holds if MPI is started in the improvement step with $\mathbf{v}_T^1$ satisfying the above condition.

6.7.3 An example

The following example applies modified policy iteration to an instance of the Gridworld model, investigating the choice of m_n.

Example 6.16. This example considers Instance 1 of the Gridworld model in Example 6.11 with $p = 0.1$ and $\epsilon = 0.0001$. The probability p is chosen to be small so that value iteration converges slowly and the impact of the choice of m_n can be examined. The algorithm is initialized with the same decision rule used to start policy iteration in the previous section. Values of m_n are chosen to be

constant, increasing and decreasing with n.

Table 6.3 gives computational results. The quantity *Evaluation equivalents* equals the sum of the number of evaluations plus twice the number of improvements because each maximization required evaluating 28 state-action pairs while each evaluation step required updating 13 values. Hence a maximization step is "roughly" twice as costly as an evaluation step.

m_n	Improvement steps (maximizations)	Evaluation equivalents	Iteration at which optimal policy first appears
0	251	502	31
2	88	352	7
10	21	252	2
20	13	286	2
50	7	364	1
n^2	10	405	3
$\max\{40-n, 0\}$	9	666	4

Table 6.3: Computational results for various specifications for m_n.

The results show that a moderate fixed value of m_n or an increasing sequence has the best performance. Moreover all choices of m_n with the exception of $m_n = \max\{40-n, 0\}$ require less computational effort than value iteration.

6.7.4 Stochastic shortest path models

In an SSP model, modified policy iteration converges when

1. initiated with a proper policy, or
2. it begins in the improvement step, or
3. when $m_0 < \infty$.

In the latter case, $\mathbf{v}'_T$ will be finite so at the subsequent improvement step, a proper policy will be identified.

6.8 Linear programming in transient models

This section discusses the formulation and properties of linear programs in transient expected total reward models.

6.8.1 The primal linear program

If $v(s)$ satisfies the transient Bellman equation

$$v(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S \setminus \Delta} p(j|s,a) v(j) \right\}$$

for all $s \in S \setminus \Delta$, then it satisfies

$$v(s) \geq \; r(s,a) + \sum_{j \in S \setminus \Delta} p(j|s,a) v(j) \tag{6.75}$$

for all $a \in A_s$ and $s \in S \setminus \Delta$[27]. By a generalization of Corollary 5.2 in the case of discounted models, any $v(s)$ satisfying (6.75) for all $a \in A_s$ and $s \in S \setminus \Delta$ is an upper bound on $v^*(s)$. Hence a solution of the transient Bellman equation may be thought of as the *smallest* (in a component-wise sense) $v(s)$ that satisfies (6.75) for all $a \in A_s$ and $s \in S \setminus \Delta$. To find such a value function, choose $\alpha(s) > 0$ for all $s \in S \setminus \Delta$ and solve the following linear program.

Primal LP formulation of a transient model

$$\text{minimize} \quad \sum_{s \in S \setminus \Delta} \alpha(s) v(s) \tag{6.76a}$$

$$\text{subject to} \quad v(s) - \sum_{j \in S \setminus \Delta} p(j|s,a) v(j) \geq r(s,a), \quad a \in A_s, \; s \in S \setminus \Delta \tag{6.76b}$$

Because the dual variables have interesting properties the dual linear program is analyzed next.

6.8.2 The dual linear program

The corresponding dual linear program may be written as:

[27] Recall that in a transient model $v(\Delta) = 0$ so that term in the summation in the Bellman equation can be removed.

Dual LP formulation of a transient model

$$\text{maximize} \quad \sum_{s\in S\setminus\Delta}\sum_{a\in A_s} r(s,a)x(s,a) \tag{6.77a}$$

$$\text{subject to} \quad \sum_{a\in A_s} x(s,a) - \sum_{j\in S\setminus\Delta}\sum_{a\in A_j} p(s|j,a)x(j,a) = \alpha(s), \quad s\in S\setminus\Delta \tag{6.77b}$$

$$x(s,a)\geq 0, \quad a\in A_s, s\in S\setminus\Delta \tag{6.77c}$$

Using the same approach as in the discounted model (see Theorem 5.23), Theorem 6.17 shows that there is a one-to-one relationship between:

1. feasible solutions to the dual problem and randomized stationary policies, and
2. *basic* feasible solutions to the dual problem and deterministic stationary policies.

Theorem 6.17. Suppose $\alpha(s) > 0$ for all $s \in S \setminus \Delta$.

1. For each $d \in D^{\text{MR}}$

$$x_d(s,a) := \sum_{j\in S\setminus\Delta} \alpha(j) \sum_{n=1}^{\infty} p^{d^\infty}(X_n = s, Y_n = a \,|\, X_1 = j) \tag{6.78}$$

   is a feasible solution to the dual linear program. That is, it satisfies the constraints (6.77b) and (6.77c).

2. Let $x(s,a)$ denote a feasible solution to the dual linear program. Let $d_{\mathbf{x}}$ denote the randomized decision rule that selects action $a \in A_s$ in state $s \in S \setminus \Delta$ with probability

$$w_{d_{\mathbf{x}}}(a|s) := \frac{x(s,a)}{\sum_{a\in A_s} x(s,a)}. \tag{6.79}$$

   Then $x_{d_{\mathbf{x}}}(s,a)$ defined by (6.78) is a feasible solution to the dual linear program and $x_{d_{\mathbf{x}}}(s,a) = x(s,a)$ for all $a \in A_s$, $s \in S \setminus \Delta$.

Noting that there are $|S-1|$ dual constraints[28] and assuming $\alpha(s) > 0$, for each $s \in S \setminus \Delta$, a basic feasible solution has $x(s,a') > 0$ for exactly one $a' \in A_s$.

[28] Not including the non-negativity constraints.

Corollary 6.8. Suppose $\alpha(s) > 0$ for all $s \in S \setminus \Delta$. Then

1. For each $d \in D^{\text{MD}}$ and $s \in S \setminus \Delta$

$$x_d(s,a) := \begin{cases} \sum_{j \in S\setminus\Delta} \alpha(j) \sum_{n=1}^{\infty} P^{d^\infty}[X_n = s, Y_n = a \,|\, X_1 = j] & \text{for} \quad a = d(s) \\ 0 & \text{for} \quad a \neq d(s) \end{cases} \tag{6.80}$$

   is a basic feasible solution of the dual linear program.

2. Let $\mathbf{x}$ be a basic feasible solution to the dual linear program. Then $d_{\mathbf{x}} \in D^{\text{MD}}$.

From these results it follows that:

Corollary 6.9. There exists an optimal basic feasible solution $x^*(s,a)$ to the dual linear program and an optimal stationary deterministic policy $d^\infty_{\mathbf{x}^*}$ where $d_{\mathbf{x}^*}(s) = a^*_s$ if $x^*(s, a^*_s) > 0$.

Observe that when $\sum_{s\in S\setminus\Delta} \alpha(s) = 1$, the expression on the right hand side of (6.78) represents the expected number of times (under policy d^∞) that the system with initial state chosen with probability $\alpha(s)$ visits state s and chooses action a[29]. Hence,

$$\sum_{s\in S\setminus\Delta} \sum_{a\in A_s} r(s,a) x_d(s,a) = \sum_{s\in S\setminus\Delta} \alpha(s) v^{d^\infty}(s). \tag{6.81}$$

This means that when $\sum_{s\in S\setminus\Delta} \alpha(s) = 1$, the objective function value when $x(s,a) = x_d(s,a)$ is the expected total reward corresponding to policy d^∞ averaged with respect to initial distribution $\alpha(s)$. An alternative derivation of equation (6.81) follows from strong duality of linear programming.

6.8.3 Examples

This section provides examples of the application of the above LP to a transient and SSP model.

[29] To see this note that

$$\sum_{n=1}^{\infty} P^{d^\infty}[X_n = s, Y_n = a \,|\, X_1 = j] = E\left[\sum_{n=1}^{\infty} I_{\{X_n = s, Y_n = a\}} \,\middle|\, X_1 = j\right].$$

Example 6.17. A simple transient example.
Suppose $S = \{s', \Delta\}$ and $A_{s'} = \{a_1, a_2\}$ with $r(s', a_1) = 5$ and $p(s'|s', a_1) = 0.2$ and $r(s', a_2) = 3$ and $p(s'|s', a_2) = 0.5$.

Setting $\alpha(s') = 1$ yields the primal linear program

$$\begin{aligned} \text{minimize} \quad & v(s') \\ \text{subject to} \quad & v(s') - 0.2v(s') \geq 5, \\ & v(s') - 0.5v(s') \geq 3. \end{aligned}$$

Note that the constraints can be rewritten as $v(s') \geq 6.25$ and $v(s') \geq 6$, respectively. Therefore the solution is $v^*(s') = 6.25$. Since equality holds in the first constraint, the optimal decision rule is $d^*(s') = a_1$.

The dual formulation is

$$\begin{aligned} \text{maximize} \quad & 5x(s', a_1) + 3x(s', a_2) \\ \text{subject to} \quad & x(s', a_1) + x(s', a_2) - 0.2x(s', a_1) - 0.5x(s', a_2) = 1 \\ & x(s', a_1) \geq 0 \text{ and } x(s', a_2) \geq 0. \end{aligned}$$

The equality constraint can be written as $0.8x(s', a_1) + 0.5x(s', a_2) = 1$. There are two basic feasible solutions: $\mathbf{x}^1 = (x^1(s', a_1), x^1(s', a_2)) = (1.25, 0)$ and $\mathbf{x}^2 = (x^2(s', a_1), x^2(s', a_2)) = (0, 2)$. Clearly, $\mathbf{x}^1$ is optimal with an objective function value of 6.25, in agreement with the solution of the primal linear program. Hence the deterministic stationary policy derived from $d^*(s') = a_1$ is optimal and under this policy, on average, the system spends 1.25 decision epochs in state s' prior to absorption in Δ.

Example 6.18. An SSP model

Consider the SSP model depicted in Figure 6.5 with $S = \{s_1, s_2, \Delta\}$, $A_{s_1} = \{a_{1,1}, a_{1,2}\}$ and $A_{s_2} = \{a_{2,1}, a_{2,2}\}$, with rewards $r(s_1, a_{1,1}) = -3$, $r(s_1, a_{1,2}) = 1$, $r(s_2, a_{2,1}) = 1$ and $r(s_2, a_{2,2}) = -2$, and non-zero transition probabilities $p(s_1|s_1, a_{1,1}) = p(s_2|s_1, a_{1,1}) = 0.5$, $p(s_2|s_1, a_{1,2}) = 1$, $p(\Delta|s_2, a_{2,1}) = 1$ and $p(s_1|s_2, a_{2,2}) = 1$.

Observe that the model is *not* transient since the stationary policy based on the decision rule $d'(s_1) = a_{1,2}, d'(s_2) = a_{2,2}$ is improper as is the stationary policy based on $d''(s_1) = a_{1,1}$ and $d''(s_2) = a_{2,2,}$. However it is a positive model.

Taking $\alpha(s_1) = \alpha(s_2) = 0.5$, the primal linear program becomes

$$\begin{aligned} \text{minimize} \quad & 0.5v(s_1) + 0.5v(s_2) \\ \text{subject to} \quad & v(s_1) - 0.5v(s_1) - 0.5v(s_2) \geq -3, \\ & v(s_1) - v(s_2) \geq 1, \end{aligned}$$

$$v(s_2) \geq 1,$$
$$-v(s_1) + v(s_2) \geq -2.$$

The corresponding dual linear program is

$$\begin{aligned}
\text{maximize} \quad & -3x(s_1, a_{1,1}) + x(s_1, a_{1,2}) + x(s_2, a_{2,1}) - 2x(s_2, a_{2,2}) \\
\text{subject to} \quad & x(s_1, a_{1,1}) + x(s_1, a_{1,2}) - 0.5x(s_1, a_{1,1}) - x(s_2, a_{2,2}) = 0.5 \\
& x(s_2, a_{2,1}) + x(s_2, a_{2,2}) - 0.5x(s_1, a_{1,1}) - x(s_1, a_{1,2}) = 0.5 \\
& x(s, a) \geq 0, \quad a \in A_s, s \in S \setminus \Delta.
\end{aligned}$$

Solving the dual yields $x(s_1, a_{1,2}) = 0.5$, $x(s_2, a_{2,1}) = 1$ with the remaining variables equal to 0. Hence the optimal policy d^* uses decision rule $d^*(s_1) = a_{1,2}$ and $d^*(s_2) = a_{2,1}$.

Note that the same policy is optimal for all values of $r(s_2, a_{2,1})$. On the other hand, when either $r(s_2, a_{2,2}) > -1$, $r(s_1, a_{1,2}) > 2$ or $r(s_1, a_{1,1}) > 1$, the primal linear program is infeasible[a]. In each of these cases the condition that the value of an improper policy equals $-\infty$ were violated so it is not an SSP model.

[a] You can easily confirm this by plotting the feasible region of the primal.

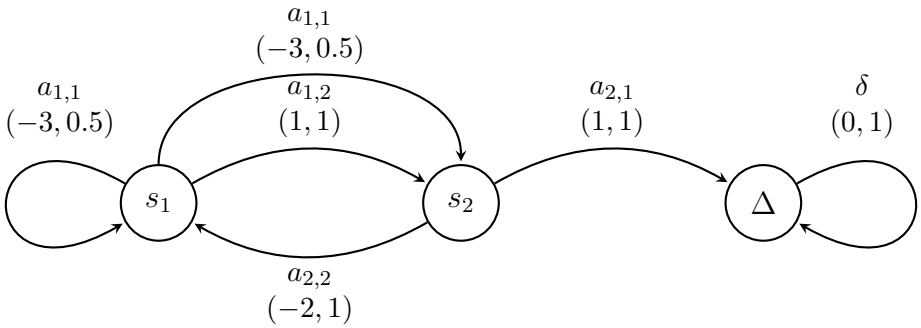

Figure 6.5: Symbolic representation of model in Example 6.18. The labels (r, p) refer, respectively, to the reward and transition probabilities associated with using the specified action in the given state.

6.9 Optimality of structured policies

This section extends results in Section 5.10 to expected total reward models in which value iteration is convergent.

6.9.1 The fundamental result

Define V^F to be a subset of V with a specific form (such as convex or monotone) and define D^F to be a subset D^{MD} in which all decision rules have a particular structure. They are *compatible* if $\mathbf{v} \in V^F$ implies that there exists a

$$d' \in \arg\operatorname*{c\text{-}max}_{d \in D^{\text{MD}}} \{\mathbf{r}_d + \mathbf{P}_d \mathbf{v}\}$$

that is in D^F. For example in an optimal stopping problem, V^F might denote the set of non-decreasing functions and D^F might denote decision rules that stop after a threshold is reached.

Theorem 6.18. Suppose

1. V^F is non-empty,
2. D^F is non-empty and compatible with V^F,
3. $\mathbf{v} \in V^F$ implies $L\mathbf{v} \in V^F$, and
4. if $\lim_{n\to\infty} \|\mathbf{v}^n - \mathbf{v}^*\| = 0$ and $\mathbf{v}^n \in V^F$ for all $n = 0, 1, \ldots$ then $\mathbf{v}^* \in V^F$.

Then there exists an optimal stationary policy $(d^*)^\infty$ with $d^* \in D^F$.

The proof is identical to that of Theorem 5.30 and is omitted. Note that for hypothesis 4 to apply requires at a minimum the convergence of value iteration, which holds in all the models considered previously in this section.

6.10 Applications

6.10.1 Gridworld navigation

Example 6.11 focused on convergence of value iteration. In this example attention shifts to the properties and structure of solutions and how they are impacted by model parameters. Solution is straightforward and it is easy to solve the model by value iteration, policy iteration, modified policy iteration or linear programming. Investigating computational aspects of these approaches is left to the reader.

Consider Instance 1 of Example 6.11 with a reward of 50 units when successfully delivering the coffee, a cost of 200 units if the robot falls down the stairs and a cost per step of 1 unit. Recall that the only available actions are "up", "left" or "right", with not every action available in each cell.

When $p = 1$ this is a deterministic shortest path problem. Its optimal policy starting in cell 13 is

$$13 \to (10 \text{ or } 14) \to 11 \to 8 \to 5 \to 2 \to 1.$$

Both paths yield $v^*(13) = 50 - 6 = 44$. Moreover, solving this problem also provides the shortest path from each cell not on the shortest path to the office.

When $p = 0.95$, a decision rule that specifies action choice in each sate is required. However starting in cell 13, the "optimal path"[30] becomes

$$13 \to 14 \to 15 \to 12 \to 9 \to 6 \to 3 \to 2 \to 1.$$

For such a policy, the expected value starting in cell 13 is $v^*(13) = 40.96$. Observe that this policy is conservative in that it aims to take a longer path so as to avoid the costly cell 7. Moreover if the robot ever occupies cell 11, the optimal policy moves it to cell 12 for the same reason.

Note that when $p = 0.65$, the optimal policy remains the same as when $p = 0.95$ but $v^*(13) = -4.63$ implying that even under the optimal policy there is a non-negligible probability of falling down the stairs.

For small values of p such as $p = 0.2$, the robot's movement becomes unreliable and the optimal policy becomes quite strange. In such cases the robot must aim in the "wrong" direction so as to actually head in the right direction. The optimal policy when $p = 0.2$ appears in Table 6.4.

State	Intended cell
2	3
3	2
4	5
5	6
6	3
8	7
9	8
10	11
11	10
12	11
13	10
14	13
15	14

Table 6.4: State and optimal direction when $p = 0.2$.

Observe that in cell 8, the optimal policy aims the robot in the counter-intuitive direction of cell 7. This is because the probability it moves in some other direction is 0.8. Note that $v^*(13) = -150.26$ corresponding to the high probability of falling down the stairs when the robot is unreliable. The reader is encouraged to investigate whether the optimal value function is monotone as a function of p, particularly as p approaches 0, when it is possible to avoid falling down the stairs.

[30] That is, the path the robot follows if the implemented action results in a movement in the intended direction.

Interesting variants to consider include models with cell-dependent movement probabilities or a tilted grid where it is more likely to move in one direction than the other.

6.10.2 Optimal stopping

This section analyzes a stationary version of the optimal stopping problem (Section 3.8) in which the system moves probabilistically between states in S' until the decision maker decides to stop. Thus $S = S' \cup \{\Delta\}$ where Δ denotes the stopped state. For each state in S' the decision maker can either continue (action C) or stop (action Q). In state Δ, the only option is to continue (action C) and remain in that state and receive a reward of 0. Therefore the action sets may be written as

$$A_s = \begin{cases} \{C, Q\} & s \in S' \\ \{C\} & s = \Delta. \end{cases}$$

For simplicity assume continuing costs c and stopping in state $s' \in S$ generates revenue $g(s)$. Hence the reward function[31] is given by:

$$r(s,a) = \begin{cases} -c & s \in S', a = C \\ g(s) & s \in S', a = Q \\ 0 & s = \Delta, a = a_C. \end{cases}$$

Note that the reward $g(s)$ is only received upon stopping, at which time the system moves to the zero reward absorbing state Δ.

Assume that transitions are governed by an underlying $|S'| \times |S'|$ transition probability matrix $\mathbf{B}$ with components $b(j|s)$. Thus the transition probabilities for the corresponding Markov decision process can be expressed as

$$p(j|s,a) = \begin{cases} b(j|s) & s \in S', a = C, j \in S' \\ 1 & s \in S', a = Q, j = \Delta \text{ or } s = j = \Delta, a = C \\ 0 & \text{otherwise.} \end{cases}$$

Note that in this model the stationary policy that always continues, that is when $d'(s) := a_0$ for all $s \in S'$, has $v^{(d')^\infty}(s) = -\infty$ for all $s \in S'$.

Model classification

The optimal stopping model can be classified as follows:

1. If $g(s) \geq 0$ for all $s \in S'$, it is a positive model.
2. If $g(s) \leq 0$ for all $s \in S'$, it is a negative model.

[31] It would be more precise to formulate this model in terms of $r(s,a,j)$ especially $r(s,a_1,\Delta) = g(s)$, but since under a_1 the transition is deterministic, representing this by $r(s,a_1)$ suffices.

3. It is not a transient model since under $(d')^\infty$ the state Δ cannot be reached.

4. The model is an SSP model. When the underlying Markov chain forms a single closed class, the policy $(d')^\infty$ is improper and satisfies condition 3 in the definition of an SSP model. This holds also for other Markov chain structures but application of condition 3 is more subtle[32].

The Bellman equation

The Bellman equation in component notation can be expressed as:

$$v(s) = \max\left\{-c + \sum_{j\in S'} b(j|s)v(j), g(s) + v(\Delta)\right\}, \quad s \in S', \tag{6.85}$$

$$v(\Delta) = v(\Delta). \tag{6.86}$$

Since $v(\Delta) = 0$, the Bellman equation can be rewritten for $s \in S'$ as:

$$v(s) = \max\left\{-c + \sum_{j\in S'} b(j|s)v(j),\ g(s)\right\}. \tag{6.87}$$

It can be solved using any iterative algorithm or using the (primal) linear program:

$$\text{minimize} \quad \sum_{s\in S'} v(s) \tag{6.88}$$

$$\text{subject to} \quad v(s) - \sum_{j\in S'} b(j|s)v(j) \geq c \tag{6.89}$$

$$v(s) \geq g(s) \tag{6.90}$$

Let $v^*(s)$ for $s \in S'$ denote the optimal value function and d^* denote a decision rule corresponding to the optimal stationary policy. Then it follows from the primal linear program that

$$d^*(s) = \begin{cases} Q & \text{if} \quad v^*(s) = g(s) \\ C & \text{if} \quad v^*(s) > g(s). \end{cases}$$

Since $(d^*)^\infty$ must stop in at least one state in S' (why?), $\mathbf{P}_{d^*}^N$ must converge to a matrix that has entries equal to zero in the first $|S'|$ columns and ones in the column corresponding to Δ. Since $v^*(\Delta) = 0$, (6.36) holds. Therefore it follows from Theorem 6.5 that the policy d^* is optimal.

[32] To see this, consider the case of of Markov chain with two closed classes and some transient states.

Optimal stopping on a random walk

The following example applies the above results to the problem of optimal stopping in a finite random walk on the integers with reflecting barriers at 1 and N.

Example 6.19. Let $S = \{1, \ldots, N\}$, c be the continuation cost, $g(s) = \alpha s^2$ and p denote the probability of a transition from s to $s+1$, except in state N where p is the probability of remaining in state N. Let $q = 1 - p$, denote the probability of a transition from state s to $s-1$, except in state 1 where it denotes the probability of staying there. Thus the (transient) Bellman equation (6.87) is given by:

$$\begin{aligned} v(1) &= \max\{-c + qv(1) + pv(2),\ g(1)\}, \\ v(s) &= \max\{-c + qv(s-1) + pv(s+1),\ g(s)\}, \quad \text{for} \quad s = 2, \ldots, N-1 \\ v(N) &= \max\{-c + qv(N-1) + pv(N),\ g(N)\}. \end{aligned}$$

The decision maker trades off the cost of continuing versus the cost of reaching the high reward states. When $\alpha > 0$ these are states with large values of s and when $\alpha < 0$ they correspond to small values of s.

Since this is an SSP model, it can be solved numerically using value iteration[a] and a sup-norm stopping criterion with $\epsilon = 0.000001$. For simplicity, initialize computation with $v(s) = 0$ for all $s \in S$.

Consider the following instances:

Instance 1: $N = 25$, $\alpha = 0.2$, $p = 0.65$ and $c = 2$,

Instance 2: $N = 25$, $\alpha = 0.2$, $p = 0.5$ and $c = 2$,

Instance 3: $N = 500$, $\alpha = 0.05$, $p = 0.65$ and $c = 10$,

Instance 4: $N = 500$, $\alpha = -0.05$, $p = 0.65$ and $c = 10$,

Instance 5: $N = 500$, $\alpha = -0.05$, $p = 0.35$ and $c = 10$.

Note that Instances 1-3 represent reward maximization problems and Instances 4-5 represent cost minimization problems. The parameters were chosen so as to obtain interesting policies.

Instance	Iterations	Continuation region
1	249	$\{8, \ldots, 24\}$
2	2	none
3	1777	$\{299, \ldots, 499\}$
4	1287	$\{500\}$
5	1677	$\{334, \ldots, 500\}$

Table 6.5: Computational results for Example 6.19.

Table 6.5 provides the number of iterations to converge and the *continuation region* in which the optimal policy chooses the action a_0 for each instance. It shows that:

1. Value iteration converged in all instances.
2. In all instances $v^*(s)$ was a monotone function of s.
3. With the exception of Instance 2 in which $v^*(s) = g(s)$ for $S = 1, \ldots, 25$, and Instance 4 where the continuation region is a singleton, the continuation region is a set of consecutive states.
4. In Instance 2, value iteration converges in 2 iterations with iterates $v^n(s) = g(s)$ for all $n \geq 1$.
5. In Instances 1 and 3, the optimal policy stops for s less than some threshold (7 and 298, respectively) and when $s = N$.
6. In Instance 4 when $g(s) < 0$ for all $s \in S$, and there is a probability of 0.65 of moving to a higher cost state, it is better to stop unless the system is in the highest cost state.
7. In Instance 5, in which the system moves to a lower cost state with probability 0.65, it is optimal to continue in high cost states.

We hypothesize that one can prove analytically that when $g(s)$ is monotone and p is constant, then $v^*(s)$ is monotone. Hence under further conditions on $g(s)$, one can prove that the continuation region is either empty, or a set of consecutive states. This is left as an exercise.

[a] This is the easiest to code and converges quickly for small N.

Selling an asset*

Consider a stationary version of the asset selling problem in Section 3.8.2. Assume you are trying to sell an asset such as a home. Offers arrive daily and you can either accept the best offer from that day (action Q) or wait another day in anticipation of better offers (action S). If you do not accept the current best offer, it is withdrawn at the end of the day.

Let the random variable X denote the offer amount which is assumed to be discrete and independent between days. Let $b(n) := P[X = n]$, $0 \leq n \leq N$, for some N and assume a cost c for holding the asset for an additional day.

Consider first, a related problem, referred to as the *modified asset selling problem*, in which all past offers remain available in perpetuity. In it, $S = S' \cup \{\Delta\}$, $S' = \{0, \ldots, N\}$ and Δ denotes the stopped state. The set of actions $A_s = \{C, Q\}$ for $s \in S'$ and

$A_\Delta = \{C\}$. Transition probabilities satisfy $p(\Delta|s, Q) = 1$ for all $s \in S'$, $p(\Delta|\Delta, C) = 1$ and

$$p(j|s, a_0) = \begin{cases} 0 & 0 \leq j < s \\ \sum_{k \leq s} p(k) & j = s \\ p(j) & s < j \leq N. \end{cases}$$

Rewards satisfy

$$r(s, a) = \begin{cases} -c & s \in S', a = C \\ s & s \in S', a = Q \\ 0 & s = \Delta, a = C. \end{cases}$$

Again, this can be classified as a positive or SSP model. The (transient) Bellman equation can be written[33] as

$$v(s) = \max \left\{ s, -c + v(s) \sum_{j=0}^{s} p(j) + \sum_{j=s+1}^{N} p(j)v(j) \right\}. \tag{6.91}$$

Instead of solving this equation numerically, the following theorem provides a relationship that can be used to easily compute the optimal policy. Its proof uses policy iteration. Let $\mu > 0$ denote the expected offer size, that is,

$$\mu = \sum_{j=1}^{N} jp(j).$$

Moreover, define

$$F(s) := \sum_{j=s+1}^{N} (j - s)p(j)$$

for $s = 0, \ldots, N$. Then, it is easy to see that $F(0) = \mu$, $F(N) = 0$ and $F(s)$ is non-increasing in s.

[33] Note typo in the Bellman equation on p. 307 in Puterman [1994]. This typo could have been avoided by carefully writing out the transition probabilities.

Theorem 6.19. Let

$$s^* = \min\{s \in \{0, 1, \ldots, N\} \mid F(s) < c\}. \tag{6.92}$$

Then the stationary policy $(d^*)^\infty$ where[a]

$$d^*(s) = \begin{cases} C & s < s^* \\ Q & s \geq s^* \end{cases} \tag{6.93}$$

is an optimal policy for the modified asset selling problem. Moreover, if $\mu < c$, $s^* = 0$ and $d^*(s) = Q$. That is, it is optimal to accept the first offer.

[a] If $s^* = 0$, the following decision rule never chooses C.

Proof. Begin policy iteration with $d(s) = Q$ for all $s \in S'$. Then $v(s) = s$ for all $s \in S'$. The improvement step chooses a $d'(s)$ as

$$\begin{aligned} d'(s) &\in \arg\max \left\{ s, -c + s\sum_{j=0}^{s} p(j) + \sum_{j=s+1}^{N} jp(j) \right\} \\ &= \arg\max \left\{ s, -c + s + \sum_{j=s+1}^{N} (j-s)p(j) \right\} = \arg\max \{0, -c + F(s)\}, \end{aligned}$$

where the first equality follows by adding and subtracting $s\sum_{j=s+1}^{N} p(j)$ in the second expression. Therefore

$$d'(s) = \begin{cases} a_1 & \text{if} \quad F(s) - c < 0 \\ \{a_0, a_1\} & \text{if} \quad F(s) - c = 0 \\ a_0 & \text{if} \quad F(s) - c > 0. \end{cases}$$

Hence $d'(s)$ has the form (6.93).

Let $d''(s)$ denote the decision rule chosen at the next iteration of policy iteration. When $\mu < c$, $s^* = 0$ and $d''(s) = a_1$ for all $s \in S'$. Hence policy iteration terminates because $d''(s) = d'(s)$ for all $s \in S$.

Now assume $\mu \geq c$. Consequently $s^* > 0$, so it follows without further computation that $v'(s) := v^{(d')^\infty}(s) = s$ for $s \geq s^*$ and $v'(s) > s$ for $s < s^*$ (Why?).

Now assume $s \geq s^*$. Then

$$\begin{aligned} &-c + v'(s)\sum_{j=0}^{s} p(j) + \sum_{j=s+1}^{N} v'(j)p(j) = -c + s\sum_{j=0}^{s} p(j) + \sum_{j=s+1}^{N} jp(j) \\ &= -c + \sum_{j=s+1}^{N} (j-s)p(j) + s = -c + F(s) + s < s, \end{aligned}$$

where the last inequality follows from the definition of s^*. Therefore $d''(s) = Q = d'(s)$ for all $s \geq s^*$.

Now assume $s < s^*$. Then

$$
\begin{aligned}
&-c + v'(s)\sum_{j=0}^{s} p(j) + \sum_{j=s+1}^{N} v'(j)p(j) \\
&= -c + v'(s)\sum_{j=0}^{s} p(j) + \sum_{j=s+1}^{s^*} v'(j)p(j) + \sum_{j=s^*+1}^{N} v'(j)p(j) \\
&> -c + s\sum_{j=0}^{s} p(j) + \sum_{j=s+1}^{s^*} jp(j) + \sum_{j=s^*+1}^{N} v'(j)p(j) \\
&= -c + s\sum_{j=0}^{s} p(j) + \sum_{j=s+1}^{s^*} jp(j) + \sum_{j=s^*+1}^{N} jp(j) \\
&= -c + \sum_{j=s+1}^{N} (j-s)p(j) + s = -c + F(s) + s > s,
\end{aligned}
$$

where the last inequality follows from the definition of s^*. Therefore $d''(s) = C = d'(s)$. Hence policy iteration terminates and the stationary policy derived from d' is optimal. □

Since the above policy is admissible in the (unmodified) asset selling problem and the value of the modified asset selling problem is necessarily greater than or equal to the value of the (unmodified) asset selling problem, it follows that:

Corollary 6.10. The stationary policy derived from the decision rule $d^*(s)$ defined by equation (6.93) is optimal in the (unmodified) asset selling problem.

Note this result extends to continuous distributions and those on $0, 1, 2 \ldots$ but it requires extension of underlying theory to continuous and countable state spaces, respectively.

To see how this works in practice consider an example in which the offer size is generated from a truncated (at $N = 20$) Poisson with parameter 10. Since $F(8) = 2.44$ and $F(9) = 1.78$ when $c = 2$, $s^* = 9$. Moreover, it follows from Figure 6.6 that if $c > 9.98$ (the mean of the truncated Poisson), it would be optimal to stop after receiving the first offer and if $c = 0$ it would be optimal to continue until receiving the maximum offer.

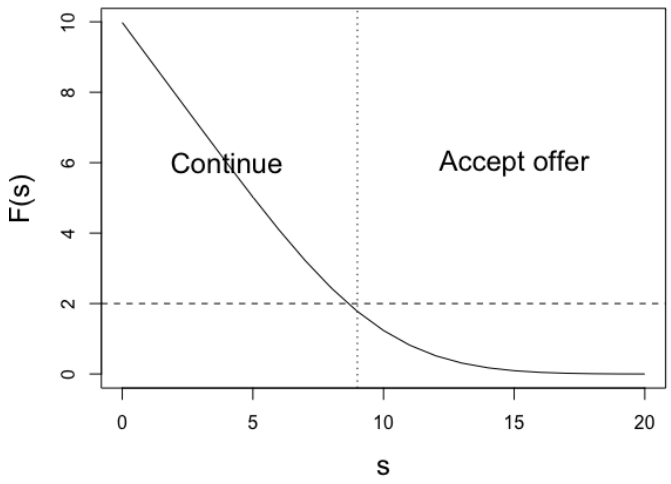

Figure 6.6: Plot of $F(s)$ vs. s in asset selling problem when offers follow a truncated Poisson distribution with parameter 10. The dashed horizontal line corresponds to $c = 2$ and the dotted vertical line to $s^* = 9$.

Bibliographic remarks

This chapter analyzes finite state and action expected total reward models focusing primarily on transient and SSP models, which arise naturally in modern applications. On the other hand, Chapter 7 in Puterman [1994] considers countable state models and restricts attention to positive and negative models.

Foundational papers include Strauch [1966] (negative models), Blackwell [1967] (positive models), and Veinott [1969a] (transient models).

Transient models were first referred to as *optimal first passage problems* and studied by Eaton and Zadeh [1962] and Derman [1970]. Bertsekas and Tsitsiklis [1991] extended them to include stochastic shortest path models and provide a rich bibliographic background.

The definition of a transient model in Veinott [1969a] differs from ours in that he requires at least one row sum of $\mathbf{P}_d$ be strictly less than 1. Thus he does not explicitly require the existence of a single reward-free absorbing state. Our use of the Euclidean norm to analyze transient models does not seem to appear in the literature. Wessels [1977] introduced the use of the weighted supremum norm for analyzing value iteration in expected total reward models.

The discussion of optimal stopping is based on Çinlar [1975] and analysis of the asset selling problem follows Karlin [1962]. Puterman [1994] analyzes an optimal parking problem where it is formulated as a countable state model and Ross [1983] analyzes an application in gambling. In the exercises, the tennis serving problem is adapted from

Norman [1985], the tennis handicapping problem is adapted from Chan and Singal [2016], and the darts handicapping problem is adapted from Chan et al. [2024]

Exercises

1. Consider the following deterministic expected total reward model with $S = \{s_1, s_2, s_3\}$; $A_{s_1} = \{a_{1,1}\}$, $A_{s_2} = \{a_{2,1,}, a_{2,2}\}$ and $A_{s_3} = \{a_{3,1}\}$; $r(s_1, a_{1,1}) = 1$, $r(s_2, a_{2,1}) = 1$, $r(s_2, a_{2,2}) = -3$, $r(s_3, a_{3,1}) = 0$ and $p(s_2|s_1, a_{1,1}) = 1$, $p(s_1|s_2, a_{2,1}) = 1$, $p(s_3|s_2, a_{2,2}) = 1$, $p(s_3|s_3, a_{3,1}) = 1$.

   (a) Represent the model graphically.

   (b) Classify the model.

   (c) Find the value of both deterministic stationary policies. What is $\mathbf{v}^*$ in each case?

   (d) Write out and solve the Bellman equation.

2. **Value iteration in a transient model.** Consider a stationary policy d^∞ in a transient model on $S = \{s_1, s_2, s_3, \Delta\}$ in which

$$\mathbf{P}_d = \begin{bmatrix} 0.3 & 0.4 & 0.3 & 0 \\ 0.8 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{r}_d = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}.$$

   Let $\mathbf{Q}_d$ denote the 3×3 sub-matrix of $\mathbf{P}_d$ restricted to the transient states $\{s_1, s_2, s_3\}$.

   (a) Verify that the model is transient.

   (b) Compute the max norm and the Euclidean norm (spectral radius) of $\mathbf{Q}_d$.

   (c) Show that for some N, all row sums of $\mathbf{Q}_d^N$ are strictly less than 1 and compute the norm of $\mathbf{Q}_d^N$.

   (d) Find the value of this policy using value iteration and interpret the result (see Section 6.5.2).

   (e) Compare the rate of convergence with respect to the span, the max norm and the Euclidean norm and describe the results.

3. Prove that in a transient model:

   (a) (Easy) For any deterministic stationary policy d^∞, L_d is a contraction mapping in the Euclidean norm.

   (b) (Harder) L_T is a contraction mapping in the Euclidean norm. Hint: Complete the following steps.

i. Use the definition of the Euclidean norm to write out $\|L_T\mathbf{v} - L_T\mathbf{u}\|_2^2$.

ii. Partition S into subsets S_1 in which $L_T\mathbf{v}(s) \geq L_T\mathbf{u}(s)$ and S_2 where $L_T\mathbf{v}(s) < L_T\mathbf{u}(s)$.

iii. Define a decision rule d' that attains $L_T\mathbf{v}$ on S_1 and $L_T\mathbf{u}$ on S_2.

iv. Show that for each $s \in S$

$$(L_T\mathbf{v}(s) - L_T\mathbf{u}(s))^2 \leq (L_{d'}\mathbf{v}(s) - L_{d'}\mathbf{u}(s))^2$$

v. Apply Part 1.

4. Show that if value iteration converges to $\mathbf{v}^*$ with $\mathbf{v}^0 = \mathbf{0}$, it converges to $\mathbf{v}^* + c\mathbf{e}$ if $\mathbf{v}^0 = c\mathbf{e}$.

5. Prove that if two Markov decision processes are positively similar, then they have the same sets of transient policies, they have the same optimal policy and value iteration converges at the same rate in each.

6. Give an example of a transition probability matrix in a transient model with norm greater than one and spectral radius less than one.

   (a) Find a positively similar model with transition probability matrix with norm less than one.

   (b) Show that value iteration converges in both the untransformed model and the transformed model for an arbitrary reward function.

   (c) What is the rate of convergence?

7. Consider a version of the Gridworld problem in Section 6.10.1 in which an additional column has been added to the right of grid displayed in Figure 3.3. Show how the optimal policy changes as a function of the probability, p, of successfully exercising the intended movement in the following two cases:

   (a) There is a penalty for falling down the stairs combined with a reward for successfully delivering the coffee;

   (b) The robot seeks to maximize the probability of successfully delivering the coffee.

   Solve the problems using value iteration and linear programming.

8. Formulate and solve the following variant of the Gridworld problem described in Section 3.2. Assume that the robot starts at the professor's office, needs to navigate to the coffee room, fill the coffee cup and then return to the professor's office. Assume in each cell that the robot can decide to move up, down, right or left. If a wall (or boundary) is in the way, an attempted move in that direction is unsuccessful and the robot remains where it is. Assume probabilistic transitions

in each state; if the robot attempts to move in a specific direction, then it executes that move with probability p and moves in another direction with probability $(1-p)/3$.

Note that to formulate and solve this model requires augmenting the state with the content of the coffee cup. Assume a reward of 50 for delivering a filled coffee cup, a penalty of 200 for falling down the stairs and a cost of one unit for each attempted move. Find a policy that maximizes the expected total reward with $p = 0.5, 0.9, 1$ and discuss how it varies as a function of p. Clearly state any assumptions made.

9. **The rational burglar.** Consider a Markov chain on $S = \{0, 1, \ldots, N\}$ with probability p of moving from s to $s+1$ and a probability of $1-p$ of moving from s to 0. Suppose in addition that states 0 and N are absorbing.

   (a) Formulate this as an optimal stopping problem in which the burglar receives a reward of $g(s) = s$ if he/she decides to stop in state s.

   (b) Why do you think this is called the burglar problem?

   (c) Classify the model.

   (d) Find an optimal policy numerically for $p = 0.01, 0.25, 0.5, 0.75, 0.99$ using policy iteration and linear programming.

   (e) Describe the structure of the optimal policy and how it varies with p.

   (f) *Formally prove the optimality of the structured policy.

10. **Sequential gambling model.** As long as you have some money to bet, you can repeatedly play a game of chance that has a win probability of p. If your fortune reaches 0 you can no longer play.

    Prior to each round of the game you may place a bet of b. If you win you receive $2b$ (the amount bet and an equal pay out) and if you lose, you receive 0. You must decide how much to bet in each round of the game given that your fortune can assume any (integer) value between 0 and N dollars.

    (a) Formulate this as an expected total reward model that seeks to develop a betting strategy that maximizes the probability of reaching N.

    (b) How is this related to the rational burglar problem and how is it different?

    (c) Classify this model.

    (d) Solve the model numerically for $N = 100$ and $p = 0.01, 0.25, 0.5, 0.75, 0.99$ and describe how the optimal strategy varies as a function of p.

    (e) *Hypothesize and prove the form of the optimal policy as a function of p.

11. Determine the optimal policy in the asset selling problem when the demand distribution is binomially and normally distributed.

12. **Lion hunting behavior.** Consider a simplified infinite horizon version of the lion hunting behavior model in Section 3.5 in which the lion seeks a hunting policy to maximize survival time. Assume a lion's energy capacity is $C = 30$ kgs, that a lion requires $d = 4$ kgs of energy per day to survive, and hunting requires 1 additional kg of energy per day. A successful hunt, which occurs with probability 0.25, yields 12 kgs of energy.

    (a) Formulate this as an expected total reward model in which the objective is to maximize the expected number of days of survival and the decision is whether to hunt or not.

    (b) Classify this model.

    (c) Provide the Bellman equation and transient Bellman equation.

    (d) Find and interpret the optimal policy and the expected survival time if the lion starts the planning horizon at its highest energy level.

    (e) Suppose there are two other species to hunt: one yields 7 kgs of energy with a catch probability of 0.5 and another which yields 20 kg of energy with a catch probability of 0.15. Reformulate and solve this more general model where the objective is to decide whether or not to hunt and if so, what species to hunt.

13. **Serving in Tennis.** Consider the problem of choosing the type of serve to use at each point in a game of tennis. Tennis scoring is described in Section 3.9.2. The salient issue is that whenever the game reaches *deuce*, the game continues until a player wins two points in a row[34]. Moreover at each point, if the first serve is "out", that is, it lands out of bounds, the server has a second serve. If the second serve is out, the server loses the point.

    From the perspective of this chapter, a tennis game must be analyzed using an infinite horizon expected total reward model because the length of the game is random.

    Consider the simplified situation where the server can use either a "fast" or "slow" serve. A fast serve lands in bounds with probability p_f and the server wins the point with probability q_f. Similarly a slow serve lands in bounds with probability p_s and the server wins the point with probability q_s. To be realistic, $q_f > q_s$ and $p_f < p_s$.

    The objective is choose a service strategy that maximizes the probability of winning the game as a function of the score and whether it is a first serve or second serve.

    (a) Formulate this problem as an expected total reward model.

[34] Deuce corresponds to a score of 3-3 in points in a game, or "40-40" in tennis jargon.

(b) Classify it.

(c) Provide the Bellman equation.

(d) Find an optimal policy numerically when $p_f = 0.4$, $p_s = 0.8$, $q_f = 0.9$ and $q_s = 0.5$.

(e) Characterize the optimal strategy as a function of the model parameters.

(f) Reformulate and solve the problem when the parameters depend on the game score. Use your knowledge of tennis (or empirical data) to determine realistic parameter choices.

14. **Handicapping in Tennis.** Consider the tennis handicapping example from Chapter 3.

(a) Provide the Bellman equation.

(b) Prove that the optimal value function is equal to the probability of the weaker player winning the match at the current state.

(c) Find an optimal policy numerically for a one-set match when $p_1 = 0.5$, $p_0 = 0.46$ and $B = 2$.

(d) With $p_1 = 0.45$ and $p_0 = 0.45$, determine the smallest value of B such that the weaker player's probability of winning the match is at least 0.5. Repeat this calculation with $p_1 = 0.6$ and $p_0 = 0.3$, comment on whether B increases or decreases, and explain why.

(e) Reformulate this problem accounting for "hot streaks". Suppose if the serving player wins the point their probability of winning the next point increases by $\epsilon > 0$. If the serving player loses, the probability of winning the point reverts to the original value.

15. **501 darts.** In 501 darts, players start with a score of 501. Darts thrown at a dart board are scored based on the region they land in and subtracted from their current score. The scoring regions are $\{1, 2, \ldots, 20\}$ (singles), $\{2, 4, \ldots, 40\}$ (doubles), $\{3, 6, \ldots, 60\}$ (triples), and $\{25, 50\}$ (single and double bullseye). Players take turns, with each turn comprising three dart throws. The first player to reach 0 wins, with the added restriction that the dart that takes the score to zero must be a double. If a player goes below zero in a given turn or ends with a score that is impossible to complete with a double (i.e., 1), then the turn ends and the player's score reverts to the score at the start of the turn. The objective is to choose a strategy that minimizes the number of turns until a player reaches zero.

(a) Formulate this problem as an infinite horizon expected total reward model, starting with the simpler model where each turn comprises a single dart.

(b) Classify it.

(c) Provide the Bellman equation.

(d) Repeat the above steps where each turn comprises three dart throws.

(e) Find an optimal policy numerically assuming that a player lands their dart in the region they are aiming for with probability 0.8, with the remaining 0.2 distributed equally among all regions that share a border with the target region, regardless of the size of those regions. This will require looking up an image of a dart board to see which regions are adjacent to which.

16. Prove in an optimal stopping problem on a random walk that when $g(s)$ is monotone and p is constant, that $v^*(s)$ is monotone and the continuation region is either empty, or a set of consecutive states.

17. Complete the details of the proof of Theorem 6.16.

18. Show by example that in an SSP, the evaluation equations will not have a solution when you choose an improper policy.

19. Show that a finite horizon model is a (transient) infinite horizon expected total reward model.

20. **Tetris.** Formulate the game of Tetris as an expected total reward Markov decision process in which the objective is to maximize the length of the game. Classify this model.