1
Markov Decision Problems

In this chapter, we introduce Markov decision problems, which are stochastic games with a single player. They serve as an appetizer. On the one hand, the basic concepts and basic proofs for zero-sum stochastic games are better understood in this simple model. On the other hand, some of the conclusions that we draw for Markov decision problems are different from those drawn for zero-sum stochastic games. This illustrates the inherent difference between single-player decision problems and multiplayer decision problems (=games). The interested reader is referred to, for example, Ross (1982) or Puterman (1994) for an exposition of Markov decision problems.

We will study both the T-stage evaluation and the discounted evaluation. We will introduce and study contracting mappings,[1] and will use such mappings to show that the decision maker has a stationary discounted optimal strategy. We will also define the concept of uniform optimality, and show that the decision maker has a stationary uniformly optimal strategy.

Definition 1.1 A *Markov decision problem*[2] is a vector $\Gamma = \langle S, (A(s))_{s \in S}, q, r \rangle$ where

- S is a finite set of states.
- For each $s \in S$, $A(s)$ is a finite set of actions available at state s. The set of pairs (state, action) is denoted by

$$SA := \{(s,a) \colon s \in S, a \in A(s)\}.$$

- $q \colon SA \to \Delta(S)$ is a transition rule.
- $r \colon SA \to \mathbf{R}$ is a payoff function.

[1] We adhere to the convention that a mapping is a function whose range is a general space or $\mathbb{R}^n$, while a function is always real-valued.

[2] Andrey Andreyevich Markov (Ryazan, Russia, June 14, 1856 – St. Petersburg, Russia, July 20, 1922) was a Russian mathematician. He is best known for his work on the theory of stochastic processes that now bear his name: Markov chains and Markov processes.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

A Markov decision problem involves a decision maker, and it evolves as follows. The problem lasts for infinitely many stages. The initial state $s_1 \in S$ is given. At each stage $t \geq 1$, the following happens:

- The current state s_t is announced to the decision maker.
- The decision maker chooses an action $a_t \in A(s_t)$ and receives the stage payoff $r(s_t, a_t)$.
- A new state s_{t+1} is drawn according to $q(\cdot \mid s_t, a_t)$, and the game proceeds to stage $t+1$.

Example 1.2 Consider the following situation. The technological level of a country can be High (H), Medium (M), or Low (L). The annual investment of the country in technological advances can also be high (2 billion dollars), medium (1 billion dollars), or low (0.5 billion dollars). The annual gain from technological level is increasing: the high, medium, and low technological level yield 10, 6, and 2 billion dollars, respectively. The technological level changes stochastically as a function of the investment in technological advancement, according to the following table:[3]

Technology level	High investment	Medium investment	Low investment
H	H	$\left[\frac{1}{2}(H), \frac{1}{2}(M)\right]$	$\left[\frac{1}{4}(H), \frac{3}{4}(M)\right]$
M	$\left[\frac{3}{5}(H), \frac{2}{5}(M)\right]$	M	$\left[\frac{2}{5}(M), \frac{3}{5}(L)\right]$
L	$\left[\frac{3}{5}(M), \frac{2}{5}(L)\right]$	$\left[\frac{2}{5}(M), \frac{3}{5}(L)\right]$	L

The situation can be presented as a Markov decision problem as follows:

- There are three states, which represent the three technological levels: $S = \{H, M, L\}$.
- There are three actions in each state, which represent the three investment levels: $A(s) = \{h, m, l\}$ for each $s \in S$.
- The transition rule is given by

[3] Here and in the sequel, a probability distribution is denoted by a list of probabilities and outcomes in square brackets, where the outcomes are written within round brackets. Thus, $\left[\frac{2}{3}(H), \frac{1}{3}(M)\right]$ means a probability distribution that assigns probability $\frac{2}{3}$ to H and probability $\frac{1}{3}$ to M.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

$$\begin{array}{lll}
q(H \mid H,h) = 1, & q(M \mid H,h) = 0, & q(L \mid H,h) = 0, \\
q(H \mid H,m) = \frac{1}{2}, & q(M \mid H,m) = \frac{1}{2}, & q(L \mid H,m) = 0, \\
q(H \mid H,l) = \frac{1}{4}, & q(M \mid H,l) = \frac{3}{4}, & q(L \mid H,l) = 0, \\
q(H \mid M,h) = \frac{3}{5}, & q(M \mid M,h) = \frac{2}{5}, & q(L \mid M,h) = 0, \\
q(H \mid M,m) = 0, & q(M \mid M,m) = 1, & q(L \mid M,m) = 0, \\
q(H \mid M,l) = 0, & q(M \mid M,l) = \frac{2}{5}, & q(L \mid M,l) = \frac{3}{5}, \\
q(H \mid L,h) = 0, & q(M \mid L,h) = \frac{3}{5}, & q(L \mid L,h) = \frac{2}{5}, \\
q(H \mid L,m) = 0, & q(M \mid L,m) = \frac{2}{5}, & q(L \mid L,m) = \frac{3}{5}, \\
q(H \mid L,l) = 0, & q(M \mid L,l) = 0, & q(L \mid L,l) = 1.
\end{array}$$

- The payoff function (in billions of dollars) is given by

$$\begin{array}{lll}
r(H,h) = 8, & r(H,m) = 9, & r(H,l) = 9\frac{1}{2}, \\
r(M,h) = 4, & r(M,m) = 5, & r(M,l) = 5\frac{1}{2}, \\
r(L,h) = 0, & r(L,m) = 1, & r(L,l) = 1\frac{1}{2}.
\end{array}$$

♦

Example 1.3 The Markov decision problem that is illustrated in Figure 1.1 is formally defined as follows:

- There are three states: $S = \{s(1), s(2), s(3)\}$.
- In state $s(1)$, there are two actions: $A(s(1)) = \{U, D\}$; in states $s(2)$ and $s(3)$, there is one action: $A(s(2)) = A(s(3)) = \{D\}$.
- Payoffs appear at the center of each entry and are given by:

$$r(s(1),U) = 10; \quad r(s(1),D) = 5; \quad r(s(2),D) = 10; \quad r(s(3),D) = -100.$$

- Transitions appear in parentheses next to the payoff and are given by:
  - If in state $s(1)$ the decision maker chooses U, the process moves to state $s(2)$, that is, $q(s(2) \mid s(1), U) = 1$.
  - If in state $s(1)$ the decision maker chooses D, the process remains in state $s(1)$, that is, $q(s(1) \mid s(1), D) = 1$.

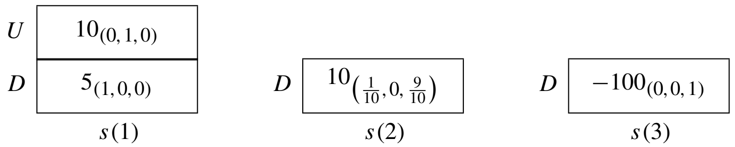

Figure 1.1 The Markov decision problem in Example 1.3.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

- From state $s(2)$, the process moves to state $s(1)$ with probability $\frac{1}{10}$ and to state $s(3)$ with probability $\frac{9}{10}$, that is, $q(s(1) \mid s(2), D) = \frac{1}{10}$ and $q(s(3) \mid s(2), D) = \frac{9}{10}$.
- Once the process reaches state $s(3)$, it stays there, that is, $q(s(3) \mid s(3), D) = 1$. ♦

1.1 On Histories

For $t \in \mathbb{N}$, the set of *histories of length t* is defined by

$$H_t := (SA)^{t-1} \times S,$$

where by convention $(SA)^0 = \emptyset$. This is the set of all histories that may occur until stage t. A typical element in H_t is denoted by h_t. The last state of history h_t is denoted by s_t. The set H_1 is identified with the state space S, and the history (s_1) is simply denoted by s_1.

We denote the set of all *histories* by

$$H := \bigcup_{t \in \mathbb{N}} H_t,$$

and the set of all *infinite histories* or *plays* by

$$H_\infty := (SA)^{\mathbb{N}}.$$

The set of plays H_∞ is a measurable space, with the sigma-algebra generated by the cylinder sets, which are defined as follows. For a history $\widetilde{h}_t = (\widetilde{s}_1, \widetilde{a}_1, \ldots, \widetilde{s}_t) \in H_t$, the *cylinder set* $C(\widetilde{h}_t) \subset H_\infty$ is the collection of all plays that start with $\widetilde{h}_t$, that is,

$$C(\widetilde{h}_t) := \{h = (s_1, a_1, s_2, a_2, \ldots) \in H_\infty \colon s_1 = \widetilde{s}_1, a_1 = \widetilde{a}_1, \ldots, s_t = \widetilde{s}_t\}.$$

For every $t \in \mathbb{N}$, the collection of all cylinder sets $(C(\widetilde{h}_t))_{\widetilde{h}_t \in H_t}$ defines a finite partition, or an algebra, on H_∞. We denote by $\mathcal{H}_t$ this algebra and by $\mathcal{H}$ the sigma-algebra on H_∞ generated by the algebras $(\mathcal{H}_t)_{t \in \mathbb{N}}$.

1.2 On Strategies

A *mixed action* at state s is a probability distribution over the set of actions $A(s)$ available at state s. The set of mixed actions at state s is therefore $\Delta(A(s))$. A strategy of the decision maker specifies how the decision maker should play after each possible history.

Definition 1.4 A *strategy* is a mapping σ that assigns to each history $h = (s_1, a_1, \ldots, a_{t-1}, s_t)$ a mixed action in $\Delta(A(s_t))$.

The set of all strategies is denoted by Σ.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

A decision maker who follows a strategy σ behaves as follows: at each stage t, given the past history $(s_1, a_1, \ldots, s_t)$, the decision maker chooses an action a_t according to the mixed action $\sigma(\cdot \mid s_1, a_1, \ldots, s_t)$.

Comment 1.5 A strategy as defined in Definition 1.4 is termed in the literature *behavior strategy*.

Comment 1.6 The fact that the choice of the decision maker depends on past play implicitly assumes that the decision maker knows the past play; that is, the decision maker observes (and remembers) all past states that the process visited, and she remembers all her past choices. In Chapter 2, we will study the model of Markov decision problems when the decision maker does not observe the state.

Comment 1.7 A strategy contains a lot of irrelevant information. Indeed, when the initial state is $s_1 = s$, it is not important what the decision maker would play if the initial state were $s' \neq s$. Similarly, if in the first stage the decision maker played the action $a_1 = a$, it is irrelevant what she would play in the second stage if she played the action $a' \neq a$ in the first stage. We nevertheless regard a strategy as a mapping defined on the set of *all* histories, because of the simplicity of the definition; otherwise we would have to define for every strategy σ and every positive integer t the set of all histories of length t that can occur with positive probability when the decision maker follows strategy σ (which depend on the definition of σ up to stage $t-1$), and define σ at stage t only for those histories.

Every strategy σ, together with the initial state s_1, defines a probability distribution $\mathbf{P}_{s_1,\sigma}$ on the space of measurable space $(H_\infty, \mathcal{H})$. To define this probability distribution formally, we define it on the collection of cylinder sets that generate $(H_\infty, \mathcal{H})$ by the rule

$$\begin{aligned} &\mathbf{P}_{s_1,\sigma}(C(\widetilde{s}_1, \widetilde{a}_1, \ldots, \widetilde{s}_{t-1}, \widetilde{a}_{t-1}, \widetilde{s}_t)) \\ &\quad := \mathbf{1}_{\{s_1 = \widetilde{s}_1\}} \cdot \prod_{k=1}^{t-1} \sigma(\widetilde{a}_k \mid \widetilde{s}_1, \widetilde{a}_1, \ldots, \widetilde{s}_1) \cdot \prod_{k=1}^{t-1} q(\widetilde{s}_{k+1} \mid \widetilde{s}_k, \widetilde{a}_k). \end{aligned} \tag{1.1}$$

Let $\mathbf{P}_{s_1,\sigma}$ be the unique probability distribution on H_∞ that agrees with this definition on cylinder sets. The fact that, in this way, we indeed obtain a unique probability distribution is guaranteed by the Carathéodory[4] Extension Theorem (see, e.g., theorem 3.1 in Billingsley (1995)).

[4] Constantin Carathéodory (Berlin, Germany, September 13, 1873 – Munich, Germany, February 2, 1950) was a Greek mathematician who spent most of his career in Germany. He made significant contributions to the theory of functions of a real variable, the calculus of variations, and measure theory. His work also includes important results in conformal representations and in the theory of boundary correspondence.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

Two simple classes of strategies are pure strategies that involve no randomization, and stationary strategies that depend only on the current state and not on the whole past history.

Definition 1.8 A strategy σ is *pure* if $|\mathrm{supp}(\sigma(h_t))| = 1$ for every history $h_t \in H$.

The set of pure strategies is denoted by Σ_{P}.

Definition 1.9 A strategy σ is *stationary* if, for every two histories $h_t = (s_1, a_1, s_2, \dots, a_{t-1}, s_t)$ and $\widehat{h}_k = (\widehat{s}_1, \widehat{a}_1, \widehat{s}_2, \dots, \widehat{a}_{k-1}, \widehat{s}_k)$ that satisfy $s_t = \widehat{s}_k$, we have $\sigma(h_t) = \sigma(\widehat{h}_k)$.

The set of stationary strategies is denoted by Σ_{S}.

A pure stationary strategy assigns to each state $s \in S$ an action in $A(s)$. Since the number of actions in $A(s)$ is $|A(s)|$, we can express the number of pure stationary strategies in terms of the data of the Markov decision problem.

Theorem 1.10 *The number of pure stationary strategies is* $\prod_{s \in S} |A(s)|$.

One can identify a stationary strategy σ with a vector $x \in \prod_{s \in S} \Delta(A(s))$. With this identification, $x(s)$ is the mixed action chosen when the current state is s. Thus, the set of stationary strategies Σ_{S} can be identified with the space $X := \prod_{s \in S} \Delta(A(s))$, which is convex and compact. For every element $x \in X$, the stationary strategy that corresponds to x is still denoted by x.

In Definition 1.4 we defined a strategy to be a mapping from histories to mixed actions. We now present another concept of a strategy that involves randomization – a mixed strategy.

Definition 1.11 A *mixed strategy* is a probability distribution over the set Σ_{P} of pure strategies.

Every strategy is equivalent to a mixed strategy. Indeed, a strategy σ is defined by $\aleph_0$ lotteries: to each history $h_t \in H$, it assigns a lottery $\sigma(h_t) \in \Delta(A(s_t))$. If the decision maker performs all the $\aleph_0$ lotteries before the play starts, then the realizations of the lotteries define a pure strategy. In particular, the strategy defines a probability distribution over the set of pure strategies.

Conversely, every mixed strategy is equivalent to a strategy. Indeed, given a mixed strategy τ, one can calculate for each history h_t the conditional probability $\sigma(a_t \mid h_t)$ that the action chosen after h_t is $a_t \in A(s_t)$. If the history h_t occurs with probability 0 under $\mathbf{P}_{s_1,\sigma}$, we set $\sigma(a_t \mid h_t)$ arbitrarily. One can show that the strategy σ is equivalent to the mixed strategy τ.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

The equivalence just described is a special case of a more general result called *Kuhn's Theorem*;[5] see, for example, Maschler, Solan, and Zamir (2020, chapter 7).

1.3 The T-Stage Payoff

The decision maker receives the stage payoff $r(s_t, a_t)$ at every stage t. How does she compare sequences of stage payoffs? We will study two methods of evaluations. The first, which we consider in this section, is the T-stage evaluation. This evaluation is relevant when the process lasts T stages, and the goal of the decision maker is to maximize her expected average payoff during these stages. The second, which we will study in the next section, is the discounted evaluation, which is relevant when the play continues indefinitely, and the goal of the decision maker is to maximize the expected discounted sum of her stage payoffs.

The expectation operator for the probability distribution $\mathbf{P}_{s_1,\sigma}$ is denoted by $\mathbf{E}_{s_1,\sigma}[\,\cdot\,]$. In particular, $\mathbf{E}_{s_1,\sigma}[r(s_t, a_t)]$ is the expected payoff at stage t.

Definition 1.12 For every positive integer $T \in \mathbb{N}$, every initial state $s_1 \in S$, and every strategy $\sigma \in \Sigma$, define the *T-stage payoff* by:

$$\gamma_T(s_1; \sigma) := \mathbf{E}_{s_1,\sigma}\left[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t)\right]. \tag{1.2}$$

Example 1.13 The Markov decision problem in this example is given in Figure 1.2.

The initial state is $s(1)$. We will calculate the T-stage payoff of every pure strategy.

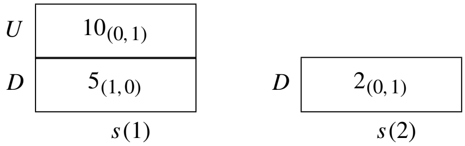

Figure 1.2 The Markov decision problem in Example 1.13.

[5] Harold William Kuhn (Santa Monica, California, July 29, 1925 – New York City, New York, July 2, 2014) was an American mathematician. He is known for the Karush–Kuhn–Tucker conditions, for Kuhn's theorem, and for developing Kuhn poker as well as the description of the Hungarian method for the assignment problem.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

The strategy σ_D that always plays D yields a payoff 5 at every stage, and therefore its T-stage payoff is 5 as well:

$$\gamma_T(s(1); \sigma_D) = 5, \quad \forall T \in \mathbb{N}.$$

The strategy σ_U that plays U in the first stage yields 10 in the first stage and 2 in all subsequent stages. Therefore,

$$\gamma_T(s(1); \sigma_U) = 10 \cdot \frac{1}{T} + 2 \cdot \frac{T-1}{T} = 2 + \frac{8}{T}, \quad \forall T \in \mathbb{N}.$$

For every $0 \le t < T$, the strategy $\sigma_{D_t U}$ that plays D in the first t stages and U in stage $t+1$ yields 5 in the first t stages, 10 in stage $t+1$, and 2 in all subsequent stages. Therefore,

$$\begin{aligned}\gamma_T(s(1); \sigma_{D_t U}) &= 5 \cdot \frac{t}{T} + 10 \cdot \frac{1}{T} + 2 \cdot \frac{T-t-1}{T} \\ &= \frac{2T + 3t + 8}{T}, \quad \forall T \in \mathbb{N}, \forall 0 \le t < T. \qquad \blacklozenge\end{aligned}$$

Definition 1.14 Let $s \in S$ and let $T \in \mathbb{N}$. The real number $v_T(s)$ is the *T-stage value at the initial state s* if

$$v_T(s) := \sup_{\sigma \in \Sigma} \gamma_T(s; \sigma). \tag{1.3}$$

Any strategy in $\text{argmax}_{\sigma \in \Sigma} \gamma_T(s; \sigma)$ is *T-stage optimal at s*.

In other words, the T-stage value at s is the maximal amount that the decision maker can get when the initial state is s, and a strategy that guarantees this quantity is T-stage optimal.

Is the supremum in Eq. (1.3) attained? That is, is there a T-stage optimal strategy? As Theorem 1.15 states, the answer is positive.

Theorem 1.15 *For every $s \in S$ and every $T \ge 1$, there is a T-stage optimal strategy at the initial state s.*

Proof In the T-stage game, the only relevant part of the strategy is its play up to stage T. In particular, for the purpose of studying the T-stage problem, we can define a strategy as a mapping $\sigma : \bigcup_{t=1}^{T} H_t \to \bigcup_{s \in S} \Delta(A(s))$, such that $\sigma(h_t) \in \Delta(A(s_t))$, for every history $h_t \in \bigcup_{t=1}^{T} H_t$. This set is a compact subset of a Euclidean space. The payoff function is continuous on this set. Since a continuous function defined on a compact set attains its maximum, the result follows. □

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

Comment 1.16 We can strengthen Theorem 1.15 and prove that, for every $s \in S$ and every $T \geq 1$, there is a T-stage *pure* optimal strategy at the initial state s (see Theorem 1.18). To see this, consider the function that maps each mixed strategy σ into the T-stage payoff $\gamma_T(s;\sigma)$. This function is linear. Indeed, let σ_1 and σ_2 be two strategies, and let σ_3 be the following strategy: toss a fair coin; if the result is Head, follow σ_1, whereas if it is Tail, follow σ_2. Then

$$\gamma_T(s;\sigma_3) = \frac{1}{2}\gamma_T(s;\sigma_1) + \frac{1}{2}\gamma_T(s;\sigma_1).$$

By the Krein–Milman[6] Theorem, a linear function that is defined on a compact space attains its maximum at an extreme point. Since the pure strategies are the extreme points of the set of mixed strategies, it follows that the function $\sigma \mapsto \gamma_T(s;\sigma)$ attains its maximum at a pure strategy.

Example 1.3, continued The quantity $\gamma_T(s(1);\sigma_{D_tU}) = \frac{2T+3t+8}{T}$ is maximized when $t = T-1$: the decision maker plays $T-1$ times D, and then she plays U once. The resulting average payoff is $5 + \frac{5}{T}$. The T-stage value at the initial state $s(1)$ is therefore $v_T(s(1)) = 5 + \frac{5}{T}$. ♦

In general, the T-stage value, as well as the T-stage optimal strategies, can be found by *backward induction*, a method that is also known as the *dynamic programming principle*. We now formalize this method.

Theorem 1.17 *For every initial state $s_1 \in S$ and every $T \geq 2$, we have*

$$v_T(s_1) = \max_{a_1 \in A(s_1)} \left\{ \frac{1}{T} r(s_1,a_1) + \frac{T-1}{T} \sum_{s_2 \in S} q(s_2 \mid s_1,a_1) v_{T-1}(s_2) \right\}. \quad (1.4)$$

Eq. (1.4) states that, to calculate the T-stage value, we can break the problem into two parts: the first stage, and the last $T-1$ stages. Since transitions and payoffs depend only on the current state and on the current action, the problem that starts at stage 2 is not affected by s_1 and a_1, the state and action at stage 1. This problem is a $(T-1)$-stage Markov decision problem, whose value $v_{T-1}(s_2)$ depends on its initial state (and not on the initial state s_1). To calculate the T-stage value, we collapse the last $T-1$ stages into a single number, the value of the $(T-1)$-stage problem that starts

[6] Mark Grigorievich Krein (Kiev, Russia, April 3, 1907 – Odessa, Ukraine, October 17, 1989) was a Soviet mathematician who is best known for his work in operator theory.
David Pinhusovich Milman (Kiev, Russia, January 15, 1912 – Tel Aviv, Israel, July 12, 1982) was a Soviet and later Israeli mathematician specializing in functional analysis.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

at stage 2, and we ask what is the optimal action in the first stage, assuming that if state s_2 is reached at stage 2, the continuation value is $v_{T-1}(s_2)$.

In Eq. (1.4) the weight of the payoff in the first stage, $r(s_1,a_1)$, is $\frac{1}{T}$, and the weight of the value of the $(T-1)$-stage problem that encapsulates the last $T-1$ stages is $\frac{T-1}{T}$. Why do we take these weights? The reason is that the quantity $r(s_1,a_1)$ represents the payoff in the first stage, while the quantity $v_{T-1}(s_2)$ captures the average payoff in $T-1$ stages: stages $2,3,\ldots,T$. The weights of each of the two quantities reflect this point.

To prove Theorem 1.17, we will consider conditional expectation. Recall that $\mathbf{E}_{s_1,\sigma}[r(s_t,a_t)]$ is the expected payoff at stage t. For every $t' \leq t$ and every history $\widetilde{h}_{t'} = (\widetilde{s}_1,\widetilde{a}_1,\ldots,\widetilde{s}_{t'}) \in H_{t'}$ with $\widetilde{s}_1 = s_1$, the quantity $\mathbf{E}_{s_1,\sigma}[r(s_t,a_t) \mid \widetilde{h}_{t'}]$ is the expected payoff at stage t, conditional that the history $\widetilde{h}_{t'}$ has occurred, that is, conditional that the action in the initial state is $\widetilde{a}_1$, the state at stage 2 is $\widetilde{s}_2$, and so on. Formally, for every history $\widetilde{h}_{t'} = (\widetilde{s}_1,\widetilde{a}_1,\ldots,\widetilde{s}_{t'}) \in H_{t'}$, the probability distribution $\mathbf{P}_{s_1,\sigma}(\cdot \mid \widetilde{h}_{t'})$ is defined as follows:

- For histories that are not longer than $\widetilde{h}_{t'}$: For every $t \leq t'$ we have

$$\mathbf{P}_{s_1,\sigma}(C(s_1,a_1,\ldots,s_t) \mid \widetilde{h}_{t'}) := \mathbf{1}_{\{s_1=\widetilde{s}_1,a_1=\widetilde{a}_1,\ldots,s_t=\widetilde{s}_t\}}.$$

- For histories that are longer than $\widetilde{h}_{t'}$: For every $t > t'$, we have

$$\begin{aligned}\mathbf{P}_{s_1,\sigma}&(C(s_1,a_1,\ldots,s_{t-1},a_{t-1},s_t) \mid h_{t'})\\ &:= \mathbf{1}_{\{s_1=\widetilde{s}_1,a_1=\widetilde{a}_1,\ldots,s_{t'}=\widetilde{s}_{t'}\}} \cdot \prod_{k=t'}^{t-1} \sigma(a_k \mid s_1,a_1,\ldots,s_k)\\ &\times \prod_{k=t'}^{t-1} q(s_{k+1} \mid s_k,a_k).\end{aligned}$$

Denote by $\mathbf{E}_{s_1,\sigma}[\cdot \mid \widetilde{h}_{t'}]$ the expectation with respect to $\mathbf{P}_{s_1,\sigma}(\cdot \mid \widetilde{h}_{t'})$.

Proof of Theorem 1.17 For $T=1$, the T-stage problem concerns the first stage only, and

$$v_1(s_1) = \max_{a_1 \in A(s_1)} r(s_1,a_1).$$

In particular, Eq. (1.4) holds. For $T \geq 2$, by definition and by the law of iterated expectations,

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

$$
\begin{aligned}
&v_T(s_1)\\
&= \max_{\sigma\in\Sigma} \mathbf{E}_{s_1,\sigma}\left[\frac{1}{T}\sum_{t=1}^{T} r(s_t,a_t)\right]\\
&= \max_{\sigma\in\Sigma} \mathbf{E}_{s_1,\sigma}\left[\frac{1}{T} r(s_1,a_1) + \frac{T-1}{T}\cdot\frac{1}{T-1}\sum_{t=2}^{T} r(s_t,a_t)\right]\\
&= \max_{\sigma\in\Sigma}\left(\mathbf{E}_{s_1,\sigma}\left[\frac{1}{T} r(s_1,a_1)\right] + \mathbf{E}_{s_1,\sigma}\left[\frac{T-1}{T}\cdot\frac{1}{T-1}\sum_{t=2}^{T} r(s_t,a_t)\mid h_2\right]\right).
\end{aligned}
\tag{1.5}
$$

The term within the maximization in these equalities depends only on the part of the strategy σ that follows the initial state s_1. This part is composed of the mixed action $\sigma(s_1)\in\Delta(A(s_1))$ that is played in the first stage and the continuation strategies played from the second stage and on. We denote these continuation strategies by $(\sigma'_{s_1,a_1})_{a_1\in A(s_1)}$. Formally, for every action $a_1\in A(s_1)$, σ'_{s_1,a_1} is a strategy in the $T-1$ stage problem that is defined by

$$
\sigma'_{s_1,a_1}(h_{t-1}) := \sigma(s_1,a_1,h_{t-1}),\quad \forall 2\le t\le T, \forall h_{t-1}=(s_2,a_2,\ldots,s_t)\in H_{t-1}.
$$

With this notation, the right-hand side in Eq. (1.5) is equal to

$$
\begin{aligned}
&\max_{\alpha\in\Delta(A(s_1))}\ \max_{(\sigma'_{s_1,a_1})_{a_1\in A_1(s_1)}} \mathbf{E}_{s_1,\alpha,(\sigma'_{s_1,a_1})_{a_1\in A_1(s_1)}}\left[\frac{1}{T} r(s_1,a_1)\right.\\
&\left.+\mathbf{E}_{s_1,\sigma'_{s_1,a_1}}\left[\frac{T-1}{T}\cdot\frac{1}{T-1}\sum_{t=2}^{T} r(s_t,a_t)\mid a_1,s_2\right]\right],
\end{aligned}
\tag{1.6}
$$

where α captures the mixed action played in the first stage. The continuation strategies $(\sigma'_{s_1,a_1})_{a_1\in A_1(s_1)}$ do not affect the payoff in the first stage $r(s_1,a_1)$. The action a_1 that is chosen in the first stage affects the continuation payoff in two ways. First, it determines the probability $q(s_2\mid s_1,a_1)$ that the state in the first stage is s_2. Second, it determines the continuation strategy σ'_{s_1,a_1}. Since the probability distribution $\mathbf{P}_{s_1,\sigma}$ conditional on a_1 and s_2 is equal to the probability distribution $\mathbf{P}_{s_2,\sigma'_{s_1,a_1}}$, it follows that we can split the maximization problem in Eq. (1.6) into two parts, and obtain that

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

$$
\begin{aligned}
v_T(s_1) = \max_{\alpha \in \Delta(A(s_1))} \Bigg(& \frac{1}{T} r(s_1,\alpha) + \sum_{s_2 \in S} q(s_2 \mid s_1, a_1) \\
& \times \max_{(\sigma'_{s_1,a_1})_{a_1 \in A_1(s_1)}} \mathbf{E}_{s_2,\sigma'_{s_1,a_1}} \left[\frac{T-1}{T} \cdot \frac{1}{T-1} \sum_{t=2}^{T} r(s_t,a_t) \right] \Bigg).
\end{aligned}
\tag{1.7}
$$

Note that

$$
v_{T-1}(s_2) = \max_{(\sigma'_{s_1,a_1})_{a_1 \in A_1(s_1)}} \mathbf{E}_{s_2,\sigma'_{s_1,a_1}} \left[\frac{1}{T-1} \sum_{t=2}^{T} r(s_t,a_t) \right];
$$

hence, the right-hand side of Eq. (1.7) is equal to

$$
\max_{\alpha \in \Delta(A(s_1))} \left(\frac{1}{T} r(s_1,\alpha) + \sum_{a_1 \in A_1(s_1)} \alpha(a_1) q(s_2 \mid s_1, a_1) \frac{T-1}{T} v_{T-1}(s_2) \right).
$$

The function within the parentheses is linear in α, and $\Delta(A(s_1))$ is a compact set whose extreme points are the Dirac measures concentrated at the points a_1 with $a_1 \in A(s_1)$. A linear function that is defined on a compact set attains its maximum in an extreme point. The result follows. □

The proof of Theorem 1.17 yields an algorithm that calculates the T-stage value and a T-stage optimal strategy σ^*. We will calculate by induction a k-stage optimal strategy σ^*_k for every $k = 1, 2, \ldots, T$. We start with $k = 1$, and calculate a one-stage optimal strategy for every initial state $s \in S$. Let $a^*_1(s) \in A(s)$ be an action that maximizes the quantity $r(s,a)$ over $a \in A(s)$, and set

$$
\sigma^*_1(s) := a^*_1(s).
$$

The value of the one-stage problem with initial state s is $v_1(s) = r(s_1, a^*_1(s))$. We continue recursively. Suppose that, for every initial state s, we already calculated $v_{k-1}(s)$ and already defined a $(k-1)$-stage optimal strategy σ^*_{k-1}. To calculate $v_k(s)$ and define a k-stage optimal strategy σ^*_k, we take

$$
\max_{a \in A(s)} \left(\frac{1}{k} r(s,a) + \frac{k-1}{k} q(s' \mid s, a) v_{k-1}(s) \right),
\tag{1.8}
$$

and denote by $a^*_k(s) \in A(s)$ an action that achieves the maximum in Eq. (1.8). This is the quantity on the right-hand side of Eq. (1.4); hence, it is equal to $v_k(s)$. We can now define an optimal strategy σ^* for the decision maker as follows:

- At stage 1, play the action $a^*_k(s_1)$.
- From stage 2 on, follow the strategy σ^*_{k-1}; that is, at each stage t, when the current state is s_1 and $T - t + 1$ stages are left, play the action $a^*_{T-t+1}(s_t)$. Formally,

$$\sigma^*(h_t) := a^*_{T-t+1}(s_t), \quad \forall h_t = (s_1, a_1, \ldots, s_t) \in \bigcup_{t=1}^{T} H_t.$$

In Exercise 1.1, the reader is asked to prove that this strategy is indeed T-stage optimal.

The proof of Theorem 1.17 relies on the linearity of the payoff function: the goal of the decision maker is to maximize a linear function of the stage payoffs. If the sets of actions and states are not finite, the theorem still holds, provided that in Eq. (1.4) we replace maximum by supremum.

Theorem 1.17 admits the following corollary.

Theorem 1.18 *The T-stage value always exists. Moreover, there exists an optimal pure strategy $\sigma \in \Sigma$.*

One can show a stronger result concerning the structure of an optimal pure strategy: there exists an optimal pure strategy σ with the property that $\sigma(h_t)$ depends on the current state s_t and on the stage t, and is independent of the rest of the history $(s_1, a_1, \ldots, s_{t-1}, a_{t-1})$ (Exercise 1.3).

1.4 The Discounted Payoff

The discounted payoff depends on a parameter $\lambda \in (0, 1]$, called the *discount factor*, which measures how money grows with time: one dollar today is worth $\frac{1}{1-\lambda}$ dollars tomorrow, $\frac{1}{(1-\lambda)^2}$ dollars the day after tomorrow, and so on. In other words, the decision maker is indifferent between getting $1 - \lambda$ dollars today and one dollar tomorrow.

Definition 1.19 For every discount factor $\lambda \in (0, 1]$, every state $s \in S$, and every strategy $\sigma \in \Sigma$, the *λ-discounted payoff* under strategy profile σ at the initial state s is

$$\gamma_\lambda(s; \sigma) := \mathbf{E}_{s,\sigma}\left[\lambda \sum_{t=1}^{\infty} (1 - \lambda)^{t-1} r(s_t, a_t)\right]. \tag{1.9}$$

The λ in Eq. (1.9) serves as a normalization factor: a player who receives one dollar at every stage evaluates this stream of payoffs as one dollar. Since there are finitely many states and actions, the payoff function r is bounded,

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

and therefore γ_λ obeys the same bound (which is independent of λ, thanks to the multiplication by λ).

The dominated convergence theorem (see, e.g., Shiryaev (1995), theorem 6.3) implies that

$$\gamma_\lambda(s;\sigma) := \lambda \sum_{t=1}^{\infty} (1-\lambda)^{t-1} \mathbf{E}_{s,\sigma}\left[r(s_t,a_t)\right].$$

Simple algebraic manipulations yield

$$\gamma_\lambda(s;\sigma) := \mathbf{E}_{s,\sigma}\left[\lambda r(s_1,a_1) + (1-\lambda)\left(\lambda \sum_{t=2}^{\infty}(1-\lambda)^{t-2} r(s_t,a_t)\right)\right]. \quad (1.10)$$

For every two states $s, s' \in S$ and every action $a \in A(s)$, set

$$\gamma_\lambda(s';\sigma_{s,a}) := \mathbf{E}_{s,\sigma}\left[\lambda \sum_{t=2}^{\infty}(1-\lambda)^{t-2} r(s_t,a_t) \mid s_1 = s,\ a_1 = a,\ s_2 = s'\right].$$

This is the expected discounted payoff from stage 2 on, when conditioning on the history at stage 2. Alternatively, this is the expected discounted payoff when the initial state is s', and the decision maker follows that part of her strategy that follows the history (s,a). If σ is a stationary strategy, then the way it plays after the first stage does not depend on the play in the first stage. Hence, in this case, for every two states $s, s' \in S$ and every action $a \in A(s)$ we have

$$\gamma_\lambda(s';\sigma_{s,a}) = \gamma_\lambda(s';\sigma).$$

From Eq. (1.10) we obtain:

$$\gamma_\lambda(s;\sigma) := \mathbf{E}_{s,\sigma}\left[\lambda r(s_1,a_1) + (1-\lambda)\gamma_\lambda(s_2;\sigma_{s_1,a_1})\right]. \quad (1.11)$$

Thus, the expected payoff is a weighted average of the payoff $r(s_1,a_1)$ at the first stage and the expected payoff $\gamma_\lambda(s_2;\sigma_{s_1,a_1})$ in all subsequent stages. When the discount factor λ is high, the weight of the first stage is high; whereas when the discount factor λ is low, the weight of the first stage is low.

Eq. (1.11) illustrates that the decision maker's payoff consists of two parts: today's payoff and the future's payoff. The discount factor indicates the relative importance of each part. The lower the discount factor, the higher the importance of the future, and therefore the decision maker should put more weight on future opportunities. The higher the discount factor, the higher the importance of the present, and the decision maker should concentrate on short-term gains.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

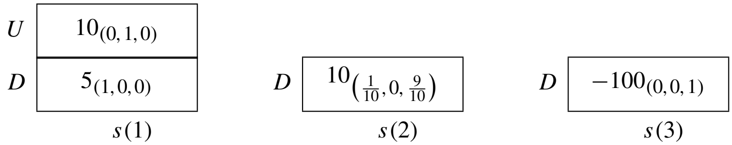

Figure 1.3 The Markov decision problem in Example 1.3.

Comment 1.20 In the proof of Theorem 1.17 we in fact showed that the T-stage payoff satisfies the following formula:

$$\gamma_T(s;\sigma) := \mathbf{E}_{s,\sigma}\left[\frac{1}{T}r(s_1,a_1) + \frac{T-1}{T}\gamma_{T-1}(s_2;\sigma_{s_1,a_1})\right]. \tag{1.12}$$

Thus, similar to the discounted payoff, the T-stage payoff is a weighted average of the payoff $r(s_1,a_1)$ at the first stage and the expected payoff $\gamma_{T-1}(s_2;\sigma_{s_1,a_1})$ in all subsequent stages, with weights $\frac{1}{T}$ and $\frac{T-1}{T}$.

Example 1.3, continued The Markov decision problem in Example 1.3 is reproduced in Figure 1.3.

The initial state is $s(1)$. The strategy σ_D that always plays D at state $s(1)$ yields a payoff 5 at every stage, and therefore its λ-discounted payoff is 5 as well. Let us calculate the λ-discounted payoff of the strategy σ_U that always plays U at state $s(1)$. Since this strategy is stationary,

$$\begin{aligned}&\gamma_\lambda(s(1);\sigma_U)\\ &\quad = 10\lambda + (1-\lambda)\left(10\lambda + (1-\lambda)\left(\frac{9}{10}(-100) + \frac{1}{10}\gamma_\lambda(s(1);\sigma_U)\right)\right).\end{aligned} \tag{1.13}$$

The term $\gamma_\lambda(s(1);\sigma_U)$ on the right-hand side is the discounted payoff from the third stage and on, if at the second stage the play moves from $s(2)$ to $s(1)$. Eq. (1.13) solves to

$$\gamma_\lambda(s(1);\sigma_U) = \frac{10\lambda + 10\lambda(1-\lambda) - 100\frac{9}{10}(1-\lambda)^2}{1-\frac{1}{10}(1-\lambda)^2}.$$

For $\lambda = 1$ (only the first day matters), we get

$$\gamma_1(s(1);\sigma_U) = 10,$$

while for λ close to 0 (the far future matters), we get

$$\lim_{\lambda\to 0}\gamma_\lambda(s(1);\sigma_U) = -100.$$

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

Since the function $\lambda \mapsto \gamma_\lambda(s(1); \sigma_U)$ is continuous, and since $\gamma_\lambda(s(1); \sigma_D) = 5$ for every $\lambda \in [0,1)$, for a high discount factor the strategy σ_U is superior to the strategy σ_D, while for a low discount factor the strategy σ_D is superior to the strategy σ_U. ♦

Definition 1.21 Let $s \in S$ and let $\lambda \in (0,1]$ be a discount factor. The real number $v_\lambda(s)$ is the *λ-discounted value at the initial state s* if

$$v_\lambda(s) := \sup_{\sigma \in \Sigma} \gamma_\lambda(s; \sigma). \tag{1.14}$$

The strategies in $\mathrm{argmax}_{\sigma \in \Sigma} \gamma_\lambda(s; \sigma)$ are said to be *λ-discounted optimal at the initial state s*.

Thus, the λ-discounted value at s is the maximal λ-discounted payoff that the decision maker can get when the initial state is s, and a strategy that guarantees this quantity is λ-discounted optimal.

In Theorem 1.17 we stated the dynamic programming principle for the T-stage decision problem. We now provide the analogous principle for the discounted problem. The proof of the result is left to the reader (Exercise 1.5).

Theorem 1.22 *For every state $s \in S$ and every discount factor $\lambda \in (0,1]$, we have*

$$v_\lambda(s) = \max_{a \in A(s)} \left\{ \lambda r(s,a) + (1-\lambda) \sum_{s' \in S} q(s' \mid s,a) v_\lambda(s') \right\}. \tag{1.15}$$

In Eq. (1.15), the weight of the payoff at the first stage is λ, while the weight of the value at the second stage is $1 - \lambda$. The reason for these weights comes from the definition of the λ-discounted payoff in Eq. (1.9). In that equation, the weight of the payoff at stage t is $\lambda(1-\lambda)^{t-1}$. In particular, the weight of the payoff at the first stage is λ, which is similar to the weight of the payoff at the first stage in Eq. (1.15). Since the sum of the weights of the payoffs in Eq. (1.9) is 1, it follows that the total weight of the payoffs at stages $2, 3, \ldots$ is $1 - \lambda$, which is the weight of the second term on the right-hand side of Eq. (1.15).

In Section 1.7, we will prove that, for every discount factor λ, there is a pure stationary strategy that is λ-discounted optimal at *all* initial states. The proof uses contracting mappings, which will be defined in Section 1.5. Moreover, we will show that there is a pure stationary strategy that is optimal for every discount factor sufficiently close to 0.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

Comment 1.23 Like we did for the T-stage problem, one can provide a direct argument for the existence of a discounted optimal strategy. Since the set of histories is countable, the set of strategies, which is $\prod_{h_t \in H} \Delta(A(s_t))$, is compact in the product topology. Moreover, the discounted payoff function is continuous in this topology. Hence, the supremum in Eq. (1.14) is attained.

1.5 Contracting Mappings

A *metric space* is a pair (X,d), where X is a set and $d\colon X \times X \to [0,\infty)$ is a *metric*, that is, d satisfies the following conditions:

- $d(x,y) = 0$ if and only if $x = y$.
- Symmetry: $d(x,y) = d(y,x)$ for all $x,y \in X$.
- Triangle inequality: $d(x,z) \leq d(x,y) + d(y,z)$ for all $x,y,z \in X$.

A sequence $(x_n)_{n \in \mathbb{N}}$ in a metric space is *Cauchy*[7] if, for every $\epsilon > 0$, there is an $n_0 \in \mathbb{N}$ such that $n_1, n_2 \geq n_0$ implies $d(x_{n_1} x_{n_2}) \leq \epsilon$. A metric space is *complete* if every Cauchy sequence has a limit. For every $m \in \mathbb{N}$, the Euclidean space $\mathbb{R}^m$ equipped with the distance induced by the Euclidean norm, the L_1-norm, or the L_∞-norm is complete. Readers who are not familiar with metric spaces can think of a metric space as $\mathbb{R}^m$ equipped with the Euclidean distance.

Definition 1.24 Let (X,d) be a metric space. A mapping $f\colon X \to X$ is *contracting* if there exists $\rho \in [0,1)$ such that $d(f(x), f(y)) \leq \rho d(x,y)$ for all $x, y \in X$.

Example 1.25 Let $\rho \in [0,1)$ and $a \in \mathbb{R}^n$. The mapping $f\colon \mathbb{R}^n \to \mathbb{R}^n$ that is defined by

$$f(x) := a + \rho x, \quad \forall x \in \mathbb{R}^n,$$

is contracting.

Theorem 1.26 *Let (X,d) be a complete metric space. Every contracting mapping $f\colon X \to X$ has a unique fixed point; that is, there exists a unique point $x \in X$ such that $x = f(x)$.*

Proof Let $f\colon X \to X$ be a contracting mapping.

[7] Augustin-Louis Cauchy (Paris, France, August 21, 1789 – Sceaux, France, May 23, 1857) was a French mathematician. He started the project of formulating and proving the theorems of calculus in a rigorous manner and was thus an early pioneer of analysis. He also developed several important theorems in complex analysis and initiated the study of permutation groups.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

Step 1: f has at most one fixed point.

If $x, y \in X$ are fixed points of f, then

$$d(x,y) = d(f(x), f(y)) \leq \rho d(x,y).$$

Since $\rho \in [0,1)$, this implies that $d(x,y) = 0$, and therefore $x = y$.

Step 2: f has at least one fixed point.

Let $x_0 \in X$ be arbitrary, and define inductively $x_{n+1} = f(x_n)$ for every $n \geq 0$. Then for any $k, m > 0$,

$$\begin{aligned} d(x_k, x_{k+m}) &\leq \sum_{l=0}^{m-1} d(x_{k+l}, x_{k+l+1}) \\ &\leq d(x_0, f(x_0))\rho^k \sum_{l=0}^{m-1} \rho^l < d(x_0, f(x_0)) \frac{\rho^m}{1-\rho}, \end{aligned}$$

where the first inequality follows from the triangle inequality, and the second inequality holds since by induction: $d(x_l, x_{l+1}) \leq \rho^l d(x_0, x_1) = \rho^l d(x_0, f(x_0))$. Thus, $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, and therefore it converges to a limit x. By the triangle inequality,

$$d(x, f(x)) \leq d(x, x_k) + d(x_k, x_{k+1}) + d(x_{k+1}, f(x)), \tag{1.16}$$

for all $k \in \mathbb{N}$. Let us show that all three terms on the right-hand side of Eq. (1.16) converge to 0 as k goes to infinity; this will imply that $d(x, f(x)) = 0$, hence $x = f(x)$, that is, x is a fixed point of f. Indeed, $\lim_{k\to\infty} d(x, x_k) = 0$ because x is the limit of $(x_k)_{k\in\mathbb{N}}$; $\lim_{k\to\infty} d(x_k, x_{k+1})$ because $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence; and finally, since f is contracting,

$$\lim_{k\to\infty} d(x_{k+1}, f(x)) = \lim_{k\to\infty} d(f(x_k), f(x)) \leq \lim_{k\to\infty} \rho d(x_k, x) = 0. \qquad \square$$

1.6 Existence of an Optimal Stationary Strategy

In this section, we prove the following result, due to Blackwell (1965).[8]

Theorem 1.27 *For every $\lambda \in (0,1]$, there exists a λ-discounted pure stationary optimal strategy.*

[8] David Harold Blackwell (Centralia, Illinois, April 24, 1919 – Berkeley, California, July 8, 2010) was an American statistician and mathematician who made significant contributions to game theory, probability theory, information theory, and Bayesian statistics.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

The existence of a λ-discounted optimal strategy was discussed in Comment 1.23, while the existence of a λ-discounted pure optimal strategy is established by the same arguments as in Comment 1.16. We now explain the intuition behind the existence of a λ-discounted pure *stationary* optimal strategy. Let h_t and $\widehat{h_{\widehat{t}}}$ be two histories that end at the same state s. Since the payoffs and transitions depend only on the current state, and not on past play, if the decision maker plays in the same way after h_t and after $\widehat{h_{\widehat{t}}}$, the evolution of the Markov decision problem after h_t is the same as after $\widehat{h_{\widehat{t}}}$. Suppose now that the optimal strategy σ prescribes to play differently after h_t and after $\widehat{h_{\widehat{t}}}$, that is, $\sigma(h_t) \neq \sigma(\widehat{h_{\widehat{t}}})$. Assume without loss of generality that the expected payoff after h_t is at least as high as the expected payoff after $\widehat{h_{\widehat{t}}}$. Define a new strategy σ_1 as follows: σ_1 is similar to σ, except that after the history $\widehat{h_{\widehat{t}}}$, it plays as σ plays after h_t. It is easy to see that $\gamma_\lambda(s_1;\sigma_1) \geq \gamma_\lambda(s_1;\sigma)$. Repeating this process over all histories shows that one can modify σ to be a stationary strategy, without lowering the λ-discounted payoff, thus establishing Theorem 1.27. The proof of Theorem 1.27 that we will provide will use a different idea – contracting mappings. This approach will be useful when we will later study stochastic games.

Before we can prove Theorem 1.27, we need a bit of preparation. Fix a function $w\colon S \to \mathbb{R}$. This function will capture the "discounted payoff from the next stage on," given the state at the next stage. Given the initial state s and the strategy σ, let $h_t \in H$ be a history with positive probability of realization, such that $\mathbf{P}_{s,\sigma}(C(h_t)) > 0$. Consider the situation in which, when the decision maker follows the strategy σ, once some history h_t is realized, the decision maker is told that after she chooses the action a_t and the new state s_{t+1} is announced, the process will terminate, and she will get a terminal payoff $w(s_{t+1})$. As in Eq. (1.15), the weights of the payoff at stage t is λ, and the weight of the terminal payoff[9] is $1-\lambda$. The expected payoff from stage t and on is then given by

$$\begin{aligned}&\mathbf{E}_{s,\sigma}\left[\lambda r^i(s_t,a_t) + (1-\lambda)\sum_{s'\in S} q(s' \mid s_t,a_t)w(s') \mid h_t\right] \\ &\quad = \mathbf{E}_{s,\sigma}\big[\lambda r^i(s_t,a_t) + (1-\lambda)w(s_{t+1}) \mid h_t\big].\end{aligned} \tag{1.17}$$

The first term in the expectation measures the expected stage payoff, while the second term measures the expected terminal payoff. Note that in Eq. (1.17)

[9] Setting the weight of the terminal payoff to $1-\lambda$ is equivalent to considering a standard discounted payoff, assuming the payoff in *all* stages after stage t is $w(s_{t+1})$.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

the expectation is a conditional expectation, given the history at stage t. The following result relates the expectation in Eq. (1.17) to the discounted payoff.

Lemma 1.28 *Let σ be a strategy, let $s \in S$, and let $w \colon S \to \mathbb{R}$ be a function. If for every $t \in \mathbb{N}$ and every $h_t \in H_t$,*

$$\mathbf{E}_{s,\sigma}\left[\lambda r(s_t, a_t) + (1-\lambda) w(s_{t+1}) \mid h_t\right] \ge w(s_t) \tag{1.18}$$

then

$$\gamma_\lambda(s;\sigma) \ge w(s). \tag{1.19}$$

If the inequality in Eq. (1.18) *is reversed for every $t \in \mathbb{N}$ and every $h_t \in H_t$, so is the inequality in* Eq. (1.19). *If the inequality in* Eq. (1.18) *is an equality for every $t \in \mathbb{N}$ and every $h_t \in H_t$, then* Eq. (1.19) *becomes an equality as well.*

Proof Recall the law of iterated expectation: for every function $f \colon S \to \mathbb{R}$, every $t \in \mathbb{N}$, and every history $h_t \in H_t$,

$$\mathbf{E}_{s,\sigma}[\mathbf{E}_{s,\sigma}[f(s_{t+1}) \mid h_t]] = \mathbf{E}_{s,\sigma}[f(s_{t+1})].$$

Taking expectations in Eq. (1.18), we deduce that

$$\mathbf{E}_{s,\sigma}[\lambda r(s_t, a_t)] \ge \mathbf{E}_{s,\sigma}[w(s_t)] - (1-\lambda)\mathbf{E}_{s,\sigma}[w(s_{t+1})], \quad \forall t \in \mathbb{N}. \tag{1.20}$$

Multiplying both sides of Eq. (1.20) by $(1-\lambda)^{t-1}$ and summing over $t \in \mathbb{N}$, we obtain Eq. (1.19):

$$\begin{aligned}\gamma_\lambda(s;\sigma) &= \sum_{t=1}^{\infty} (1-\lambda)^{t-1} \mathbf{E}_{s,\sigma}[\lambda r(s_t, a_t)] \\ &\ge \sum_{t=1}^{\infty} (1-\lambda)^{t-1} \big(\mathbf{E}_{s,\sigma}[w(s_t)] - (1-\lambda)\mathbf{E}_{s,\sigma}[w(s_{t+1})]\big) \qquad (1.21) \\ &= w(s),\end{aligned}$$

where the last equality holds because the sum involved is telescopic.

If the inequality in Eq. (1.18) is reversed for every $t \in \mathbb{N}$ and every $h_t \in H_t$, then the inequality in Eq. (1.20) is reversed as well, and therefore so is the equality in Eq. (1.21). The last conclusion follows from the first two statements. □

We need the following technical result.

Lemma 1.29 *Let $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Then*

$$\left| \max_{1 \le i \le n} x_i - \max_{1 \le i \le n} y_i \right| \le \max_{1 \le i \le n} |x_i - y_i|.$$

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

Proof Without loss of generality, we can assume that $\max_{1\le i\le n} x_i \ge \max_{1\le i\le n} y_i$. Suppose also that $x_{i_0} = \max_{1\le i\le n} x_i$ and $y_{i_1} = \max_{1\le i\le n} y_i$. Then

$$\begin{aligned}\left|\max_{1\le i\le n} x_i - \max_{1\le i\le n} y_i\right| &= \max_{1\le i\le n} x_i - \max_{1\le i\le n} y_i\\ &= x_{i_0} - y_{i_1}\\ &\le x_{i_0} - y_{i_0}\\ &\le \max_{1\le i\le n} |x_i - y_i|. \end{aligned}$$ □

Proof of Theorem 1.27 We define a mapping $T: \mathbb{R}^S \to \mathbb{R}^S$, prove that it is contracting, and conclude that it has a unique fixed point w. We then show that the decision maker has a pure stationary strategy x^* such that $\gamma_\lambda(s; x^*) = w(s)$ for every initial state $s \in S$, and that $\gamma_\lambda(s;\sigma) \le w(s)$ for every initial state $s \in S$ and every strategy σ.

For every vector $w = (w(s))_{s\in S} \in \mathbb{R}^S$, define

$$(T(w))(s) := \max_{a\in A(s)}\left(\lambda r(s,a) + (1-\lambda)\sum_{s'\in S} q(s' \mid s,a) w(s)\right).$$

Step 1: The mapping T is contracting.

Let $w, u \in \mathbb{R}^S$. By Lemma 1.29,

$$\begin{aligned}|(T(w))(s) - (T(u))(s)| &= \left|\max_{a\in A(s)}\left(\lambda r(s,a) + (1-\lambda)\sum_{s'\in S} q(s' \mid s,a) w(s')\right)\right.\\ &\qquad \left. - \max_{a\in A(s)}\left(\lambda r(s,a) + (1-\lambda)\sum_{s'\in S} q(s' \mid s,a) u(s')\right)\right|\\ &\le \max_{a\in A(s)}\left|\left(\lambda r(s,a) + (1-\lambda)\sum_{s'\in S} q(s' \mid s,a) w(s')\right)\right.\\ &\qquad \left. - \left(\lambda r(s,a) + (1-\lambda)\sum_{s'\in S} q(s' \mid s,a) u(s')\right)\right|\\ &= \max_{a\in A(s)} (1-\lambda)\sum_{s'\in S} q(s' \mid s,a)|w(s') - u(s')|\\ &\le (1-\lambda)\|w-u\|_\infty. \end{aligned}$$

It follows that $\|T(w) - T(u)\|_\infty \le (1-\lambda)\|w-u\|_\infty$, hence T is contracting. By Theorem 1.26, T has a unique fixed point w. For each $s \in S$, let $a_s \in A(s)$ be an action that maximizes the expression

$$\lambda r(s,a) + (1-\lambda)\sum_{s'\in S} q(s' \mid s,a)w(s).$$

There might be more than one such action. Then,

$$(T(w))(s) = \lambda r(s,a_s) + (1-\lambda)\sum_{s'\in S} q(s' \mid s,a_s)w(s). \tag{1.22}$$

Let x_* be the pure stationary strategy that plays the action a_s at state s, for every $s \in S$. We prove that $w(s) = v_\lambda(s)$ for every $s \in S$, and that x^* is λ-discounted optimal.

Step 2: $\gamma_\lambda(s;x^*) = w(s)$ for every initial state $s \in S$.
This follows from Eq. (1.22) and Lemma 1.28.

Step 3: $\gamma_\lambda(s;\sigma) \leq w(s)$ for every strategy σ and every initial state $s \in S$.
By the definition of $T(w)$,

$$\begin{aligned}(T(w))(s_t) &= \max_{a\in A(s_t)} (\lambda r(s_t,a) + (1-\lambda)w(s_{t+1})))\\ &\geq \mathbf{E}_{s_t,\sigma}\left[\lambda r(s_t,a) + (1-\lambda)w(s_{t+1})\right],\end{aligned}$$

for all $t \in \mathbb{N}$. The claim follows from Lemma 1.28. □

We in fact proved the following characterization of the set of optimal strategies in Markov decision problems, whose proof is left for the reader (Exercise 1.16). In this characterization and later in the book, we will use the following notations:

$$r(s,x(s)) := \sum_{a\in A(s)}\left(\prod_{i\in I} x^i(s,a^i)\right) r(s,a), \quad \forall s \in S, x(s) \in \prod_{i\in I}\Delta(A^i(s)),$$

$$q(s' \mid s,x(s)) := \sum_{a\in A(s)}\left(\prod_{i\in I} x^i(s,a^i)\right) q(s' \mid s,a),$$

$$\forall s,s' \in S, x(s) \in \prod_{i\in I}\Delta(A^i(s)).$$

The quantity $\prod_{i\in I} x^i(s,a^i)$ is the probability that under the mixed action profile $x(s)$, the action profile a is chosen. Therefore, $r(s,x(s))$ is the expected stage payoff at stage s when the players play the stationary strategy profile x, and $q(s' \mid s,x(s))$ is the probability that the play moves from s to s' when the players play the stationary strategy profile x.

Theorem 1.30 *Let $\Gamma = \langle S,(A(s))_{s\in S},q,r\rangle$ be a Markov decision problem, and let $\lambda \in (0,1]$ be a discount factor. A stationary strategy x is λ-discounted*

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

optimal at all initial states if and only if for every state $s \in S$ the mixed action $x(s)$ satisfies

$$v_\lambda(s) = \lambda r(s, x(s)) + (1-\lambda) \sum_{s' \in S} q(s' \mid s, x(s)) v_\lambda(s').$$

1.7 Uniform Optimality

For each $s \in S$ consider the function $\lambda \mapsto v_\lambda(s)$, which assigns to each discount factor its discounted value. How does this function depend on λ? Can it be equal to $\sin(\lambda)$ or e^λ? In this section we will answer this question, among others.

Recall that a function $f\colon \mathbb{R} \to \mathbb{R}$ is *rational* if it is the ratio of two polynomials.

Theorem 1.31 *Two rational functions $f, g\colon \mathbb{R} \to \mathbb{R}$ either coincide, or they (i.e., their graphs) have finitely many intersection points: the set $\{x \in \mathbb{R}\colon f(x) = g(x)\}$ is either $\mathbb{R}$ or finite.*

Proof Let $f = \frac{P_1}{Q_1}$ and $g = \frac{P_2}{Q_2}$, where P_1, Q_1, P_2, and Q_2 are polynomials. Then

$$\begin{aligned}\left\{x \in \mathbb{R}\colon f(x) = g(x)\right\} &= \left\{x \in \mathbb{R}\colon \frac{P_1(x)}{Q_1(x)} = \frac{P_2(x)}{Q_2(x)}\right\} \\ &= \left\{x \in \mathbb{R}\colon P_1(x)Q_2(x) - P_2(x)Q_1(x) = 0\right\}.\end{aligned}$$

That is, $\{x \in \mathbb{R}\colon f(x) = g(x)\}$ is the set of all zeroes of a polynomial. Since a nonzero polynomial has finitely many zeros, the result follows. □

An $n \times n$ matrix $Q = (Q_{ij})_{i,j \in \{1,\dots,n\}}$ is *stochastic* if the sum of entries in every row is 1, that is, $\sum_{j=1}^n Q_{ij} = 1$ for all $i \in \{1, 2, \dots, n\}$. Let Id denote the identity matrix.

Theorem 1.32 *For every stochastic matrix Q and every $\lambda \in (0,1]$, the matrix $Id - (1-\lambda)Q$ is invertible, that is, the inverse matrix $(Id - (1-\lambda)Q)^{-1}$ exists.*

Proof Setting $P := Id - (1-\lambda)Q$ and $R := \sum_{k=0}^{\infty} (1-\lambda)^k Q^k$, we note that $P \cdot R = Id$, and therefore P is invertible.

Alternatively, $P_{ii} > 0$ for every $i \in \{1, 2, \dots, n\}$ and $P_{ij} \leq 0$ for every $i, j \in \{1, 2, \dots, n\}$ such that $i \neq j$, which implies that P is invertible. □

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

Theorem 1.33 *For any fixed pure stationary strategy x and any fixed initial state $s \in S$, the function $\lambda \mapsto \gamma_\lambda(s;x)$ is rational.*

Our proof for Theorem 1.33 is valid for any stationary (and not necessarily pure) strategy (see Exercise 1.12).

Proof Recall that a pure stationary strategy is a vector of actions, one for each state. Fix a pure stationary strategy $x = (a_s)_{s \in S}$. Denote by Q the transition matrix induced by x. This is a matrix with $|S|$ rows and $|S|$ columns, with entries (s,s') given by

$$Q_{s,s'} = q(s' \mid s, a_s).$$

Using the matrix Q, we can easily calculate the distribution of the state s_t at stage t. Suppose that one chooses an initial state according to a probability distribution $p \in \Delta(S)$ (which is expressed as a row vector), and then one plays the action a_s. What is the probability that the next state will be s'? This probability is $\sum_{s \in S} p_s q(s' \mid s, a_s)$, which is the s' coordinate of the vector pQ. Similarly, since $(pQ)_s$ is the probability that the state at stage 2 is s, the probability that the state at stage 3 is s' is given by $\sum_{s \in S} (pQ)_s q(s' \mid s, a_s)$, which is the s' coordinate of the vector pQ^2. By induction, it follows that the probability that the state at stage t is s' is the s' coordinate of the vector pQ^{t-1}.

For a state $s \in S$ denote by $\mathbf{1}(s) = (0, \ldots, 0, 1, 0, \ldots, 0)$ the row vector with the s coordinate equal to 1 and all the other coordinates equal to 0. Then $\mathbf{1}(s)Q^{t-1}$ represents the probability distribution of the state s_t at stage t, given that the initial state is s. Therefore, the λ-discounted payoff can be expressed as

$$\gamma_\lambda(s;x) = \sum_{t=1}^{\infty} \lambda(1-\lambda)^{t-1} \mathbf{1}(s) Q^{t-1} R,$$

where R is the row vector $(r(s,a_s))_{s \in S}$. Therefore,

$$\begin{aligned} \gamma_\lambda(s;x) &= \lambda \mathbf{1}(s) \left(\sum_{t=1}^{\infty} (1-\lambda)^{t-1} Q^{t-1} \right) R \\ &= \lambda \mathbf{1}(s) (I - (1-\lambda)Q)^{-1} R. \end{aligned}$$

By Theorem 1.32 the matrix $I - (1-\lambda)Q$ is invertible, and by Cramer's rule,[10] the inverse matrix $(I - (1-\lambda)Q)^{-1}$ can be represented as the ratio of

[10] Gabriel Cramer (Geneva, Italy, July 31, 1704 – Bagolns-sur-Cèze, France, January 4, 1752) was a mathematician from the Republic of Geneva. In addition to presenting Cramer's rule for the calculation of the inverse of a matrix, Cramer worked on algebraic curves.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

two polynomials in the entries of the matrix $I - (1-\lambda)Q$. We conclude that for every fixed pure stationary strategy x, the function $\lambda \mapsto \gamma_\lambda(s;x)$ is rational. □

We can now prove a general structure theorem regarding the value function.

Corollary 1.34 *For any fixed state $s \in S$, the function $\lambda \mapsto v_\lambda(s)$ is continuous. Moreover, there exist $K \in \mathbb{N}$ and $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_K = 1$ such that for every $k = 0, 1, \ldots, K-1$, the following holds:*

- *The restriction of the function $\lambda \mapsto v_\lambda(s)$ to the interval $(\lambda_k, \lambda_{k+1})$ is rational.*
- *There is a pure stationary strategy $x_k \in \prod_{s \in S} A(s)$ that is λ-discounted optimal for all $\lambda \in (\lambda_k, \lambda_{k+1})$.*

Proof Let Σ_{SP} denote the finite set of all pure stationary strategies. For any fixed pure stationary strategy $x \in \Sigma_{SP}$ and any fixed state $s \in S$, consider the function $\lambda \mapsto \gamma_\lambda(s;x)$, which we denote by $\gamma_\bullet(s;x)$. By Theorem 1.33, $\gamma_\bullet(s;x)$ is a rational function; in particular, $\gamma_\bullet(s;x)$ is continuous. Since there exists a pure stationary optimal strategy, the λ-discounted value at the initial state s is given by

$$v_\lambda(s) = \max_{x \in \Sigma_{SP}} \gamma_\lambda(s;x).$$

Since the function $\lambda \mapsto v_\lambda(s)$ is the maximum of a finite family of rational functions, it is continuous.

By Theorem 1.31, two distinct rational functions intersect in finitely many points. Let Λ_s be the set of all intersection points of the rational functions $(\gamma_\bullet(s;x))_{x \in \Sigma_{SP}}$, and set $\Lambda := \bigcup_{s \in S} \Lambda_s$. Since the set Σ_{SP} is finite, the set Λ_s is finite for every state $s \in S$, and so the set Λ is finite as well. Add the points 0 and 1 to the set Λ, and denote $\Lambda = \{\lambda_0, \lambda_1, \ldots, \lambda_K\}$ where $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_K = 1$.

Fix $k \in \{0, 1, \ldots, K-1\}$. By the choice of λ_k and λ_{k+1}, for every state $s \in S$ the functions $(\gamma_\bullet(s;x))_{x \in \Sigma_{SP}}$ have no common intersection point in the interval $(\lambda_k, \lambda_{k+1})$. Let $x_k \in \Sigma_{SP}$ be a pure stationary strategy that is λ-discounted optimal for some $\lambda \in (\lambda_k, \lambda_{k+1})$. We claim that x_k is λ'-discounted optimal at all initial states, for every $\lambda' \in (\lambda_k, \lambda_{k+1})$, as needed. Indeed, since x_k is λ-discounted optimal at all initial states, for every fixed pure stationary strategy $x \in \Sigma_{SP}$ and every fixed state $s \in S$, either $\gamma_\gamma(s;x_k) > \gamma_\lambda(s;x)$, or $\gamma_\lambda(s;x_k) = \gamma_\lambda(s;x)$. In the former case, since the set of intersection points of the functions $\gamma_\bullet(s;x_k)$ and $\gamma_\bullet(s;x)$ is disjoint from $(\lambda_k, \lambda_{k+1})$, it follows that $\gamma_{\lambda'}(s;x_k) > \gamma_{\lambda'}(s;x)$ for every $\lambda' \in (\lambda_k, \lambda_{k+1})$. In the latter case, for the same reason $\gamma_{\lambda'}(s;x_k) = \gamma_{\lambda'}(s;x)$ for every $\lambda' \in (\lambda_k, \lambda_{k+1})$. Hence, x_k is indeed λ'-discounted optimal for all $\lambda' \in (\lambda_k, \lambda_{k+1})$. □

The significance of Corollary 1.34 is that the decision maker does not need to know precisely the discount factor for her to play optimally. If all the decision maker knows is that the discount factor is within an interval in which a specific pure stationary strategy x is optimal, by following x she ensures that she plays optimally, regardless of the exact value of the discount factor.

In particular, we get the following.

Corollary 1.35 *There is a pure stationary strategy that is optimal for every discount factor sufficiently close to* 0.

In many situations the decision maker is patient, that is, her discount factor is close to 0. For example, countries negotiating a peace treaty are often patient. Another example concerns an investor who may execute many transactions along the day, sometimes even selling a stock that she bought earlier in the day. For such an investor, one period of the game may last one hour or one minute, and subsequently her discount factor is quite close to 0. When the discount factor is close to 0, by Corollary 1.35, to play optimally the decision maker does not need to know the exact value of the discount factor.

Definition 1.36 A strategy σ is *uniformly optimal* at the initial state s if there is a $\lambda_0 > 0$ such that σ is λ-discounted optimal at the initial state s for every $\lambda \in (0, \lambda_0)$.

In the literature, uniformly optimal strategies are also called *Blackwell optimal*. By Corollary 1.35, we deduce the following result.

Theorem 1.37 *In every Markov decision problem, there is a pure stationary strategy that is uniformly optimal at all initial states.*

If $f\colon (0,1] \to \mathbb{R}$ is a bounded rational function, then the limit $\lim_{\lambda \to 0} f(\lambda)$ exists. We therefore deduce that the discounted value is continuous at 0.

Corollary 1.38 $\lim_{\lambda \to 0} v_\lambda(s)$ *exists for every initial state* $s \in S$.

1.8 Comments and Extensions

Markov decision problems were first studied by Blackwell (1962). The model, as introduced by Definition 1.1, include finitely many states, and the set of actions available at each state is finite. Markov decision problems with general state and action sets were considered in the literature, and the existence of T-stage optimal strategies as well as of stationary λ-discounted optimal

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

strategies was established under various topological conditions on the set SA of pairs (state, action) and continuity conditions on the payoff function and on the transition rule. For more details, the reader is referred to Puterman (1994).

By Theorem 1.34, for every state $s \in S$ the function $\lambda \mapsto v_\lambda(s)$ is piecewise rational. A natural goal is to characterize the set of all functions that can arise as the value function of some Markov decision problem. Such a characterization was provided by Lehrer et al. (2016).

Here we considered two types of evaluations for the decision maker: the T-stage evaluation and the discounted evaluations. Other evaluations have also been considered, see Puterman (1994, Section 5.4), where algorithms for approximating optimal strategies for various evaluations are described.

We proved that the limit $\lim_{\lambda \to 0} v_\lambda(s)$ of the discounted value exists for every initial state s. We did not touch upon the convergence of the T-stage value as T goes to infinity, namely, $\lim_{T \to \infty} v_T(s)$. For Markov decision problems with finitely many states and actions, the fact that $\lim_{T \to \infty} v_T(s)$ exists and is equal to $\lim_{\lambda \to 0} v_\lambda(s)$ follows from a result of Hardy and Littlewood, see Korevaar (2004, chapter I.7). We will not prove this result directly, as it will follow from a much more general result that we will obtain later in this book (see Theorem 9.13 on Page 139). A rich literature extends this result to Markov decision problems with general state and action sets, see, for example, Lehrer and Sorin (1992), Monderer and Sorin (1993), and Lehrer and Monderer (1994).

When the decision maker follows a uniformly optimal strategy, she guarantees that the discounted payoff is close to the value. This does not rule out the possibility that the payoff fluctuates along the play: during some long blocks of stages, the payoff is high; in other long blocks of stages, the payoff is low; and the blocks are arranged in such a way that the average payoff is close to the value. Sorin et al. (2010) proved that this is not the case: if the decision maker follows a uniformly optimal strategy, then for every sufficiently large positive integer m there is a $T \in \mathbb{N}$ such that for every $t \geq T$, the expected average payoff in stages $t, t+1, \ldots, t+T-1$ is close to $\lim_{\lambda \to 0} v_\lambda(s_1)$.

1.9 Exercises

Exercise 1.3 is used in the solution Exercise 5.1.

1. Prove that the strategy σ^* that is described after the proof of Theorem 1.17 is T-stage optimal.
2. In this exercise, we bound the variation of the sequence of the T-stage values $(v_T(s))_{T \in \mathbb{N}}$. Prove that for every $T, k \in \mathbb{N}$ and every state $s \in S$,

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

$$|(T+k)v_{T+k}(s) - Tv_T(s)| \leq k\|r\|_\infty.$$

3. Let Γ be a Markov decision problem. Prove that there is a pure T-stage optimal strategy with the following property: the action played in each stage depends only on the current stage and on the number of stages left. That is, for every $t \in \{1, 2, \ldots, T\}$ and every two histories $h_t = (s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ and $h'_t = (s'_1, a'_1, \ldots, s'_{t-1}, a'_{t-1}, s'_t)$, if $s_t = s'_t$, then $\sigma(h_t) = \sigma(h'_t)$.
4. For $\lambda \in (0, 1]$, calculate the λ-discounted value and the λ-discounted optimal strategy at the initial state $s(1)$ in Example 1.3.
5. Prove Theorem 1.22 about the dynamic programming principle for discounted decision problems: For every initial state $s \in S$ and every discount factor $\lambda \in (0, 1]$,

$$v_\lambda(s) = \max_{a \in A(s)} \left\{ \lambda r(s,a) + (1-\lambda) \sum_{s' \in S} q(s' \mid s, a) v_\lambda(s') \right\}.$$

6. Find the discounted payoff of each pure stationary strategy in the following Markov decision problem and determine the discounted value for every discount factor.

$$\begin{array}{c|c|} \cline{2-2} U & 1_{\left(\frac{2}{3}, \frac{1}{3}\right)} \\ \cline{2-2} D & 0_{(0,1)} \\ \cline{2-2} \multicolumn{1}{c}{} & \multicolumn{1}{c}{s(1)} \end{array} \qquad \begin{array}{c|c|} \cline{2-2} U & 2_{\left(\frac{1}{2}, \frac{1}{2}\right)} \\ \cline{2-2} D & 3_{(1,0)} \\ \cline{2-2} \multicolumn{1}{c}{} & \multicolumn{1}{c}{s(2)} \end{array}$$

7. Let σ_1 and σ_2 be two pure stationary strategies. Let σ_3 be a stationary strategy that at every state s chooses an action a that maximizes

$$\lambda r(s,a) + (1-\lambda) \sum_{s' \in S} q(s' \mid s, a) \max\{\gamma_\lambda(s'; \sigma_1), \gamma_\lambda(s'; \sigma_2)\}.$$

Prove that

$$\gamma_\lambda(s; \sigma_3) \geq \max\{\gamma_\lambda(s; \sigma_1), \gamma_\lambda(s; \sigma_2)\}, \quad \forall s \in S.$$

8. Let Γ be a Markov decision problem and let s be a state. In view of Comment 1.20, is it true that for $\lambda = \frac{1}{T}$ we have $v_\lambda(s) = v_T(s)$? If so, prove it. If not, explain why an inequality does not necessarily hold.
9. Show that every contracting mapping is continuous.
10. Show that for every polynomial P there exist a Markov decision problem and an initial state s such that $v_\lambda(s) = P(\lambda)$ for all $\lambda \in (0, 1]$.

11. Let σ be a strategy in a Markov decision problem Γ, and let $\lambda \in (0,1)$. Prove that σ is λ-discounted optimal at the initial state s if and only if the following condition holds: For every history $h_t \in H$ that satisfies $\mathbf{P}_{s,\sigma}(h_t) > 0$ and every action $a' \in A(s_t)$ that satisfies $\sigma(a' \mid h_t) > 0$,

$$a' \in \operatorname{argmax}_{a \in A(s_t)} \left\{ \lambda r(s_t, a) + (1-\lambda) \sum_{s' \in S} q(s' \mid s_t, a) v_\lambda(s') \right\}.$$

12. Prove that for each fixed stationary (not necessarily pure) strategy x, the function $\lambda \mapsto \gamma_\lambda(s; x)$ is rational.
13. Find a Markov decision problem that satisfies the following two properties:

- There is a strategy σ that is λ-discounted optimal for $\lambda = 1$ and for every λ sufficiently close to 0, but is not optimal for $\lambda = \frac{1}{2}$.
- There is a strategy σ' that is not λ-discounted optimal for $\lambda = 1$ and for every λ sufficiently close to 0, but is optimal for $\lambda = \frac{1}{2}$.

14. Let $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ be two closed sets. A *correspondence* $F\colon X \rightrightarrows Y$ is a mapping that assigns to each point $x \in X$ a subset $F(x) \subseteq Y$. We say that the correspondence F has non-empty values if $F(x) \neq \emptyset$ for every $x \in X$. The graph of a correspondence F is $\text{Graph}(F) = \{(x,y) \in \mathbb{R}^{n+m} \colon y \in F(x)\}$.

    Let $X \subseteq \mathbb{R}^n$ be a compact set. Let $F\colon X \times X \rightrightarrows \mathbb{R}$ and $G\colon X \rightrightarrows X$ be two correspondences with non-empty values and compact graphs and let $\lambda \in (0,1)$. Prove that there exists a unique function $f\colon X \rightrightarrows \mathbb{R}$ such that

$$f(x) = \max_{y \in G(x)} (F(x,y) + \lambda f(y)).$$

15. Let $\Gamma = \langle S, (A(s))_{s \in S}, q, r \rangle$ be a Markov decision problem, and consider the following linear program in the variables $(v(s))_{s \in S}$:

$$\text{Minimize} \quad \sum_{s \in S} v(s)$$

$$\text{Subject to} \quad v(s) \geq \lambda r(s,a) + (1-\lambda) \sum_{s' \in S} q(s' \mid s, a) v(s'), \ \forall s \in S, a \in A.$$

    Show that the solution $(v(s))_{s \in S}$ of this linear program has the property that $v(s)$ is the λ-discounted value at the initial state s.

16. Prove Theorem 1.30: Let $\Gamma = \langle S, (A(s))_{s \in S}, q, r \rangle$ be a Markov decision problem, let $\lambda \in (0,1]$ be a discount factor, and let $v_\lambda(s)$ be the λ-discounted value at the initial state s, for every $s \in S$. A stationary

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press

strategy x is λ-discounted optimal at all initial states if and only if, for every state $s \in S$, the mixed action $x(s)$ satisfies

$$v_\lambda(s) = \lambda r(s, x(s)) + (1 - \lambda) \sum_{s' \in S} q(s' \mid s, x(s)) v_\lambda(s').$$

17. Let $\Gamma = \langle S, (A(s))_{s \in S}, q, r \rangle$ be a Markov decision problem where S is countable, $A(s)$ is finite for every $s \in S$, and r is bounded. Prove that for every $\lambda \in (0, 1]$ the λ-discounted value exists at all initial states, and moreover the decision maker has a pure stationary λ-discounted optimal strategy.
18. Does $\lim_{\lambda \to 0} v_\lambda(s)$ exist in every Markov decision problem $\Gamma = \langle S, (A(s))_{s \in S}, q, r \rangle$ for every $s \in S$, where S is countable, $A(s)$ is finite for every $s \in S$, and r is bounded? Prove or provide a counterexample.

https://doi.org/10.1017/9781009029704.002 Published online by Cambridge University Press