A number of individuals have read various versions of the manuscript, full or in parts and helped me to reduce the number of mistakes by sending corrections. They include Dimitri Bertsekas, Gábor Balázs, Bernardo Avila Pires, Warren Powell, Rich Sutton, Nikos Vlassis, Hengshuai Yao and Shimon Whiteson. Thank You! Of course, all the remaining mistakes are mine. If I have left out someone from the above list, this was by no means intentional. If this is the case, please remind me in an e-mail (better yet, send me some comments or suggestions). Independently of whether they have contacted me before or not, readers are encouraged to e-mail me if they find errors, typos or they just think that some topic should have been included (or left out). I plan to periodically update the text and I will try to accommodate all the requests. Finally, I wish to thank Remi Munos, and Rich Sutton, my closest collaborators over the last few years, from whom I have learned and continue to learn a lot. I also wish to thank all my students, the members of RLAI group and all researchers of RL who continue to strive to push the boundaries of what we can do with reinforcement learning. This book is made possible by you.

A The theory of discounted Markovian decision processes

The purpose of this section is to give a short proof of the basic results of the theory of Markovian decision processes. All the results will be worked out for the discounted expected total cost criterion. First, we give a short overview of contraction mappings and Banach's fixed-point theorem. Next, we show how this powerful result can be applied to proof a number of basic results about value functions and optimal policies.

A.1 Contractions and Banach's fixed-point theorem

We start with some basic definitions which we will need in the rest of this section.

Definition 1 (Norm)**.** Let V be a vector space over the reals. Then $f : V \to \mathbb{R}_0^+$ is a norm on V provided that the following hold:

1. If $f(v) = 0$ for some $v \in V$ then $v = 0$;
2. For any $\lambda \in \mathbb{R}$, $v \in V$, $f(\lambda v) = |\lambda| f(v)$;
3. For any $v, u \in V$, $f(v + u) \leq f(v) + f(u)$.

A vector space together with a norm is called a *normed vector space*.

According to the definition a norm is a function that assigns a nonnegative number to each vector. This number is often called the "length" or just the "norm" of the vector. The norm of a vector v is often denoted by $\|v\|$.

Example 3: *Here are a few examples of norms over the vector space* $V = (\mathbb{R}^d, +, \lambda\cdot)$.

1. *ℓ^p norms: For* $p \geq 1$,
$$\|v\|_p = \left(\sum_{i=1}^{d} |v_i|^p \right)^{1/p}.$$

2. *ℓ^∞ norm:*
$$\|v\|_\infty = \max_{1 \leq i \leq d} |v_i|.$$

3. *The weighted variants of these norms are defined as follows:*
$$\|v\|_p = \begin{cases} \left(\sum_{i=1}^{d} \frac{|v_i|^p}{w_i} \right)^{1/p}, & \text{if } 1 \leq p < \infty; \\ \max_{1 \leq i \leq d} \frac{|v_i|}{w_i}, & \text{if } p = \infty, \end{cases}$$
where $w_i > 0$.

4. *The matrix-weighted* 2*-norm is defined as follows:*
$$\|v\|_P^2 = v^T P v.$$
Here P *is a fixed, positive definite matrix.*

Similarly, one can define norms over spaces of functions. For example, if V *is the vector space of functions over the domain* $\mathcal{X}$ *which are bounded then*
$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|.$$

(A function is called bounded exactly when $\|f\|_\infty < +\infty$*.)*

We will be interested in the convergence of sequences in normed vector spaces.

Definition 2 (Convergence in norm)**.** Let $V = (V, \|\cdot\|)$ be a normed vector space. Let $v_n \in V$ be a sequence of vectors ($n \in \mathbb{N}$). The sequence $(v_n; n \geq 0)$ is said to converge to the vector v in the norm $\|\cdot\|$ if $\lim_{n\to\infty} \|v_n - v\| = 0$. This will be denoted by $v_n \to_{\|\cdot\|} v$.

Note that in a d-dimensional vector space $v_n \to_{\|\cdot\|} v$ is the same as requiring that for each $1 \leq i \leq d$, $v_{n,i} \to v_i$ (here $v_{n,i}$ denotes the i^{th} component of v_n). However, this does not hold for infinite dimensional vector spaces. Take for example $\mathcal{X} = [0, 1]$ and the space of bounded

functions over $\mathcal{X}$. Let

$$f_n(x) = \begin{cases} 1, & \text{if } x < 1/n; \\ 0, & \text{otherwise .} \end{cases}$$

Define f so that $f(x) = 0$ if $x \neq 0$ and $f(0) = 1$. Then $f_n(x) \to f(x)$ for each x (i.e., f_n converges to $f(x)$ *pointwise*). However, $\|f_n - f\|_\infty = 1 \not\to 0$.

If we have a sequence of real-numbers $(a_n; n \geq 0)$, we can test if the sequence converges *without the knowledge of the limiting value* by verifying if it is a *Cauchy sequence*, i.e., whether $\lim_{n\to\infty} \sup_{m\geq n} |a_n - a_m| = 0$. ('Sequences with vanishing oscillations' is possibly a more descriptive name for Cauchy sequences.) It is a quite notable property of the real numbers that every Cauchy sequence of reals assumes a limit.

The extension of the concept of Cauchy sequences to normed vector spaces is straightforward:

Definition 3 (Cauchy sequence)**.** Let $(v_n; n \geq 0)$ be a sequence of vectors of a normed vector-space $V = (V, \|\cdot\|)$. Then v_n is called a Cauchy-sequence if $\lim_{n\to\infty} \sup_{m\geq n} \|v_n - v_m\| = 0$.

Normed vector spaces where all Cauchy sequences are convergent are special: one can find examples of normed vector spaces such that some of the Cauchy sequences in the vector space do not have a limit.

Definition 4 (Completeness)**.** A normed vector space V is called *complete* if every Cauchy sequence in V is convergent in the norm of the vector space.

To pay tribute to Banach, the great Polish mathematician of the first half of the 20th century, we have the following definition:

Definition 5 (Banach space)**.** A complete, normed vector space is called a *Banach space.*

One powerful result in the theory of Banach spaces concerns contraction mappings, or contraction operators. These are special Lipschitzian mappings:

Definition 6. Let $V = (V, \|\cdot\|)$ be a normed vector space. A mapping $T : V \to V$ is called *L-Lipschitz* if for any $u, v \in V$,

$$\|Tu - Tv\| \leq L\|u - v\|.$$

A mapping T is called a *non-expansion* if it is Lipschitzian with $L \leq 1$. It is called a *contraction* if it is Lipschitzian with $L < 1$. In this case, L is called the contraction factor of T and T is called an L-contraction.

Note that if T is Lipschitz, it is also continuous in the sense that if $v_n \to_{\|\cdot\|} v$ then also $Tv_n \to_{\|\cdot\|} Tv$. This is because $\|Tv_n - Tv\| \leq L\|v_n - v\| \to 0$ as $n \to \infty$.

Definition 7 (Fixed point)**.** Let $T : V \to V$ be some mapping. The vector $v \in V$ is called a *fixed point* of T if $Tv = v$.

Theorem 1 (Banach's fixed-point theorem)**.** *Let V be a Banach space and $T : V \to V$ be a contraction mapping. Then T has a unique fixed point. Further, for any $v_0 \in V$, if $v_{n+1} = Tv_n$ then $v_n \to_{\|\cdot\|} v$, where v is the unique fixed point of T and the convergence is geometric:*

$$\|v_n - v\| \leq \gamma^n \|v_0 - v\|.$$

Proof. Pick any $v_0 \in V$ and define v_n as in the statement of the theorem. We first demonstrate that (v_n) converges to some vector. Then we will show that this vector is a fixed point of T. Finally, we show that T has a single fixed point.
Assume that T is a γ-contraction. To show that (v_n) converges it suffices to show that (v_n) is a Cauchy sequence (since V is a Banach, i.e., complete normed vector-space). We have

$$\begin{aligned}
\|v_{n+k} - v_n\| &= \|Tv_{n-1+k} - Tv_{n-1}\| \\
&\leq \gamma \|v_{n-1+k} - v_{n-1}\| = \gamma \|Tv_{n-2+k} - Tv_{n-2}\| \\
&\leq \gamma^2 \|v_{n-2+k} - v_{n-2}\| \\
&\vdots \\
&\leq \gamma^n \|v_k - v_0\| .
\end{aligned}$$

Now,

$$\|v_k - v_0\| \leq \|v_k - v_{k-1}\| + \|v_{k-1} - v_{k-2}\| + \ldots + \|v_1 - v_0\|$$

and, by the same logic as used before, $\|v_i - v_{i-1}\| \leq \gamma^{i-1} \|v_1 - v_0\|$. Hence,

$$\|v_k - v_0\| \leq \left(\gamma^{k-1} + \gamma^{k-2} + \ldots + 1\right) \|v_1 - v_0\| \leq \frac{1}{1-\gamma} \|v_1 - v_0\|.$$

Thus,

$$\|v_{n+k} - v_n\| \leq \gamma^n \left(\frac{1}{1-\gamma} \|v_1 - v_0\| \right),$$

and so

$$\lim_{n \to \infty} \sup_{k \geq 0} \|v_{n+k} - v_n\| = 0,$$

showing that $(v_n; n \geq 0)$ is indeed a Cauchy sequence. Let v be its limit.
Now, let us go back to the definition of the sequence $(v_n; n \geq 0)$:

$$v_{n+1} = Tv_n.$$

Taking the limes of both sides, on the one hand, we get that $v_{n+1} \to_{\|\cdot\|} v$. On the other

hand, $Tv_n \to_{\|\cdot\|} Tv$, since T is a contraction, hence it is continuous. Thus, the left-hand side converges to v, while the right-hand side converges to Tv, while the left and right-hand sides are equal. Therefore, we must have $v = Tv$, showing that v is a fixed point of T.
Let us consider the problem of uniqueness of the fixed point of T. Let us assume that v, v' are both fixed points of T. Then, $\|v - v'\| = \|Tv - Tv'\| \le \gamma\|v - v'\|$, or $(1-\gamma)\|v - v'\| \le 0$. Since a norm takes only nonnegative values and $\gamma < 1$, we get that $\|v - v'\| = 0$. Thus, $v - v' = 0$, or $v = v'$, finishing the proof of the first part of the statement.
For the second part, we have

$$\begin{aligned}
\|v_n - v\| &= \|Tv_{n-1} - Tv\| \\
&\le \gamma\|v_{n-1} - v\| = \gamma\|Tv_{n-2} - Tv\| \\
&\le \gamma^2\|v_{n-2} - v\| \\
&\vdots \\
&\le \gamma^n\|v_0 - v\|.
\end{aligned}$$

□

A.2 Application to MDPs

For the purpose of this section, we define V^* by

$$V^*(x) = \sup_{\pi \in \Pi_{\text{stat}}} V^\pi(x), \quad x \in \mathcal{X}.$$

Thus, $V^*(x)$ is an upper bound on the value that we can achieve by choosing some stationary policy π. Note that if the supremum was taken over the larger class of all policies, we could possibly get a larger function. However, in the case of MDPs considered in this section, these two optimal value functions are actually the same. Although, this is not hard to prove, we omit the proof.
Let $B(\mathcal{X})$ be the space of bounded functions with domain $\mathcal{X}$:

$$B(\mathcal{X}) = \{\, V : \mathcal{X} \to \mathbb{R} \,:\, \|V\|_\infty < +\infty \,\}.$$

In what follows, we will view $B(\mathcal{X})$ as a normed-vector space with the norm $\|\cdot\|_\infty$. It is easy to show that $(B(\mathcal{X}), \|\cdot\|_\infty)$ is complete: If $(V_n; n \ge 0)$ is a Cauchy sequence in it then for any $x \in \mathcal{X}$, $(V_n(x); n \ge 0)$ is also a Cauchy sequence over the reals. Denoting by $V(x)$ the limit of $(V_n(x))$, one can show that $\|V_n - V\|_\infty \to 0$. Vaguely speaking, this holds because $(V_n; n \ge 0)$ is a Cauchy sequence in the norm $\|\cdot\|_\infty$ so the rate of convergence of $V_n(x)$ to

$V(x)$ is independent of x.
Pick any stationary policy π. Remember that the Bellman operator underlying π, $T^\pi : B(\mathcal{X}) \to B(\mathcal{X})$, is defined by

$$(T^\pi V)(x) = r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{X}} \mathcal{P}(x, \pi(x), y) V(y), \quad x \in \mathcal{X}.$$

Note that T^π is well-defined: If $U \in B(\mathcal{X})$, then $T^\pi U \in B(\mathcal{X})$ holds true.
It is easy to see that V^π as defined by (7) is a fixed point to T^π:

$$\begin{aligned} V^\pi(x) &= \mathbb{E}\left[R_1 | X_0 = x\right] + \gamma \sum_{y \in \mathcal{X}} \mathcal{P}(x, \pi(x), y) \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+2} | X_1 = y\right] \\ &= T^\pi V^\pi(x). \end{aligned}$$

It is also easy to see that T^π is a contraction in $\|\cdot\|_\infty$:

$$\begin{aligned} \|T^\pi U - T^\pi V\|_\infty &= \gamma \sup_{x \in \mathcal{X}} \left| \sum_{y \in \mathcal{X}} \mathcal{P}(x, \pi(x), y)(U(y) - V(y)) \right| \\ &\le \gamma \sup_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \mathcal{P}(x, \pi(x), y) \left|U(y) - V(y)\right| \\ &\le \gamma \sup_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \mathcal{P}(x, \pi(x), y) \left\|U - V\right\|_\infty \\ &= \gamma \left\|U - V\right\|_\infty, \end{aligned}$$

where the last line follows from $\sum_{y \in \mathcal{X}} \mathcal{P}(x, \pi(x), y) = 1$.
It follows that in order to find V^π, one can construct the sequence V_0, $T^\pi V_0$, $(T^\pi)^2 V_0$, ..., which, by Banach's fixed-point theorem will converge to V^π at a geometric rate.
Now, recall the definition of the Bellman optimality operator: $T^* : B(\mathcal{X}) \to B(\mathcal{X})$,

$$(T^* V)(x) = \sup_{a \in \mathcal{A}} \left\{ r(x, a) + \gamma \sum_{y \in \mathcal{X}} \mathcal{P}(x, a, y) V(y) \right\}, \quad x \in \mathcal{X}. \tag{42}$$

Again, T^* is well-defined. We now show that T^* is also a γ-contraction with respect to the supremum norm $\|\cdot\|_\infty$.
To see this first note that

$$\left|\sup_{a \in \mathcal{A}} f(a) - \sup_{a \in \mathcal{A}} g(a)\right| \le \sup_{a \in \mathcal{A}} \left|f(a) - g(a)\right|,$$

which can be seen using an elementary case analysis. Using this inequality and then pro-

ceeding as with the analysis of T^π we get,

$$\begin{aligned}\|T^*U - T^*V\|_\infty &\le \gamma \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \sum_{y\in\mathcal{X}} \mathcal{P}(x,a,y)\,|U(y)-V(y)| \\ &\le \gamma \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \sum_{y\in\mathcal{X}} \mathcal{P}(x,a,y)\,\|U-V\|_\infty \\ &= \gamma\,\|U-V\|_\infty,\end{aligned}$$

thus proving the statement. Here, the last equality follows by $\sum_{y\in\mathcal{X}} \mathcal{P}(x,a,y) = 1$.
The main result of this section is the following theorem:

Theorem 2. *Let V be the fixed point of T^* and assume that there is policy π which is greedy w.r.t V: $T^\pi V = T^* V$. Then $V = V^*$ and π is an optimal policy.*

Proof. Pick any stationary policy π. Then $T^\pi \le T^*$ in the sense that for any function $V \in B(\mathcal{X})$, $T^\pi V \le T^* V$ holds ($U \le V$ means that $U(x) \le V(x)$ holds for any $x \in \mathcal{X}$). Thus, $V^\pi = T^\pi V^\pi \le T^* V^\pi$, i.e., $V^\pi \le T^* V^\pi$. Since $T^* U \le T^* V$ follows from $U \le V$, we also have $T^* V^\pi \le (T^*)^2 V^\pi$. Chaining the inequalities, we get $V^\pi \le (T^*)^2 V^\pi$. Continuing this way, we get for all $n \ge 0$ that $V^\pi \le (T^*)^n V^\pi$. Since T^* is a contraction, the right-hand side converges to V, the unique fixed point of T^* (at this stage we cannot know if $V = V^*$ or not). Thus, $V^\pi \le V$. Since π was arbitrary, we get that $V^* \le V$.
Pick now a policy π such that $T^\pi V = T^* V$. Since V is the fixed-point of T^*, we have $T^\pi V = V$. Since T^π has a unique fixed point, V^π, we have $V^\pi = V$, showing that $V^* = V$ and that π is an optimal policy. □

In the statement of the theorem, we were careful in assuming that a greedy policy w.r.t. V exists. Note that this always holds for finite action spaces, and it will hold for infinite action spaces under some extra (continuity) assumptions.
The following theorem serves as the basis of the policy iteration algorithm:

Theorem 3 (Policy improvement theorem)**.** *Choose some stationary policy π_0 and let π be greedy w.r.t. V^{π_0}: $T^\pi V^{\pi_0} = T^* V^{\pi_0}$. Then $V^\pi \ge V^{\pi_0}$, i.e., π is an improvement upon π_0. In particular, if $T^* V^{\pi_0}(x) > V^{\pi_0}(x)$ for some state x then π strictly improves upon π_0 at x: $V^\pi(x) > V^{\pi_0}(x)$. On the other hand, when $T^* V^{\pi_0} = V^{\pi_0}$ then π_0 is an optimal policy.*

Proof. We have $T^\pi V^{\pi_0} = T^* V^{\pi_0} \ge T^{\pi_0} V^{\pi_0} = V^{\pi_0}$. Applying T^π to both sides, we get $(T^\pi)^2 V^{\pi_0} \ge T^\pi V^{\pi_0} \ge V^{\pi_0}$. Continuing this way, we get that for any $n \ge 0$, $(T^\pi)^n V^{\pi_0} \ge V^{\pi_0}$. Taking the limit of both sides, we get that $V^\pi \ge V^{\pi_0}$.
For the second part, notice that we have $(T^\pi)^n V^{\pi_0}(x) \ge T^* V^{\pi_0}(x) > V^{\pi_0}(x)$. Hence, taking the limit, we have $V^\pi(x) \ge T^* V^{\pi_0}(x) > V^{\pi_0}(x)$.