Markov Decision Processes: Optimality

Csaba Szepesvári

Wednesday 8th February, 2023

1 Definitions

Fix a finite MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ and consider the continuing case with the discounted total expected reward criterion with discount factor $0 \le \gamma < 1$. Let Π be the set of all policies of M (we allow history dependent policies, as well). We will use ML to denote the set of memoryless policies of M.

Definition 1.1 (Optimality)**.** A policy $\pi \in \Pi$ of M is said to be optimal in M, if for any other policy $\pi' \in \Pi$ of M, it holds that

$$v_\pi(s) \ge v_{\pi'}(s), \qquad \text{for all } s \in \mathcal{S}\,. \tag{1}$$

A shorthand notation for Equation (1) is[1]

$$v_\pi \ge v_{\pi'}\,.$$

Note that here π and π' are arbitrary policies; they could also be history dependent.

Whether an optimal policy even exists in an MDP is not obvious at this stage: The requirement is that a single policy should be at least as good as any other policy at any state. The answer to this question will be positive, but this comes later. First, we reformulate optimality with the help of optimal value functions:

Definition 1.2 (The optimal value function)**.** The optimal value function $v^* : \mathcal{S} \to \mathbb{R}$, of M is defined via[2]

$$v^*(s) = \sup_{\pi \in \Pi} v_\pi(s)\,, \qquad s \in \mathcal{S}\,.$$

It follows that for any policy $\pi \in \Pi$,

$$v_\pi \le v^*\,. \tag{2}$$

Note that v^* takes on finite values. In particular, for any $s \in \mathcal{S}$,

$$|v^*(s)| \le \frac{r_{\max}}{1-\gamma}\,,$$

where $r_{\max} = \max\{|r'| \,:\, r' \in \mathcal{R}\}$.

The following is another immediate corollary to the above definitions.

Corollary 1.3. *The following are equivalent:*

1. *Policy* $\pi \in \Pi$ *is optimal in* M*;*
2. $v_\pi \geq v^*$*;*
3. $v_\pi = v^*$.

Proof. We show that (1) implies (2), which implies (3), which implies (1).
(1) ⇒ (2): Assume that π is optimal in M. Then, for any $\pi' \in \Pi$ and $s \in \mathcal{S}$

$$v_\pi(s) \geq v_{\pi'}(s)\,.$$

Taking the supremum of both sides with respect to $\pi' \in \Pi$, we get

$$v_\pi(s) \geq \sup_{\pi' \in \Pi} v_{\pi'}(s) = v^*(s)\,.$$

Since $s \in \mathcal{S}$ arbitrary, it follows that $v_\pi \geq v^*$ holds.
(2) ⇒ (3): Assume that $\pi \in \Pi$ is such that $v_\pi \geq v^*$. We also have $v_\pi \leq v^*$. Putting these together, $v^* \geq v_\pi \geq v^*$, hence all of them are equal.
(3) ⇒ (1): Assume that $\pi \in \Pi$ is such that $v_\pi = v^*$. Let $\pi' \in \Pi$ any policy, $s \in \mathcal{S}$ any state. We know that $v^*(s) \geq v_{\pi'}(s)$. Hence, $v_\pi(s) = v^*(s) \geq v_{\pi'}(s)$. Since π' and $s \in \mathcal{S}$ were arbitrary, π is optimal. □

Definition 1.4 (ε-optimality)**.** Let $\varepsilon > 0$. A policy $\pi \in \Pi$ is $\underline{\varepsilon\text{-optimal at state } s \in \mathcal{S}}$, if

$$v_\pi(s) \geq v^*(s) - \varepsilon\,.$$

A policy $\pi \in \Pi$ is $\underline{\varepsilon\text{-optimal}}$, if it is simultaneously ε-optimal at every state, i.e.

$$v_\pi(s) \geq v^*(s) - \varepsilon\,, \qquad s \in \mathcal{S}\,.$$

While it is clear that for any state $s \in \mathcal{S}$, there exists a policy $\pi \in \Pi$ that is ε-optimal at state $s \in \mathcal{S}$ (this follows from the definition of v^*), it is not obvious whether there exists any policy that is ε-optimal simultaneously at all the states. Equivalently, it is not obvious whether the set of ε-optimal policies is empty or not.

2 The optimality equation

Recall that $\mathcal{H}_0 = \mathcal{S}$ and for $t \geq 1$, $\mathcal{H}_t = \mathcal{H}_{t-1} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S}$. We first need a claim, which extends the Bellman equation for general policies:

Proposition 2.1. *Let* $\pi \in \Pi$*, and in particular,* $\pi = (\pi_t)_{t \geq 0}$*,* $\pi_t : \mathcal{H}_t \to \mathcal{M}_1(\mathcal{A})$*. Then,*

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi_0(a|s) \left\{ r(s,a) + \gamma \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} p(r', s'|s,a) v_{\pi^{(s,a,r')}}(s') \right\}, \qquad s \in \mathcal{S}\,, \tag{3}$$

where for any $(s,a,r') \in \mathcal{S} \times \mathcal{A} \times \mathcal{R}$*, we define* $\pi^{(s,a,r')} = (\pi_t^{(s,a,r')})_{t \geq 0}$ *as follows:*

$$\pi_0^{(s,a,r')}(s_1) = \pi_1(s,a,r',s_1)\,, \qquad s_1 \in \mathcal{S}\,,$$

and for arbitrary $t \geq 1$ *and* $h_t = (s_1, a_1, r_2, s_2, a_2, \ldots, a_t, r_{t+1}, s_{t+1}) \in \mathcal{H}_t$*,*

$$\pi_t^{(s,a,r')}(h_t) = \pi_{t+1}(s,a,r',s_1,a_1,r_2,s_2,a_2,\ldots,a_t,r_{t+1},s_{t+1})\,.$$

Proof. The proof works essentially the same way as the proof of the Bellman equation worked for memoryless policies and, as such, is left as an exercise. □

The meaning of the result is as follows: $\pi^{(s,a,r')}$ is the policy that will be followed after π is followed through one transition from state s, where the action chosen by π happens to be a, and the reward that is incurred through the transition is r'. Then, Equation (3) just states that the value of π at state s is the sum of the expected immediate reward that is incurred while following π from s and the expected total discounted value of policy $\pi^{(s,A_0,R_1)}$, where A_0 represents the first (random) action taken and R_1 represents the first (random) reward.

Theorem 2.2. *We have*

$$v^*(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a) v^*(s') \right\}, \qquad s \in \mathcal{S}\,. \tag{4}$$

We call Equation (4) the *Bellman optimality equation*.

Proof. We first prove that $v^*(s)$ is less than or equal to the right-hand side of Equation (4), and then we prove that it is at least as large as this right-hand side. From these two, the equality then follows.

(Part 1:) Fix an arbitrary policy $\pi \in \Pi$ and an arbitrary state $s \in \mathcal{S}$. By Proposition 2.1 and Equation (2),

$$\begin{aligned} v_\pi(s) &= \sum_{a\in\mathcal{A}} \pi_0(a|s) \left\{ r(s,a) + \gamma \sum_{r'\in\mathcal{R},s'\in\mathcal{S}} p(r',s'|s,a) v_{\pi^{(s,a,r')}}(s') \right\} \\ &\le \sum_{a\in\mathcal{A}} \pi_0(a|s) \left\{ r(s,a) + \gamma \sum_{r'\in\mathcal{R},s'\in\mathcal{S}} p(r',s'|s,a) v^*(s') \right\}. \end{aligned}$$

Using the fact that π was arbitrary, we can take the supremum of both the sides, and the inequality will still hold. (Note that for the right hand side, the supremum over $\pi = (\pi_t)_{t\ge 0}$ is equivalent to taking the supremum over π_0.) This means that

$$\begin{aligned} v^*(s) = \sup_\pi v_\pi(s) &\le \sup_{\pi_0} \sum_{a\in\mathcal{A}} \pi_0(a|s) \left\{ r(s,a) + \gamma \sum_{r'\in\mathcal{R},s'\in\mathcal{S}} p(r',s'|s,a) v^*(s') \right\} \\ &= \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{r'\in\mathcal{R},s'\in\mathcal{S}} p(r',s'|s,a) v^*(s') \right\}. \end{aligned}$$

Since $s \in \mathcal{S}$ was arbitrary, we are done with the first part of the proof.

(Part 2:) Let us now show that the inequality also holds in the reverse direction. First define, for each state $s \in \mathcal{S}$, $\pi^{(s)} = (\pi_t^{(s)})_{t\ge 0}$ to be a policy that is ε-optimal at state s. That is,

$$v_{\pi^{(s)}}(s) \ge v^*(s) - \varepsilon\,. \tag{5}$$

Now, consider the policy $\pi = (\pi_t)_{t\ge 0}$, that in time step $t = 0$ and when the state is $s \in \mathcal{S}$, takes an action that maximizes the right-hand side of Equation (4), and then when it arrives at state $s' \in \mathcal{S}$ (regardless of the reward incurred), it follows policy $\pi^{(s')}$ for the remaining time steps:

$$\begin{aligned} \pi_0(s) &\in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s') \right\}, && s \in \mathcal{S}, \\ \pi_t(s_0, a_0, r_1, s_1, a_1, \ldots, r_t, s_t) &= \pi_{t-1}^{(s_1)}(s_1, a_1, \ldots, r_t, s_t), && t \geq 1. \end{aligned}$$

Note that for an arbitrary $r' \in \mathcal{R}$, $s \in \mathcal{S}$, $t \geq 0$, and $h_t = (s_0, a_0, r_1, s_1, \ldots, r_t, s_t) \in \mathcal{H}_t$,

$$\pi_t^{(s,\pi_0(s),r')}(h_t) = \pi_t^{(s_0)}(h_t).$$

(The above equation directly follows from the definitions of the policies $\pi^{(s,\pi_0(s),r')}$ and π.) From the above equality, it follows that for an arbitrary state $s' \in \mathcal{S}$, $\mathbb{P}_{\delta_{s'},\pi^{(s,\pi_0(s),r')}} = \mathbb{P}_{\delta_{s'},\pi^{(s')}}$ (why?) and hence

$$v_{\pi_t^{(s,\pi_0(s),r')}}(s') = v_{\pi^{(s')}}(s').$$

Combining this with Proposition 2.1, we obtain

$$\begin{aligned} v_\pi(s) &= r(s,\pi_0(s)) + \gamma \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} p(r', s'|s, \pi_0(s)) v_{\pi^{(s')}}(s') \\ &\geq r(s,\pi_0(s)) + \gamma \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} p(r', s'|s, \pi_0(s)) (v^*(s') - \varepsilon) && \text{(by Equation (5))} \\ &= \left\{ r(s,\pi_0(s)) + \gamma \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} p(r', s'|s, \pi_0(s)) v^*(s') \right\} - \gamma\varepsilon \\ &= \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} p(r', s'|s, a) v^*(s') \right\} - \gamma\varepsilon, \end{aligned}$$

where the last equality follows from the definition of π_0. Combining this with Equation (2), i.e. $v^*(s) \geq v_\pi(s)$, for all states $s \in \mathcal{S}$ and any policy $\pi \in \Pi$, we get

$$v^*(s) \geq v_\pi(s) \geq \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} p(r', s'|s, a) v^*(s') \right\} - \gamma\varepsilon.$$

Since ε was arbitrary,

$$v^*(s) \geq \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} p(r', s'|s, a) v^*(s') \right\}$$

also holds, and the proof is finished by noting that $s \in \mathcal{S}$ was also arbitrary. □

3 Operators, contractions, and fixed points

Note that $v, v_\pi \in \mathbb{R}^{\mathcal{S}}$ (i.e., they are functions mapping $\mathcal{S}$ to $\mathbb{R}$). Now define $T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ as follows: for $v \in \mathbb{R}^{\mathcal{S}}$, $f = T(v)$ should satisfy

$$f(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v(s') \right\}, \qquad s \in \mathcal{S}.$$

Following the standard convention, we will write Tv instead of $T(v)$. This removes some clutter. With this, the above equality can be also written as

$$(Tv)(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right\}, \qquad s \in \mathcal{S}\,. \tag{6}$$

Since T maps functions to functions, it is a "higher order function". In math, such higher order functions are often called *operators* and we will follow this convention. We collect all the above into a definition:

Definition 3.1 (Bellman optimality operator)**.** The Bellman optimality operator underlying the finite MDP M, equipped with the discounted total expected reward criterion, is the operator $T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ that satisfies Equation (6).

With the help of T, the result of Theorem 2.2 can be written in the shorter form

$$v^* = Tv^*\,.$$

Definition 3.2 (Fixed points)**.** Given a function G whose range is the same as its domain, if there exists an element x in the domain of G that satisfies $G(x) = x$, we say that x is *a fixed point* of G. The set of fixed points of G is denoted by $\textsc{Fixed}(G)$.

With this notation, we have $v^* \in \textsc{Fixed}(T)$. Are there any other fixed points of T?

Theorem 3.3. *The optimal value function v^* is the unique fixed point of T.*

To prove this result, we introduce the notion of contractions and we recall the powerful *contraction mapping theorem*. Let $\|\cdot\|$ be any norm on $\mathbb{R}^{\mathcal{S}}$, where $\mathbb{R}^{\mathcal{S}}$ is treated as the $|\mathcal{S}|$-dimensional Euclidean space. Recall that $\mathbb{R}^{\mathcal{S}}$ is a complete space with $\|\cdot\|$.[3]

Definition 3.4 (Lipchitz maps)**.** Let $p, p' \ge 1$ be natural numbers, $\|\cdot\|$ be a norm on $\mathbb{R}^p$, and $\|\cdot\|'$ be a norm on $\mathbb{R}^{p'}$. Then the map $F : \mathbb{R}^p \to \mathbb{R}^{p'}$ is *L-Lipschitz* with respect to these norms, if

$$\|F(x) - F(y)\|' \le L\,\|x - y\|\,, \qquad x, y \in \mathbb{R}^p\,.$$

Note that we allow $p \ne p'$ in this definition. This will be useful later.

Proposition 3.5. *If F is Lipschitz, then it is also continuous.*

Proof. To show continuity, it is enough to prove that for any $x_n \to x$ ($x_n, x \in \mathbb{R}^p$), we also have $F(x_n) \to F(x)$. Recalling that pointwise and norm-wise convergence are the same thanks to the finiteness of $p' > 0$, it suffices to show that $\|F(x_n) - F(x)\| \to 0$ as $n \to \infty$. But this is immediate from the Lipschitzness of F:

$$\|F(x_n) - F(x)\|' \le L\|x_n - x\| \to 0 \text{ as } n \to \infty.$$

□

Definition 3.6 (Contractions, non-expansions)**.** A map F that is L-Lipschitz with $L \le 1$ is called a *non-expansion*, while if $L < 1$, the map F is called a *contraction*. Any value $0 \le \alpha < 1$ such that F is α-Lipschitz is called a contraction factor of F.

Theorem 3.7 (Contraction mapping theorem)**.** *Let $\mathcal{S}$ be finite. Let $F : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ and $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^{\mathcal{S}}$. Assume that F is a contraction with the contraction factor $\alpha \in [0, 1)$. Then,* $\textsc{Fixed}(F)$ *has a unique element and this element is the limit of the sequence $x_{n+1} = F(x_n)$, where $x_0 \in \mathbb{R}^{\mathcal{S}}$ is arbitrary and $n \geq 0$. Finally, for $n \geq 0$ it holds that $\|x_n - x\| \leq \alpha^n \|x_0 - x\|$.*

This result is also known as Banach's fixed point theorem.

Proof. Let us first show that $(x_n)_{n\geq 0}$ converges to some vector that is the fixed point of F. For this, we show that it is a Cauchy sequence, i.e., a sequence of decreasing oscillations, since as it is well known, Cauchy sequences in finite dimensional sequences have a limit.[4] Thus, we need to show that

$$\lim_{n\to\infty} \sup_{m\geq n} \|x_m - x_n\| = 0 \,.$$

Fix $m \geq n \geq 0$. By the triangle inequality,

$$\begin{aligned} \|x_m - x_n\| &= \|(x_m - x_{m-1}) + (x_{m-1} - x_{m-2}) + \cdots + (x_{n+1} - x_n)\| \\ &\leq \|x_m - x_{m-1}\| + \|x_{m-1} - x_{m-2}\| + \cdots + \|x_{n+1} - x_n\| \,. \end{aligned}$$

Now, for $i \geq 1$, $x_{i+1} - x_i = F(x_i) - F(x_{i-1})$. Hence, $\|x_{i+1} - x_i\| = \|F(x_i) - F(x_{i-1})\| \leq \alpha\|x_i - x_{i-1}\| \leq \cdots \leq \alpha^i \|x_1 - x_0\|$. Thus,

$$\|x_m - x_n\| \leq (\alpha^{m-1} + \cdots + \alpha^n)\|x_1 - x_0\| \leq \frac{\alpha^n}{1-\alpha}\|x_1 - x_0\| \,.$$

Since m was arbitrary, it follows that we also have

$$\sup_{m\geq n} \|x_m - x_n\| \leq \frac{\alpha^n}{1-\alpha}\|x_1 - x_0\| \to 0, \text{ as } n \to \infty \,.$$

This implies that x_n is convergent. Let $x \in \mathbb{R}^{\mathcal{S}}$ be its limit.

We now show that $x \in \textsc{Fixed}(F)$. For this, note that $F(x_n) \to F(x)$ as $n \to \infty$ because F is continuous (cf. Proposition 3.5). But we also have $F(x_n) = x_{n+1} \to x$ as $n \to \infty$. Hence, $x = F(x)$, proving that $x \in \textsc{Fixed}(F)$.

Next, we show that $\textsc{Fixed}(F)$ has a single element. Indeed, for any $x, x' \in \textsc{Fixed}(F)$,

$$\|x - x'\| = \|F(x) - F(x')\| \leq \alpha\|x - x'\| \,.$$

Reordering the above equation and dividing by $(1-\alpha) > 0$, we get that $\|x - x'\| \leq 0$. Since $\|x - x'\| \geq 0$ also holds, $\|x - x'\| = 0$. Hence $x = x'$ by the properties of norms.

For the last part of the theorem note that for $n \geq 1$,

$$\|x_n - x\| = \|F(x_{n-1}) - F(x)\| \leq \alpha\|x_{n-1} - x\| \leq \cdots \leq \alpha^n \|x_0 - x\| \,.$$

□

The bound $\|x_n - x\| \leq \alpha^n \|x_0 - x\|$ reassures us that iteratively applying F gives rise to a sequence that rapidly converges to the unique fixed point of F, with the errors decreasing from the initial error $\|x_0 - x\|$ at a geometric rate. This suggest that if we ever need a good approximation to a fixed point to some contraction map F, we should start from some x_0, hopefully, not too far from x and then keep applying F in an iterative fashion. But since x is unknown, if we want to know how far x_n is from x, the above bound is not useful. But we can use that x_n is getting close to x to get a bound that is free of x:

Proposition 3.8. *For the setting of the contraction mapping theorem,* $\|x_n - x\| \leq \frac{\alpha^n}{1-\alpha^n}\|x_0 - x_n\|$.

Note that the above bound is almost as good as the bound $\|x_n - x\| \leq \alpha^n \|x_0 - x\|$. This is because the term $1/(1-\alpha^n)$ converges rapidly (from above) to 1. In particular, $1/(1-\alpha^n) \leq 1 + 2\alpha^n$ assuming that n is large enough so that $\alpha^n \leq 0.5$.

Proof. We already know that $\|x_n - x\| \leq \alpha^n \|x_0 - x\|$. We also have

$$\|x_0 - x\| \leq \|x_0 - x_n\| + \|x_n - x\| .$$

Hence,

$$\|x_n - x\| \leq \alpha^n \|x_0 - x_n\| + \alpha^n \|x_n - x\| .$$

Reordering and solving for $\|x_n - x\|$ gives

$$\|x_n - x\| \leq \frac{\alpha^n}{1-\alpha^n}\|x_0 - x_n\| .$$

□

An alternative to the previous result is as follows:

Proposition 3.9. *We have*

$$\|x_n - x\| \leq \frac{\|x_n - x_{n+1}\|}{1-\alpha} \leq \frac{\alpha^n}{1-\alpha}\|x_1 - x_0\| .$$

Proof. We have $\|x_n - x\| = \|x_n - F(x)\| \leq \|x_n - F(x_n)\| + \|F(x_n) - F(x)\| \leq \|x_n - F(x_n)\| + \alpha\|x_n - x\|$. Reordering and solving for $\|x_n - x\|$ gives the first bound. The second follows by noting that $\|x_n - x_{n+1}\| \leq \alpha^n \|x_0 - x_1\|$. □

The final tool that will be useful is the following:

Proposition 3.10. *Let $p, p', p'' \geq 1$ be three natural numbers, and $F : \mathbb{R}^p \to \mathbb{R}^{p'}$ and $G : \mathbb{R}^{p'} \to \mathbb{R}^{p''}$ be two functions. Fix the norms $\|\cdot\|$, $\|\cdot\|'$, and $\|\cdot\|'''$ on $\mathbb{R}^p$, $\mathbb{R}^{p'}$, and $\mathbb{R}^{p''}$, respectively. Assume that for some $L, L' > 0$, F is L-Lipschitz and G is L'-Lipschitz with respect to the respective norms. Then $G \circ F : \mathbb{R}^p \to \mathbb{R}^{p''}$ is LL'-Lipschitz.*[5]

Proof. Take $x, y \in \mathbb{R}^p$. Then,

$$\begin{aligned} \|G(F(x)) - G(F(y))\| &\leq L'\|F(x) - F(y)\| && (G \text{ is } L'\text{-Lipschitz}) \\ &\leq LL'\|x - y)\| . && (F \text{ is } '\text{-Lipschitz}) \end{aligned}$$

□

With these tools, we are ready to give the proof of Theorem 3.3.

Proof of Theorem 3.3. We want to show that $\textsc{Fixed}(T) = \{v^*\}$. By the contraction mapping theorem, it suffices to show that T is a contraction. For this, we need to choose a norm first. Let us choose the maximum norm: $\|v\| = \max_i |v_i|$.

Also, we view T as the composition of a number of maps. These are $P : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined using

$$(Pv)(s,a) = \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s')\,,$$

the map $L : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined by

$$(Lq)(s,a) = r(s,a) + \gamma q(s,a)\,,$$

and the map $M : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$, defined by

$$(Mq)(s) = \max_{a\in\mathcal{A}} q(s,a)\,.$$

It is easy to see[6] that $T = M \circ L \circ P$ and also that P is a non-expansion, L is a γ-contraction, and M is a non-expansion. Hence, by Proposition 3.10, T is a γ-contraction. Then, the contraction mapping theorem implies that T has a unique fixed-point. □

During this proof, we also proved the following:

Proposition 3.11. *T is a γ-contraction with respect to the maximum norm (in short, T is a max-norm contraction).*

Next, fix a memoryless policy $\pi \in \mathrm{ML}$ and define the operator $T_\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ using

$$(T_\pi v)(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right\}\,, \qquad s \in \mathcal{S}\,.$$

Then, we also have the following:

Proposition 3.12. *For any memoryless policy $\pi \in \mathrm{ML}$, the operator T_π is a max-norm contraction with the contraction factor γ, and its unique fixed point is v_π.*

Proof. That v_π is the fixed point of T_π has been shown before. That T_π is a max-norm contraction with contraction factor γ follows as the proof of the previous result. (Alternatively, this can be a corollary of the previous proposition that T is a contraction. To see this, define a new MDP $M' = (\mathcal{S}, \{1\}, \mathcal{R}, p')$ with $p'(r', s'|s, 1) = \sum_{a\in\mathcal{A}} \pi(a|s)p(r', s'|s, a)$. Now, the Bellman optimality operator T' in this MDP is the same as T_π, i.e. $T' = T_\pi$. Since T' is a contraction with contraction factor γ, it follows that T_π is a contraction with the same contraction factor). □

4 The fundamental theorem of MDPs

Theorem 4.1 (Fundamental theorem)**.** *If $\pi \in \mathrm{ML}$ is a memoryless policy such that*

$$\pi(s) \in \arg\max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)v^*(s') \right\}$$

then π is an optimal policy of M, i.e.,

$$v_\pi = v^*\,.$$

Proof. Take a policy π as in the theorem statement. By definition, we have $T_\pi v^* = Tv^*$ and by Theorem 2.2, $Tv^* = v^*$. Hence, $T_\pi v^* = v^*$. Let $k \geq 0$. It follows that $T_\pi^{k+1} v^* = T_\pi^k v^* = T_\pi^{k-1} v^* = \cdots = T_\pi v^* = v^*$. Letting $k \to \infty$, by Proposition 3.12, we have that $T_\pi^{k+1} v^* \to v_\pi$. Hence, $v_\pi = v^*$, which was the result to be proven. □

The result is called the fundamental theorem, as it allows us to restrict the search for optimal policies to the set of memoryless policies, which is a much more restricted set than the set of all policies.

Definition 4.2 (Greedy policy)**.** Let $v \in \mathbb{R}^{\mathcal{S}}$. Then a (memoryless) policy π is called greedy with respect to v, if

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v(s') \right\} .$$

Alternatively, π is greedy with respect to v if $T_\pi v = Tv$.

With this definition, what the fundamental theorem states is that any policy that is greedy with respect to v^* is optimal.

Notes

1. Generally, for functions $f, g : \mathcal{D} \to \mathbb{R}$, we write $f \le g$, if $f(x) \le g(x)$ for all $x \in \mathcal{D}$.

2. Recall that for $A \subset \mathbb{R}$, $\sup A$ means the unique real number that is the smallest among all the "upper bounds" on A. It is a defining property of the real numbers that this least upper bound exists and is unique. If the set A was closed (i.e., it contains the limit points of any sequence that takes values in A), then $\sup A = \max A$ (recall that $\max A$ is simply the largest element of A). But for an open set A, such as $A = [a, b)$ for some $a < b$ (i.e., A is an interval open from above), $\max A$ does not exist, while $\sup A = b$.

3. Without the loss of generality, we may assume that $\mathcal{S} = [S] := \{1, 2, \dots, S\}$, where $S = |\mathcal{S}|$. We can then identify $v \in \mathbb{R}^{\mathcal{S}}$ with the vector $(v(1), \dots, v(S))^\top \in \mathbb{R}^S$. Here, we use $u^\top$ to denote the transpose of u, which is needed because $(v(1), \dots, v(S))$, by default, denotes a row-vector, while, by convention, elements of $\mathbb{R}^S$ are column vectors. Recall some vector space operations: Adding vectors, multiplying vectors by a scalar constant (a real valued constant to be precise), and finally multiplying vectors by matrices. Also recall that $0 \in \mathbb{R}^S$ denotes the all-zero vector $(0, 0, \dots, 0)^\top$. Further, recall that a norm $\|\cdot\|$ on the Euclidean space $\mathbb{R}^S$ is an $\mathbb{R}^S \to [0, \infty)$ function such that for any $u, v \in \mathbb{R}^S$ and $\alpha \in \mathbb{R}$, the following hold: *(i)* $\|v\| = 0$ if and only if $v = 0$; *(ii)* $\|\alpha v\| = |\alpha|\|v\|$ (positive homogeneity); and *(iii)* $\|u + v\| \le \|u\| + \|v\|$ (triangle inequality). Examples of norms are the maximum norm, $\|v\| = \max_{i \in [S]} |v_i|$, the p-norms, $\|v\|_p = (\sum_{i \in [S]} |v_i|^p)^{1/p}$, etc. We will often use the maximum norm here. Recall that for $x_n \in \mathbb{R}^S$ (for $n \ge 0$) and some $x \in \mathbb{R}^S$, $x_n \to x$ if $(x_n)_i \to (x)_i$ holds for all $i \in S$ (here, $(y)_i$ denotes the ith component of the vector y). Sometimes, this mode of convergence is called componentwise. (Sometimes it is also called pointwise.) These terminologies (componentwise or pointwise convergence) are mainly for distinguishing this notion of convergence from other notions of convergence, such as convergence in norm: We say that x_n converges to x in norm $\|\cdot\|$, if $\|x_n - x\| \to 0$. Thanks to S being finite, the two modes of convergence are the same.

4. It is known that $\mathbb{R}^S$ is a *complete vector space*, regardless the norm chosen. This means that any sequence $(v_n)_{n \ge 0}$ taking values in $\mathbb{R}^S$ that is Cauchy in the sense that $\lim_{n \to \infty} \sup_{m \ge n} \|v_n - v_m\| \to 0$, it holds that $(v_n)_{n \ge 0}$ converges to some element v of $\mathbb{R}^S$. Let us call $\varepsilon_n := \sup_{m \ge n} \|v_n - v_m\|$ the oscillation of $(v_t)_{t \ge 0}$ after $t = n$. With this terminology we see that a sequence is Cauchy if its oscillations $(\varepsilon_n)_{n \ge 0}$ are vanishing (which is another way of saying that they are converging to zero).

5. The map $G \circ F$ is defined as the composition of F and G: First, F is applied to some input and then G is applied on the value returned by F. Formally, $(G \circ F)(x) = G(F(x))$, for any $x \in \mathbb{R}^p$.

6. That P is a nonexpansion can be proven as follows: Let $v, v' \in \mathbb{R}^{\mathcal{S}}$. Then $\|Pv - Pv'\| = \|P(v - v')\|$. Now, for $d = v - v'$, we have $|(Pd)(s,a)| = |\sum_{s' \in \mathcal{S}} p(s'|s,a) d(s')| \le \sum_{s' \in \mathcal{S}} p(s'|s,a)|d(s')| \le \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{s'' \in \mathcal{S}} |d(s'')| = \|d\| \sum_{s' \in \mathcal{S}} p(s'|s,a) = \|d\|$. Taking the maximum over $(s,a) \in \mathcal{S} \times \mathcal{A}$, we get $\|Pd\| \le \|d\|$.

   That L is a γ-contraction is left as an exercise. That M is a nonexpansion holds because of the following argument: Let $q, q' \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$. Fix $s \in \mathcal{S}$. Assume, without the loss of generality, that $(Mq)(s) \ge (Mq')(s)$. Let $a_0 \in \mathcal{A}$ be such that $(Mq)(s) = q(s, a_0)$. Then, $0 \le |(Mq)(s) - (Mq')(s)| = (Mq)(s) - (Mq')(s) = \max_{a \in \mathcal{A}} q(s,a) - \max_{a \in \mathcal{A}} q'(s,a) \le \max_{a \in \mathcal{A}} q(s,a) - q'(s,a_0) = q(s,a_0) - q'(s,a_0) = |q(s,a_0) - q'(s,a_0)| \le \max_{a \in \mathcal{A}} |q(s,a) - q'(s,a)| \le \|q - q'\|$. Since $s \in \mathcal{S}$ was arbitrary, $\|Mq - Mq'\| \le \|q - q'\|$, finishing the proof.