Notes on finite MDPs, Bellman equations and policy improvement à la Sutton/Barto

Jens Flemming*

February 2, 2024

Key words: Markov decision processes, Bellman equations, policy improvement theorem, reinforcement learning, dynamic programming

Abstract

The book *Reinforcment Learning: An Introduction* by Sutton and Barto [1] is the standard text book for introductory courses to reinforcement learning. Next to concrete algorithms and extensive examples the book contains several fundamental results related to Markov decision processes (MDPs) and Bellman equations in Chapters 3 and 4. Unfortunately some proofs are missing, some theorems lack precise formulation, and for some results the line of arguments is quite garbled.

In this note we provide all missing proofs, give precise formulations of theorems and untangle the line of arguments. Further, we avoid using random variables and their expected values. Since we (like Sutton/Barto) restrict our attention to finite MDPs all expected values can be made explicit avoiding overloaded notation and murky conclusions.

This article bridges the gap between introductory literature like Sutton/Barto and research literature containing exact formulations and proofs of relevant results, but being less accessible to beginners due to higher generality and complexity.

1 Notation

We use same notation as Sutton/Barto in [1, Chapters 3 and 4]. For a finite Markov decision process $\mathcal{S}$ denotes the finite set of states, $\mathcal{A}(s)$ denotes the

*Zwickau University of Applied Sciences, Faculty of Physical Engineering/Computer Sciences , D-08012 Zwickau, Germany, jens.flemming@fh-zwickau.de.

finite set of action available in state $s \in \mathcal{S}$, and $\mathcal{R} \subseteq \mathbb{R}$ denotes the finite set of rewards.

For discrete time steps $t = 0, 1, 2, \ldots$ the interaction between agent and environment yields a trajectory

$$s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, \ldots, \tag{1.1}$$

with initial state s_0 and initial action a_0 leading to reward r_1 and state s_1, and so on.

The environment dynamics p map each tuple $(s', r, s, a) \in \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A}(s)$ to the probability that state s' and reward r are observed if action a is taken in state s. Consequently,

$$p(s', r, s, a) \in [0, 1] \qquad \text{for all } s' \in \mathcal{S},\ r \in \mathcal{R},\ s \in \mathcal{S},\ a \in \mathcal{A}(s) \tag{1.2}$$

and

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r, s, a) = 1 \qquad \text{for all } s \in \mathcal{S},\ a \in \mathcal{A}(s). \tag{1.3}$$

A policy π maps each pair (a, s) to a probability that action a is taken by the agent in state s. Thus,

$$\pi(a, s) \in [0, 1] \qquad \text{for all } a \in \mathcal{A}(s),\ s \in \mathcal{S} \tag{1.4}$$

and

$$\sum_{a \in \mathcal{A}(s)} \pi(a, s) = 1 \qquad \text{for all } s \in \mathcal{S}. \tag{1.5}$$

2 Return

Definition 1. A state $s \in \mathcal{S}$ is an *end state* if $\mathcal{A}(s) = \emptyset$, that is, if there is no action the agent can take in that state.

Definition 2. A reinforcement learning task is *episodic* if $\mathcal{S}$ contains an end state. Else, the task is *continuing*.

If in an episodic task the agent reaches an end state, interaction with the environment stops and the trajectory is finite. If there is no end state (continuing task) or the end state is never reached by the agent, the trajectory is an infinite sequence. Sometimes trajectories are called *episodes*, where the latter puts some emphasis on the interactions and the former on the interactions' results.

In [1, Section 3.3] reinforcement learning tasks are called episodic if 'the agent–environment interaction [...] break[s] naturally into identifiable episodes'. From this somewhat imprecise definition it is not clear whether tasks with end states are called episodic even if there exist policies never reaching an end state. Such tasks with end states but possibly infinite trajectories have to be handle with care.

For episodic tasks one usually is interested in the total reward collected by the agent during one episode.

Definition 3. The *return of an episode in an episodic reinforcement learning task* is

$$\sum_{t=1}^{T} r_t, \tag{2.1}$$

if the trajectory reaches an end state after T steps, and

$$\sum_{t=1}^{\infty} r_t, \tag{2.2}$$

if the trajectory is infinite and the sum converges.

Note that for infinite trajectories of episodic tasks return may be undefined.

Definition 4. The *return of a continuing reinforcement learning task* is

$$\sum_{t=1}^{\infty} \gamma^{t-1} r_t \tag{2.3}$$

with *discounting* parameter $\gamma \in [0, 1)$.

The return of a continuing task always is a finite value, because the reward set $\mathcal{R}$ is finite:

$$\left| \sum_{t=1}^{\infty} \gamma^{t-1} r_t \right| \leq \left(\sum_{t=1}^{\infty} \gamma^{t-1} \right) \max_{r \in \mathcal{R}} |r| = \frac{1}{1-\gamma} \max_{r \in \mathcal{R}} |r|. \tag{2.4}$$

If the agent takes action $a \in \mathcal{A}(s)$ in the initial state s and the environment answers with reward r and new state s', then the return g of corresponding trajectory $s, a, r, s', \ldots$ can be computed from the return g' of the subtrajectory starting at s':

$$g = r + \gamma g' \tag{2.5}$$

with $\gamma = 1$ for episodic tasks and $\gamma < 1$ for continuing tasks.

3 Value functions

Definition 5. The *state value function* v_π for a policy π maps each state $s \in \mathcal{S}$ to the return the agent will obtain on average if starting at state s and following policy π.

Definition 6. The *action value function* q_π for a policy π maps each pair (s, a) of state $s \in \mathcal{S}$ and action $a \in \mathcal{A}(s)$ to the return the agent will obtain on average if starting at state s, taking action a, and then following policy π.

For continuing tasks each policy has well-defined state and action value functions. For episodic tasks only policies allowing for finite trajectories only always have well-defined value functions. If the policy of an episodic task allows for infinite trajectories, there might be no value functions, because return might be undefined.

Obviously,

$$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a, s)\, q_\pi(s, a) \qquad \text{for all } s \in \mathcal{S} \tag{3.1}$$

and, by equation (2.5),

$$q_\pi(s, a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r, s, a) \left(r + \gamma\, v_\pi(s')\right) \qquad \text{for all } s \in \mathcal{S},\ a \in \mathcal{A}(s). \tag{3.2}$$

for the value functions of a policy π, where $\gamma = 1$ for episodic tasks and $\gamma < 1$ for continuing tasks.

The explicit formula for state values has the structure

$$\begin{aligned} v_\pi(s) = {} & \text{expected reward after first action} \\ & + \gamma \times \text{expected reward after second action} \\ & + \gamma^2 \times \text{expected reward after third action} \\ & + \dots . \end{aligned} \tag{3.3}$$

Making expected rewards explicit we see

$$\begin{aligned} v_\pi(s) = &\sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \pi(a,s)\, p(s',r,s,a)\, r \\ &+ \gamma \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \sum_{\substack{a'\in\mathcal{A}(s')\\ s''\in\mathcal{S}\\ r'\in\mathcal{R}}} \pi(a,s)\, p(s',r,s,a)\, \pi(a',s')\, p(s'',r',s',a')\, r' \\ &+ \gamma^2 \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \sum_{\substack{a'\in\mathcal{A}(s')\\ s''\in\mathcal{S}\\ r'\in\mathcal{R}}} \sum_{\substack{a''\in\mathcal{A}(s'')\\ s'''\in\mathcal{S}\\ r''\in\mathcal{R}}} \dots \\ &+ \dots . \end{aligned} \tag{3.4}$$

For $\gamma < 1$ this value is well-defined because expected rewards are bounded by $\max_{r\in\mathcal{R}} |r|$, cf. equation (2.4). For $\gamma = 1$ (episodic tasks) the sum of expected rewards may converge or not. The standard situation for convergence is that all possible trajectories starting at s have at most T steps for some $T \in \mathbb{N}$ independent of the concrete trajectory. For all but finitely many expected rewards we then have empty sums $\sum_{a\in\mathcal{A}(s_{\text{end}})} \dots = 0$, because the action set $\mathcal{A}(s_{\text{end}})$ of an end state s_{end} is empty.

Analogously, for action values we have

$$\begin{aligned} q_\pi(s,a) = &\sum_{\substack{s'\in\mathcal{S}\\ r\in\mathcal{R}}} p(s',r,s,a)\, r \\ &+ \gamma \sum_{\substack{s'\in\mathcal{S}\\ r\in\mathcal{R}}} \sum_{\substack{a'\in\mathcal{A}(s')\\ s''\in\mathcal{S}\\ r'\in\mathcal{R}}} p(s',r,s,a)\, \pi(a',s')\, p(s'',r',s',a')\, r' \\ &+ \gamma^2 \sum_{\substack{s'\in\mathcal{S}\\ r\in\mathcal{R}}} \sum_{\substack{a'\in\mathcal{A}(s')\\ s''\in\mathcal{S}\\ r'\in\mathcal{R}}} \sum_{\substack{a''\in\mathcal{A}(s'')\\ s'''\in\mathcal{S}\\ r''\in\mathcal{R}}} \dots \\ &+ \dots . \end{aligned} \tag{3.5}$$

4 Optimal policies

Definition 7. Policy π_1 *is at least as good as* policy π_2 if

$$v_{\pi_1}(s) \geq v_{\pi_2}(s) \qquad \text{for all } s \in \mathcal{S}. \tag{4.1}$$

Definition 8. A policy π is an *optimal policy* if it is at least as good as every other policy.

The following theorem lacks a proof in Sutton/Barto. There exist (at least) two variants of the proof, one based on Banach's fixed-point theorem, the other using Zorn's lemma. Here we follow the second variant because it does not require introduction of metrics or vector space norms.

Theorem 9. *There always is an optimal policy.*

Proof. The set of all policies has no linear order, because we cannot compare every policy to every other policy in terms of 'at least as good as', cf. Definition 7. But we may find pairs of policies for which we can say that the one policy is at least as good as the other. From such pairs we may form increasing chains of policies (from bad to good). If we can show that every such chain of policies has a maximal element (a best policy), then the Zorn lemma yields existence of a maximal element w. r. t. the whole set of policies. That is, Zorn's lemma then guarantees existence of an optimal policy.

It remains to show that every chain of policies has a maximal element. For finite chains obviously the last element is the maximal one. For infinite chains we argue as follows: Each infinite chain is a bounded set of functions over a finite set (all pairs of states and actions), because policies take values in $[0, 1]$. Bounded subsets of finite-dimensional spaces always contain a convergent sequence, say $\pi_1, \pi_2, \ldots$ (Bolzano-Weierstrass theorem). Let π be the pointwise limit of the sequence, that is,

$$\pi(a, s) := \lim_{n\to\infty} \pi_n(a, s) \qquad \text{for all } s \in \mathcal{S},\, a \in \mathcal{A}(s). \tag{4.2}$$

Now π is a policy, too (values in $[0, 1]$, sum over actions is 1). The sequence of corresponding state value functions $v_1, v_2, \ldots$ converges, too, because bounded increasing sequences of functions over a finite set always converge. Denote the pointwise limit by v. Taking equation (3.4) for v_n and π_n and letting $n \to \infty$ on both sides yields (3.4) for v and π (interchange of limit and infinite summation is okay here because expected rewards for $\pi_1, \pi_2, \ldots$ are uniformly bounded). Thus, v is the state value function for π. By construction of v the policy π is at least as good as $\pi_1, \pi_2, \ldots$. Because upper bounds of subsequences also are upper bounds of the whole chain, π is at least as good as all policies in the chain under consideration. □

As a consequence of Definition 7 all optimal policies share one and the same optimal state value function. From equation (3.2) we easily deduce that they also share a common action value function.

5 Policy improvement

Definition 10. A policy π_{g} is *greedy w. r. t. an action value function* q_π of a policy π, if

$$\pi_{\mathrm{g}}(a,s) > 0 \quad \Rightarrow \quad q_\pi(s,a) = \max_{\tilde{a}\in\mathcal{A}(s)} q_\pi(s,\tilde{a}) \tag{5.1}$$

holds for all $s \in \mathcal{S}$ and all $a \in \mathcal{A}(s)$.

Greedy policies always choose an action with highest value given that all further actions are chosen following π. Greedy policies may be deterministic or not.

Theorem 11 (Policy improvement theorem)**.** *For $\gamma < 1$ each greedy policy w. r. t. the action value function of some policy π is at least as good as π. The same holds true for $\gamma = 1$ if for both the greedy policy and π there is $T \in \mathbb{N}$ such that all possible trajectories reach an end state within at most T time steps.*

Proof. Let π_{g} be some greedy policy w. r. t. q_π. Then

$$\pi_{\mathrm{g}}(a,s) > 0 \quad \Rightarrow \quad q_\pi(s,a) = \max_{\tilde{a}\in\mathcal{A}(s)} q_\pi(s,\tilde{a}) \qquad \text{for all } s \in \mathcal{S} \tag{5.2}$$

and, thus,

$$\begin{aligned} v_\pi(s) &= \sum_{a\in\mathcal{A}(s)} \pi(a,s)\, q_\pi(s,a) \\ &\leq \left(\sum_{a\in\mathcal{A}(s)} \pi(a,s)\right) \max_{\tilde{a}\in\mathcal{A}(s)} q_\pi(s,\tilde{a}) \\ &= \left(\sum_{a\in\mathcal{A}(s)} \pi_{\mathrm{g}}(a,s)\right) \max_{\tilde{a}\in\mathcal{A}(s)} q_\pi(s,\tilde{a}) \\ &= \sum_{a\in\mathcal{A}(s)} \pi_{\mathrm{g}}(a,s)\, q_\pi(s,a) \end{aligned} \tag{5.3}$$

for all $s \in \mathcal{S}$.

We define shorthands $p^{(0)} := p(s',r,s,a)$, $p^{(1)} := p(s'',r',s',a')$,... as well as $\pi_{\mathrm{g}}^{(0)} := \pi_{\mathrm{g}}(a,s)$, $\pi_{\mathrm{g}}^{(1)} := \pi_{\mathrm{g}}(a',s')$,... and alternatingly apply (5.3)

and (3.2) $N+1$ times to obtain

$$
\begin{aligned}
v_\pi(s) &\le \sum_{a\in\mathcal{A}(s)} \pi_{\mathrm{g}}^{(0)}\, q_\pi(s,a) \\
&= \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)} \left(r+\gamma\, v_\pi(s')\right) \\
&\le \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)} \left(r+\gamma \sum_{a'\in\mathcal{A}(s')} \pi_{\mathrm{g}}^{(1)}\, q_\pi(s',a')\right) \\
&= \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)}\, r + \gamma \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \sum_{a'\in\mathcal{A}(s')} \pi_{\mathrm{g}}^{(0)}\, p^{(0)}\, \pi_{\mathrm{g}}^{(1)}\, q_\pi(s', a_{\mathrm{g}}(s')) \\
&= \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)}\, r + \gamma \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \sum_{\substack{a'\in\mathcal{A}(s')\\ s''\in\mathcal{S}\\ r'\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)}\, \pi_{\mathrm{g}}^{(1)}\, p^{(1)} \left(r'+\gamma\, v_\pi(s'')\right) \\
&= \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)}\, r + \gamma \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \sum_{\substack{a'\in\mathcal{A}(s')\\ s''\in\mathcal{S}\\ r'\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)}\, \pi_{\mathrm{g}}^{(1)}\, p^{(1)}\, r' \\
&\qquad + \gamma^2 \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \sum_{\substack{a'\in\mathcal{A}(s')\\ s''\in\mathcal{S}\\ r'\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)}\, \pi_{\mathrm{g}}^{(1)}\, p^{(1)}\, v_\pi(s'') \\
&= \dots \\
&= \sum_{n=0}^{N} \gamma^n \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \cdots \sum_{\substack{a^{(n)}\in\mathcal{A}(s^{(n)})\\ s^{(n+1)}\in\mathcal{S}\\ r^{(n)}\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)} \cdots \pi_{\mathrm{g}}^{(n)}\, p^{(n)}\, r^{(n)} \\
&\qquad + \gamma^{N+1} \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \cdots \sum_{\substack{a^{(N)}\in\mathcal{A}(s^{(N)})\\ s^{(N+1)}\in\mathcal{S}\\ r^{(N)}\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)} \cdots \pi_{\mathrm{g}}^{(N)}\, p^{(N)}\, v_\pi\big(s^{(N+1)}\big)
\end{aligned}
\tag{5.4}
$$

for arbitrary $N \in \mathbb{N}$.

The first summand $\sum_{n=0}^{N} \gamma^n \dots$ coincides with the first $N+1$ summands in (3.4) for $v_{\pi_{\mathrm{g}}}$. The second summand $\gamma^{N+1} \dots$ is zero for episodic tasks

($\gamma = 1$) with finite trajectories if N is chosen large enough. Corresponding summands $N+2$, $N+3$,... in (3.4) are zero, too. This proves the theorem for $\gamma = 1$.

For continuing tasks ($\gamma < 1$) we see that

$$v_\pi(s) - \sum_{n=0}^{N} \gamma^n \sum_{\substack{a\in\mathcal{A}(s)\\ s'\in\mathcal{S}\\ r\in\mathcal{R}}} \cdots \sum_{\substack{a^{(n)}\in\mathcal{A}(s^{(n)})\\ s^{(n+1)}\in\mathcal{S}\\ r^{(n)}\in\mathcal{R}}} \pi_{\mathrm{g}}^{(0)}\, p^{(0)} \cdots \pi_{\mathrm{g}}^{(n)}\, p^{(n)}\, r^{(n)} \tag{5.5}$$

is bounded by

$$\gamma^{N+1} \max_{\tilde{s}\in\mathcal{S}} |v_\pi(\tilde{s})|. \tag{5.6}$$

For $N \to \infty$ the expression in (5.5) converges to $v_\pi(s) - v_{\pi_{\mathrm{g}}}(s)$ and corresponding bound in (5.6) converges to zero. Thus, $v_\pi(s) - v_{\pi_{\mathrm{g}}}(s) \le 0$. □

In Sutton/Barto the above theorem is formulated in Section 4.2 without precisely stating the assumptions. The prove is provided in form of an 'idea' only, neglecting the nasty, but important details. The transition from finitely to infinitely many steps there is not possible for episodic tasks allowing for infinite trajectories as the following example shows.

Example 12. Consider a 1-by-3 gridworld with cells numbered 1, 2, 3 from left to right, that is states are $\mathcal{S} = \{1, 2, 3\}$ and action sets are $\mathcal{A}(1) = \mathcal{A}(2) = \{\mathrm{L}, \mathrm{R}\}$ (go left or go right) and $\mathcal{A}(3) = \emptyset$ (end state). If the agent reaches state 3 in a step, reward is 1, else reward is 0. Going to the left in state 1 (hitting the wall) results in state 1 again and zero reward.

This is clearly an episodic task, so we choose $\gamma = 1$, resulting in return being either 1 or 0 for each trajectory. More precisely, all finite trajectories will have return 1, all infinite trajectories will have return 0. An example of a policy with infinite trajectories is to jump back and forth between cells 1 and 2, never going to cell 3.

An optimal policy clearly is 'always go to the right'. Corresponding state value function v_* satisfies $v_*(1) = 1 = v_*(2)$ and $v_*(3) = 0$. Corresponding action values all equal 1, too.

Since all actions have equal value, all policies are greedy w. r. t. the optimal action value function. Thus, by the policy improvement theorem (neglecting its assumptions) all policies have to be optimal. But state values for jumping back and forth between cells 1 und 2 obviously all are zero, that is, it's not on optimal policy.

This example shows, that the policy improvement theorem may fail for episodic task with (some) infinite trajectories.

Corollary 13. *Every optimal policy is greedy w. r. t. the optimal action value function.*

Proof. If π is an optimal policy that is not greedy w. r. t. the optimal action value function q_*, then there are a state $\bar{s}$ and an action $\bar{a} \in \mathcal{A}(\bar{s})$ such that

$$\pi(\bar{a}, \bar{s}) > 0 \qquad \text{and} \qquad q_*(\bar{s}, \bar{a}) < \max_{\tilde{a} \in \mathcal{A}(\bar{s})} q_*(\bar{s}, \tilde{a}) \tag{5.7}$$

(cf. Definition 10).

Now let π_{g} be some policy greedy w. r. t. q_*. Then

$$\begin{aligned} v_*(\bar{s}) &= \sum_{a \in \mathcal{A}(\bar{s})} \pi(a, \bar{s})\, q_*(\bar{s}, a) \\ &< \sum_{a \in \mathcal{A}(\bar{s})} \pi(a, \bar{s}) \max_{\tilde{a} \in \mathcal{A}(\bar{s})} q_*(\bar{s}, \tilde{a}) \\ &= \left(\sum_{a \in \mathcal{A}(\bar{s})} \pi(a, \bar{s}) \right) \max_{\tilde{a} \in \mathcal{A}(\bar{s})} q_*(\bar{s}, \tilde{a}) \\ &= \left(\sum_{a \in \mathcal{A}(\bar{s})} \pi_{\mathrm{g}}(a, \bar{s}) \right) \max_{\tilde{a} \in \mathcal{A}(\bar{s})} q_*(\bar{s}, \tilde{a}) \\ &= \sum_{a \in \mathcal{A}(\bar{s})} \pi_{\mathrm{g}}(a, \bar{s})\, q_*(\bar{s}, a) \\ &= v_{\pi_{\mathrm{g}}}(\bar{s}), \end{aligned} \tag{5.8}$$

which contradicts optimality of v_*. □

The assertion of the corollary is formulated in Chapter 3 of Sutton/Barto although the policy improvement theorem only appears in Chapter 4. In Sutton/Barto the corollary seems to be derived from the optimal Bellman equations (7.1), where Sutton/Barto seem to use the non-obvious fact that each solution to the optimal Bellman equations is an optimal value function. But the optimal Bellman equations aren't sufficient (only necessary) for a value function to be optimal. To get sufficiency, too, one needs the corollary or some other non-trivial idea. It's unclear how Sutton/Barto arrived at the corollary without the policy improvement theorem. See below for more detailled discussion of the solutions to the optimal Bellman equations.

6 Computing value functions

Given a policy π, from the relations (3.1) and (3.2) between state values and action values we immediately obtain equations

$$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a, s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r, s, a) \left(r + \gamma \, v_\pi(s') \right) \tag{6.1}$$

for all $s \in \mathcal{S}$ and

$$q_\pi(s, a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r, s, a) \left(r + \gamma \sum_{a' \in \mathcal{A}(s')} \pi(a', s') \, q_\pi(s', a') \right) \tag{6.2}$$

for all $s \in \mathcal{S}$ and all $a \in \mathcal{A}(s)$. These are the *Bellman equations* for state values and action values, respectively.

The Bellman equations for state values and action values both are systems of linear equations allowing to compute the value functions for all arguments without any need for exploration (if we know the environment dynamics p). In case of state values there are as many equations as there are states. In case of action values there are as many equations as there are state–action pairs.

Theorem 14. *For $\gamma < 1$ the only solutions to the Bellman equations* (6.1) *and* (6.2) *for a policy π are the value functions v_π and q_π.*

Proof. Above derivation of the Bellman equations shows that the value functions are solution. Thus, we only have to show uniqueness of solutions. Uniqueness for state values automatically implies uniqueness for action values due to (3.1) and (3.2).

Assume there are two solutions v_1 and v_2 to the Bellman equations for

state values. Then for each $s \in \mathcal{S}$ we have

$$\begin{aligned}
\big|v_1(s) - v_2(s)\big| &= \Bigg| \sum_{a\in\mathcal{A}(s)} \pi(a,s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s',r,s,a) \big(r + \gamma\, v_1(s')\big) \\
&\qquad - \sum_{a\in\mathcal{A}(s)} \pi(a,s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s',r,s,a) \big(r + \gamma\, v_2(s')\big) \Bigg| \\
&= \gamma \left| \sum_{a\in\mathcal{A}(s)} \pi(a,s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s',r,s,a) \big(v_1(s') - v_2(s')\big) \right| \\
&\le \gamma \sum_{a\in\mathcal{A}(s)} \pi(a,s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s',r,s,a) \big|v_1(s') - v_2(s')\big| \\
&\le \gamma \sum_{a\in\mathcal{A}(s)} \pi(a,s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s',r,s,a) \max_{s''\in\mathcal{S}} \big|v_1(s'') - v_2(s'')\big| \\
&= \gamma \sum_{a\in\mathcal{A}(s)} \pi(a,s) \max_{s''\in\mathcal{S}} \big|v_1(s'') - v_2(s'')\big| \\
&= \gamma \max_{s''\in\mathcal{S}} \big|v_1(s'') - v_2(s'')\big|. \qquad (6.3)
\end{aligned}$$

Now taking the maximum over all s we see

$$\max_{s\in\mathcal{S}} \big|v_1(s) - v_2(s)\big| \le \gamma \max_{s\in\mathcal{S}} \big|v_1(s) - v_2(s)\big|. \qquad (6.4)$$

But for $\gamma \in [0,1)$ this is only possible if $v_1(s) = v_2(s)$ for all s. Thus, there cannot be two different solutions to the Bellman equations. □

For $\gamma = 1$ a policy's value functions are solutions to the Bellman equations, if the value functions exist. But there might be other solutions, too.

The Bellman equations combined with the policy improvement theorem give rise to the policy iteration algorithm, cf. [1, Section 4.3].

7 Computing optimal value functions

Theorem 15. *Every policy π that is greedy w. r. t. to its own action value function q_π satisfies the* optimal Bellman equations

$$v_\pi(s) = \max_{a\in\mathcal{A}(s)} \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s',r,s,a) \big(r + \gamma\, v_\pi(s')\big) \qquad (7.1)$$

for all $s \in \mathcal{S}$ *and*

$$q_\pi(s,a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r, s, a) \left(r + \gamma \max_{a' \in \mathcal{A}(s')} q_\pi(s', a') \right) \tag{7.2}$$

for all $s \in \mathcal{S}$ *and all* $a \in \mathcal{A}(s)$.

Proof. Since the policy is greedy w. r. t. to its own action value function, by (3.1) we have

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a, s)\, q_\pi(s, a) = \max_{a \in \mathcal{A}} q_\pi(s, a) \qquad \text{for all } s \in \mathcal{S}.$$

This equality together with (3.2) yields both the optimal Bellman equations for state values and the optimal Bellman equations for action values. □

Sutton/Barto derive the optimal Bellman equations in Section 3.6 from

$$v_*(s) = \max_{a \in \mathcal{A}(s)} q_\pi(s, a) \tag{7.3}$$

with some optimal policy π. But this is a consequenc of Corollary 13, whose availability in this section of Sutton/Barto is dubious, cf. discussion below Corollary 13.

Optimal Bellman equations are systems of nonlinear equations. At the moment we only know that they have at least one solution (the optimal value functions, by Corollary 13). But it's unclear whether there could exist more solutions.

Theorem 16. *For* $\gamma < 1$ *the optimal value functions are the only solutions to the optimal Bellman equations* (7.1) *and* (7.2).

Proof. For proving uniqueness of solutions we need the inequality

$$\Big| \max_x f(x) - \max_x g(x) \Big| \le \max_x \big| f(x) - g(x) \big| \tag{7.4}$$

for arbitrary functions f and g taking arguments x from a finite set. To see that this inequality is true, assume $\max_x f(x) \ge \max_x g(x)$ (else, switch the roles of f and g) and let $\bar{x}$ be a maximizer of $f(x)$. Then

$$\begin{aligned} \Big| \max_x f(x) - \max_x g(x) \Big| &= f(\bar{x}) - \max_x g(x) \\ &\le f(\bar{x}) - g(\bar{x}) \\ &\le \max_x \big| f(x) - g(x) \big|. \end{aligned}$$

Now assume there are two solutions v_1 and v_2 to the optimal Bellman equations for state values. Then for each $s \in \mathcal{S}$ we have

$$\begin{aligned}
\big|v_1(s) - v_2(s)\big| &= \Bigg| \max_{a\in\mathcal{A}(s)} \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r, s, a) \big(r + \gamma\, v_1(s')\big) \\
&\qquad - \max_{a\in\mathcal{A}(s)} \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r, s, a) \big(r + \gamma\, v_2(s')\big) \Bigg| \\
&= \gamma \max_{a\in\mathcal{A}(s)} \Bigg| \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r, s, a) \big(v_1(s') - v_2(s')\big) \Bigg| \\
&\leq \gamma \max_{a\in\mathcal{A}(s)} \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r, s, a) \big|v_1(s') - v_2(s')\big| \\
&\leq \gamma \max_{a\in\mathcal{A}(s)} \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r, s, a) \max_{s''\in\mathcal{S}} \big|v_1(s'') - v_2(s'')\big| \\
&= \gamma \max_{a\in\mathcal{A}(s)} \max_{s''\in\mathcal{S}} \big|v_1(s'') - v_2(s'')\big| \\
&= \gamma \max_{s''\in\mathcal{S}} \big|v_1(s'') - v_2(s'')\big|.
\end{aligned}$$

Now taking the maximum over all s we see

$$\max_{s\in\mathcal{S}} \big|v_1(s) - v_2(s)\big| \leq \gamma \max_{s\in\mathcal{S}} \big|v_1(s) - v_2(s)\big|.$$

But for $\gamma \in [0, 1)$ this is only possible if $v_1(s) = v_2(s)$ for all s. Thus, there cannot be two different solutions to the optimal Bellman equations. □

Corollary 17. *A policy is optimal if and only if it is greedy w. r. t. its own action value function.*

Proof. That every optimal policy is greedy w. r. t. to its value function is stated by Corollary 13.

The other way round, if some policy is greedy w. r. t. to its own value function, then its value functions satisfy the optimal Bellman equations. But the optimal Bellman equations only have one solution, the optimal value functions. Thus, the policy has to be optimal. □

This corollary formulates an important stopping criterion for policy iteration algorithms. Sutton/Barto do not clearly state this important result. Some hints in this direction are given in Chapters 3 and 4. The clearest statement is in Section 4.6, but without proof. The reader of Sutton/Barto

does not see that this result is non-trivial since it is based on both the policy improvement theorem (via Corollary 13) and the uniqueness theorem for solutions of the optimal Bellman equations.

References

[1] Richard S. Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge (Massachusetts), London (England), second edition, 2018.