Reinforcement Learning: Exercises

Leon Lang David Quarel

April 2026

What you'll learn

- Understand Markov Decision Processes and the goal of the agent, for known environments.
- Understand the Bellman equation.
- Understand the policy improvement theorem, and how we can use it to iteratively solve for an optimal policy.
- Prove properties involving the Bellman equations, including the existence of optimal policies, the policy improvement theorem, rate of convergence of Bellman updates, and the convergence of Q-learning.

Setup

In this exercise sheet, we prove a variety of results that are behind the ARENA Intro to RL materials.

An **agent** interacts with an **environment** in discrete time steps $t = 0, 1, 2, \ldots$. On timestep t, the agent observes the current state $s_t \in \mathcal{S}$ and selects an action $a_t \in \mathcal{A}$ according to its policy. The environment then responds with a reward $r_{t+1} \in \mathbb{R}$ and a new state $s_{t+1} \in \mathcal{S}$. (The reward and next state are indexed by $t+1$ because they are produced by the environment after the agent's action.) The goal is to act so as to maximize expected discounted future reward.

These environments are called **Markov decision processes** (MDPs). They satisfy two key properties: (i) **stationarity** — the transition and reward functions do not change over time, and (ii) the **Markov property** — the distribution over the next state and reward depends only on the current state and action, not on the full history of past interactions. Together, these properties mean that the current state s_t is a sufficient summary of the past for the purpose of choosing optimal actions.

Definition 0.1 (Spaces and notation)**.**

- $\mathcal{S}$ — finite **state space**

- $\mathcal{A}$ — finite **action space**
- $\gamma \in (0,1)$ — **discount factor**
- ΔX — set of all probability distributions over a set X
- $[\![P]\!]$ — **Iverson bracket**: 1 if P is true, 0 if false

Definition 0.2 (Environment)**.** An MDP **environment** is specified by a pair (T, R):

- $T : \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$ is the **transition kernel**: given state s and action a, the next state is drawn $s' \sim T(\cdot \mid s, a)$.
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the **reward function**: on a transition from s to s' under action a, the agent receives reward $R(s, a, s')$. [a]

[a] Noting that $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ is a finite set, this implies that the rewards are bounded above by $R_{\max} = \max_{s,a,s'} R(s, a, s')$.

Definition 0.3 (Policy)**.** A **policy** $\pi : \mathcal{S} \to \Delta\mathcal{A}$ maps each state to a probability distribution over actions. Given state s:

- $\pi(\cdot \mid s)$ is a distribution over $\mathcal{A}$,
- $\pi(a \mid s) \in [0, 1]$ is the probability of choosing action a,
- the agent samples $a \sim \pi(\cdot \mid s)$.

A policy is **deterministic** if $\pi(a \mid s) \in \{0, 1\}$ for all a, s. In this case, we abuse notation and write $\pi(s) := \arg\max_{a \in \mathcal{A}} \pi(a \mid s)$ for the unique action selected at state s.

Definition 0.4 (Trajectory)**.** When policy π interacts with environment (T, R) starting from initial state s_0, the resulting **trajectory** $(s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, \ldots)$ is generated by

$$a_k \sim \pi(\cdot \mid s_k), \qquad s_{k+1} \sim T(\cdot \mid s_k, a_k), \qquad r_{k+1} = R(s_k, a_k, s_{k+1}),$$

for $k = 0, 1, 2, \ldots$. We write $\mathbb{E}_\pi[\,\cdot\,]$ for expectations over such trajectories.

1 The Bellman equation

Definition 1.1 (Return)**.** The **return** from time t is the discounted sum of future rewards:

$$G_t \;:=\; \sum_{k=t+1}^{\infty} \gamma^{k-t-1}\, r_k \;=\; r_{t+1} + \gamma\, r_{t+2} + \gamma^2\, r_{t+3} + \cdots .$$

It satisfies the recursion

$$G_t \;=\; r_{t+1} + \gamma\, G_{t+1}.$$

Definition 1.2 (Value function)**.** The **value function** of a policy π is the expected return when starting from state s and acting according to π:

$$\begin{aligned} V_\pi(s) \;&=\; \mathbb{E}_\pi[G_t \mid s_t = s] \\ &= \mathbb{E}_\pi\big[\, r_{t+1} + \gamma\, r_{t+2} + \gamma^2\, r_{t+3} + \cdots \,\big|\, s_t = s\big] \\ &= \sum_{a_t \in \mathcal{A}} \pi(a_t \mid s) \sum_{s_{t+1} \in \mathcal{S}} T(s_{t+1} \mid s, a_t) \sum_{a_{t+1} \in \mathcal{A}} \pi(a_{t+1} \mid s_{t+1}) \\ &\quad \sum_{s_{t+2} \in \mathcal{S}} T(s_{t+2} \mid s_{t+1}, a_{t+1}) \cdots \\ &\quad \big[\, R(s, a_t, s_{t+1}) + \gamma\, R(s_{t+1}, a_{t+1}, s_{t+2}) + \gamma^2\, R(s_{t+2}, a_{t+2}, s_{t+3}) + \cdots \big]. \end{aligned}$$

The expectation averages the return over all future trajectories $(a_t, s_{t+1}, a_{t+1}, s_{t+2}, \ldots)$ generated by sampling $a_k \sim \pi(\cdot \mid s_k)$ and $s_{k+1} \sim T(\cdot \mid s_k, a_k)$, starting from $s_t = s$.

Exercise 1.1. Prove the **Bellman equation**: for all $s \in \mathcal{S}$,

Note

Bellman Equation

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma\, V_\pi(s')\big).$$

Hint: condition on the first action and the first transition, then separate the immediate reward from the remaining return.

Solution to Exercise 1.1

Starting from the definition of V_π and using the recursion $G_t = r_{t+1} + \gamma\, G_{t+1}$:

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid s_t = s] = \mathbb{E}_\pi[r_{t+1} + \gamma\, G_{t+1} \mid s_t = s]\,.$$

Now condition on the first action $a_t = a$ and the first transition $s_{t+1} = s'$:

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} T(s' \mid s, a) \\ \times \mathbb{E}_\pi\big[r_{t+1} + \gamma\, G_{t+1} \;\big|\; s_t = s,\, a_t = a,\, s_{t+1} = s'\big]\,.$$

Given $s_t = s$, $a_t = a$, and $s_{t+1} = s'$, the immediate reward is deterministic: $r_{t+1} = R(s, a, s')$. For the remaining return, by the Markov property the future trajectory from time $t+1$ onward depends only on $s_{t+1} = s'$, so

$$\mathbb{E}_\pi\big[G_{t+1} \;\big|\; s_t = s,\, a_t = a,\, s_{t+1} = s'\big] = \mathbb{E}_\pi\big[G_{t+1} \;\big|\; s_{t+1} = s'\big] = V_\pi(s').$$

Substituting both back gives

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma\, V_\pi(s')\big).$$

Definition 1.3 (Action-value function)**.** The **action-value function** (or Q-function) of π is the expected return starting from state s, taking action a, and acting according to π thereafter:

$$Q_\pi(s, a) \;=\; \mathbb{E}_\pi[G_t \mid s_t = s,\, a_t = a]\,.$$

Exercise 1.2.

(a) Express $V_\pi(s)$ in terms of Q_π.

(b) Express $Q_\pi(s, a)$ in terms of V_π.

(c) Using these relations, derive a recursive equation for Q_π analogous to the Bellman equation above.

Solution to Exercise 1.2

(a) Starting from the definition of $V_\pi(s)$ and conditioning on the first action $a_t = a$:

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi[G_t \mid s_t = s] \\ &= \sum_{a \in \mathcal{A}} \pi(a \mid s)\, \mathbb{E}_\pi[G_t \mid s_t = s,\, a_t = a] \\ &= \sum_{a \in \mathcal{A}} \pi(a \mid s)\, Q_\pi(s, a). \end{aligned}$$

(b) Starting from the definition of $Q_\pi(s, a)$ and using the recursion $G_t = r_{t+1} +$

$\gamma\, G_{t+1}$, we condition on the first transition $s_{t+1} = s'$:

$$\begin{aligned} Q_\pi(s,a) &= \mathbb{E}_\pi[G_t \mid s_t = s,\, a_t = a] \\ &= \sum_{s' \in \mathcal{S}} T(s' \mid s,a)\, \mathbb{E}_\pi\big[r_{t+1} + \gamma\, G_{t+1} \;\big|\; s_t = s,\, a_t = a,\, s_{t+1} = s'\big] \\ &= \sum_{s' \in \mathcal{S}} T(s' \mid s,a)\big(R(s,a,s') + \gamma\, V_\pi(s')\big), \end{aligned}$$

where the last step uses $r_{t+1} = R(s,a,s')$ (deterministic given s_t, a_t, s_{t+1}) and the Markov property $\mathbb{E}_\pi[G_{t+1} \mid s_{t+1} = s'] = V_\pi(s')$.

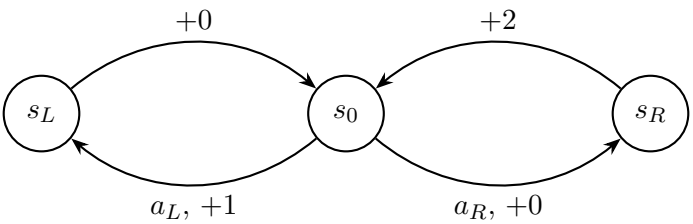
(c) Substituting Exercise 1.2(a) into Exercise 1.2(b), replacing $V_\pi(s')$ by its expression in terms of Q_π:

$$Q_\pi(s,a) = \sum_{s' \in \mathcal{S}} T(s' \mid s,a) \left(R(s,a,s') + \gamma \sum_{a' \in \mathcal{A}} \pi(a' \mid s')\, Q_\pi(s',a') \right).$$

This is the Bellman equation for the action-value function. Note also that substituting Exercise 1.2(b) into Exercise 1.2(a) recovers the Bellman equation for V_π from Exercise 1.1.

2 An example MDP

Consider the MDP with state space $\mathcal{S} = \{s_0, s_L, s_R\}$, action space $\mathcal{A} = \{a_L, a_R\}$, and deterministic transitions as shown:



From s_0, action a_L leads deterministically to s_L with reward $+1$, and action a_R leads deterministically to s_R with reward $+0$. From s_L (resp. s_R), any action returns to s_0 with reward $+0$ (resp. $+2$).

Let π_L denote the policy that always selects a_L, and π_R the policy that always selects a_R.

Exercise 2.1. Using the Bellman equation from Exercise 1.1, verify that

$$V_{\pi_L}(s_0) = \frac{1}{1-\gamma^2}, \qquad V_{\pi_R}(s_0) = \frac{2\gamma}{1-\gamma^2}.$$

Solution to Exercise 2.1

Under π_L all transitions are deterministic, so the Bellman equation collapses to

$$V_{\pi_L}(s_0) = R(s_0, a_L, s_L) + \gamma\, V_{\pi_L}(s_L) = 1 + \gamma\, V_{\pi_L}(s_L),$$
$$V_{\pi_L}(s_L) = R(s_L, a_L, s_0) + \gamma\, V_{\pi_L}(s_0) = 0 + \gamma\, V_{\pi_L}(s_0).$$

Substituting the second into the first gives $V_{\pi_L}(s_0) = 1 + \gamma^2\, V_{\pi_L}(s_0)$, hence

$$V_{\pi_L}(s_0) = \frac{1}{1-\gamma^2}.$$

Similarly, under π_R:

$$V_{\pi_R}(s_0) = R(s_0, a_R, s_R) + \gamma\, V_{\pi_R}(s_R) = 0 + \gamma\, V_{\pi_R}(s_R),$$
$$V_{\pi_R}(s_R) = R(s_R, a_R, s_0) + \gamma\, V_{\pi_R}(s_0) = 2 + \gamma\, V_{\pi_R}(s_0).$$

Substituting yields $V_{\pi_R}(s_0) = \gamma\big(2 + \gamma\, V_{\pi_R}(s_0)\big) = 2\gamma + \gamma^2\, V_{\pi_R}(s_0)$, hence

$$V_{\pi_R}(s_0) = \frac{2\gamma}{1-\gamma^2}.$$

Exercise 2.2. Determine the best action from state s_0 as a function of the discount factor $\gamma \in (0,1)$. Show that the breakeven point is $\gamma = \frac{1}{2}$.

Solution to Exercise 2.2

Since any action from s_L or s_R deterministically returns to s_0 with the same reward, the value of any policy at s_0 is determined entirely by the action chosen at s_0. Hence comparing $V_{\pi_L}(s_0)$ and $V_{\pi_R}(s_0)$ suffices:

$$V_{\pi_L}(s_0) - V_{\pi_R}(s_0) \;=\; \frac{1-2\gamma}{1-\gamma^2}.$$

Since $1-\gamma^2 > 0$ for $\gamma \in (0,1)$, the sign is determined by $1-2\gamma$:

- $\gamma < \frac{1}{2}$: a_L is optimal — the immediate reward of 1 outweighs a discounted future reward of 2.
- $\gamma > \frac{1}{2}$: a_R is optimal — the agent is patient enough to wait one step for the larger reward.
- $\gamma = \frac{1}{2}$: both actions are equally good (breakeven), and each gives $V^*(s_0) = \frac{4}{3}$.

3 Bellman operators and the existence of optimal policies

Banach fixed-point theorem

In this and the following exercises, we will make use of the well-known Banach fixed point theorem:

Theorem 3.1 (Banach Fixed Point Theorem)**.** *Let (X, d) be a complete metric space and let $F : X \to X$ be a contraction mapping, i.e. there exists $\gamma \in [0, 1)$ such that*

$$d(F(x), F(y)) \leq \gamma\, d(x, y) \quad \text{for all } x, y \in X.$$

Then F has a unique fixed point $x^* \in X$ satisfying $F(x^*) = x^*$. Moreover, for any initial point $x_0 \in X$, one has $\lim_{n\to\infty} d(F^n(x_0), x^*) = 0$ where F^n is the n-fold composition of F with itself. That is, $F^n(x_0)$ converges to x^* with respect to the metric d.

The problem

Definition 3.2 (Bellman optimality operator)**.** Let $\mathcal{V} = \{V : \mathcal{S} \to \mathbb{R}\}$ denote the vector space of value functions, equipped with the sup-norm $\|V\|_\infty = \max_{s\in\mathcal{S}} |V(s)|$. The **Bellman optimality operator** $\mathcal{B} : \mathcal{V} \to \mathcal{V}$ is defined by

$$(\mathcal{B}V)(s) \coloneqq \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma\, V(s')\big).$$

Exercise 3.1. Let $f, g : \mathcal{A} \to \mathbb{R}$ be real-valued functions on a finite set $\mathcal{A}$. Prove that

$$\left|\max_{a\in\mathcal{A}} f(a) - \max_{a\in\mathcal{A}} g(a)\right| \leq \max_{a\in\mathcal{A}} |f(a) - g(a)|.$$

Solution to Exercise 3.1

Without loss of generality, assume $\max_a f(a) \geq \max_a g(a)$. Let $a^* \in \arg\max_{a\in\mathcal{A}} f(a)$. Then:

$$\begin{aligned}
\left|\max_{a\in\mathcal{A}} f(a) - \max_{a\in\mathcal{A}} g(a)\right| &= \max_{a\in\mathcal{A}} f(a) - \max_{a\in\mathcal{A}} g(a) \\
&= f(a^*) - \max_{a\in\mathcal{A}} g(a) \\
&\leq f(a^*) - g(a^*) \\
&= |f(a^*) - g(a^*)| \\
&\leq \max_{a\in\mathcal{A}} |f(a) - g(a)|.
\end{aligned}$$

Exercise 3.2. Using Exercise 3.1, prove that $\mathcal{B}$ is a **contraction mapping** with contraction factor γ, i.e. for all $V, W \in \mathcal{V}$,

$$\|\mathcal{B}V - \mathcal{B}W\|_\infty \leq \gamma \, \|V - W\|_\infty.$$

Solution to Exercise 3.2

Fix an arbitrary state $s \in \mathcal{S}$. Define for each $a \in \mathcal{A}$:

$$f(a) \coloneqq \sum_{s' \in \mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma \, V(s')\big),$$
$$g(a) \coloneqq \sum_{s' \in \mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma \, W(s')\big).$$

Then $(\mathcal{B}V)(s) = \max_a f(a)$ and $(\mathcal{B}W)(s) = \max_a g(a)$. By Exercise 3.1:

$$\begin{aligned}
\big|(\mathcal{B}V)(s) - (\mathcal{B}W)(s)\big| &\leq \max_{a \in \mathcal{A}} \big|f(a) - g(a)\big| \\
&= \max_{a \in \mathcal{A}} \left| \sum_{s' \in \mathcal{S}} T(s' \mid s, a) \gamma \left(V(s') - W(s')\right) \right| \\
&\leq \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s' \mid s, a) \left|V(s') - W(s')\right| \\
&\leq \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s' \mid s, a) \, \|V - W\|_\infty \\
&\leq \gamma \, \|V - W\|_\infty,
\end{aligned}$$

where the last step uses $\sum_{s'} T(s' \mid s, a) = 1$. Since this holds for every $s \in \mathcal{S}$, taking the maximum over s gives $\|\mathcal{B}V - \mathcal{B}W\|_\infty \leq \gamma \, \|V - W\|_\infty$.

Exercise 3.3. Using Theorem 3.1, conclude that there exists a unique $V^* \in \mathcal{V}$ satisfying $\mathcal{B}V^* = V^*$, and that for any initial $V_0 \in \mathcal{V}$, the iterates $\mathcal{B}^n V_0 \to V^*$ as $n \to \infty$.

Solution to Exercise 3.3

The space $(\mathcal{V}, \|\cdot\|_\infty)$ is a finite-dimensional normed vector space, hence complete. By Exercise 3.2, $\mathcal{B}$ is a contraction on $\mathcal{V}$ with factor $\gamma < 1$. Theorem 3.1 then gives the existence of a unique fixed point $V^* = \mathcal{B}V^*$, and convergence $\mathcal{B}^n V_0 \to V^*$ for any $V_0 \in \mathcal{V}$.

Exercise 3.4. Show that for any policy π, the operator $\mathcal{B}_\pi : \mathcal{V} \to \mathcal{V}$ defined by

Note

Bellman Policy Operator

$$(\mathcal{B}_\pi V)(s) := \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} T(s' \mid s, a)\big(R(s,a,s') + \gamma\, V(s')\big)$$

is also a contraction with factor γ. Conclude that V_π is its unique fixed point and that $\mathcal{B}_\pi^n V_0 \to V_\pi$ for any initial $V_0 \in \mathcal{V}$ as $n \to \infty$.
Remark. This shows that the numerical policy evaluation algorithm from the ARENA materials converges to the correct solution.

Solution to Exercise 3.4

Fix $V, W \in \mathcal{V}$ and an arbitrary state $s \in \mathcal{S}$. Then:

$$\begin{aligned}
\big|(\mathcal{B}_\pi V)(s) - (\mathcal{B}_\pi W)(s)\big| &= \left|\sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} T(s' \mid s, a)\,\gamma\big(V(s') - W(s')\big)\right| \\
&\le \gamma \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} T(s' \mid s, a)\,\big|V(s') - W(s')\big| \\
&\le \gamma \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} T(s' \mid s, a)\,\|V - W\|_\infty \\
&\le \gamma\,\|V - W\|_\infty,
\end{aligned}$$

where the last step uses $\sum_{s'} T(s' \mid s, a) = 1$ and $\sum_a \pi(a \mid s) = 1$. Taking the maximum over s gives $\|\mathcal{B}_\pi V - \mathcal{B}_\pi W\|_\infty \le \gamma\,\|V - W\|_\infty$, so $\mathcal{B}_\pi$ is a contraction with factor γ.
By Theorem 3.1, $\mathcal{B}_\pi$ has a unique fixed point. By the Bellman equation (Section 1), V_π satisfies $\mathcal{B}_\pi V_\pi = V_\pi$, so V_π is this unique fixed point. The convergence statement also follows from Theorem 3.1.

Exercise 3.5. Let $V \in \mathcal{V}$ be any value function, and let π_V be a greedy policy with respect to V, i.e.

$$\pi_V(s) \in \arg\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} T(s' \mid s, a)\big(R(s,a,s') + \gamma\, V(s')\big).$$

Show that $\mathcal{B}_{\pi_V} V = \mathcal{B} V$.

Solution to Exercise 3.5

Since π_V is deterministic, $\pi_V(a \mid s) = [\![a = \pi_V(s)]\!]$, so the sum over a in $\mathcal{B}_{\pi_V}$ collapses to the single term $a = \pi_V(s)$. For every $s \in \mathcal{S}$:

$$\begin{aligned}(\mathcal{B}_{\pi_V} V)(s) &= \sum_{s' \in \mathcal{S}} T(s' \mid s, \pi_V(s))\big(R(s, \pi_V(s), s') + \gamma\, V(s')\big) \\ &= \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma\, V(s')\big) \\ &= (\mathcal{B}V)(s),\end{aligned}$$

where the second equality holds because $\pi_V(s)$ selects a maximizing action by definition.

Exercise 3.6. Define the **greedy policy** with respect to V^* by $\pi^* = \pi_{V^*}$ as in the previous part. Show that $V_{\pi^*} = V^*$, i.e. the value function of π^* equals the fixed point of $\mathcal{B}$.
Hint: show that V^* satisfies the same fixed-point equation as V_{π^*}.

Solution to Exercise 3.6

We have $\mathcal{B}_{\pi^*} V^* = \mathcal{B}V^* = V^*$ by Exercise 3.5 and Exercise 3.3. So V^* is a fixed point of $\mathcal{B}_{\pi^*}$. By Exercise 3.4, V_{π^*} is the unique fixed point of $\mathcal{B}_{\pi^*}$. Therefore $V_{\pi^*} = V^*$.

Exercise 3.7. Conclude that π^* is an **optimal policy**: for every policy π and every state $s \in \mathcal{S}$,

$$V_{\pi^*}(s) \geq V_\pi(s).$$

Hint: show that $\mathcal{B}V_\pi \geq V_\pi$ pointwise, then iterate and take the limit.

Solution to Exercise 3.7

Let π be any policy. Since the max over actions is at least as large as any weighted average:

$$(\mathcal{B}V)(s) \geq (\mathcal{B}_\pi V)(s) \quad \text{for all } V \in \mathcal{V},\ s \in \mathcal{S}.$$

In particular, since V_π is a fixed point of $\mathcal{B}_\pi$ (Exercise 3.4):

$$(\mathcal{B}V_\pi)(s) \geq (\mathcal{B}_\pi V_\pi)(s) = V_\pi(s) \quad \text{for all } s \in \mathcal{S}.$$

Note that $\mathcal{B}$ is monotone: if $V(s) \geq W(s)$ for all s, then $(\mathcal{B}V)(s) \geq (\mathcal{B}W)(s)$. Applying $\mathcal{B}$ to both sides and using monotonicity, then iterating:

$$V_\pi \leq \mathcal{B}V_\pi \leq \mathcal{B}^2 V_\pi \leq \cdots \leq \mathcal{B}^n V_\pi.$$

By Exercise 3.3, $\mathcal{B}^n V_\pi \to V^*$ as $n \to \infty$. Taking the limit:

$$V_\pi(s) \le V^*(s) = V_{\pi^*}(s) \quad \text{for all } s \in \mathcal{S},$$

where the last equality is Exercise 3.6. Since π was arbitrary, π^* is optimal.

4 Policy improvement theorem

Given a policy π, define the **improved policy** π' greedily with respect to V_π:

Note

Policy Improvement

$$\pi'(s) \in \arg\max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma\, V_\pi(s')\big).$$

Exercise 4.1. In proving Exercise 3.7, we already saw that for all $s \in \mathcal{S}$, we have $(\mathcal{B}V_\pi)(s) \ge V_\pi(s)$. When does equality hold for all states?

Solution to Exercise 4.1

We have

$$\begin{aligned}(\mathcal{B}V_\pi)(s) &= \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma\, V_\pi(s')\big) \\ &\ge \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} T(s' \mid s, a)\big(R(s, a, s') + \gamma\, V_\pi(s')\big) \\ &= V_\pi(s),\end{aligned}$$

where we used the Bellman equation in the last step. If equality holds at every state s, then V_π is a fixed point of $\mathcal{B}$, which by uniqueness (Exercise 3.3) means $V_\pi = V^* = V_{\pi^*}$, i.e. π is already optimal by Exercise 3.7.

Exercise 4.2. Show that $V_{\pi'}(s) \ge V_\pi(s)$ for all $s \in \mathcal{S}$.
Hint: show that $\mathcal{B}_{\pi'} V_\pi \ge V_\pi$ pointwise using Exercise 3.5, then iterate $\mathcal{B}_{\pi'}$ and take the limit.

Solution to Exercise 4.2

By Exercise 3.5, $\mathcal{B}_{\pi'} V_\pi = \mathcal{B} V_\pi \ge V_\pi$ pointwise, where the inequality is from Exercise 4.1. Note that $\mathcal{B}_{\pi'}$ is monotone: if $V(s) \ge W(s)$ for all s, then

$(\mathcal{B}_{\pi'}V)(s) \geq (\mathcal{B}_{\pi'}W)(s)$. Applying $\mathcal{B}_{\pi'}$ to both sides and iterating:

$$V_\pi \leq \mathcal{B}_{\pi'}V_\pi \leq \mathcal{B}_{\pi'}^2 V_\pi \leq \cdots \leq \mathcal{B}_{\pi'}^n V_\pi.$$

By Exercise 3.4, $\mathcal{B}_{\pi'}^n V_\pi \to V_{\pi'}$ as $n \to \infty$. Taking the limit gives $V_{\pi'}(s) \geq V_\pi(s)$ for all $s \in \mathcal{S}$.

Exercise 4.3. The **policy iteration** algorithm generates a sequence of policies $\pi_0, \pi_1, \pi_2, \ldots$ where each π_{k+1} is the greedy policy with respect to V_{π_k}. Show that this sequence converges to an optimal policy π^* in a finite number of steps.
Hint: how many deterministic policies are there?

Solution to Exercise 4.3

By Exercise 4.2, the sequence of value functions is monotonically improving: $V_{\pi_{k+1}}(s) \geq V_{\pi_k}(s)$ for all s and all k. Since each π_k is a deterministic policy and there are only $|\mathcal{A}|^{|\mathcal{S}|}$ deterministic policies, the sequence must eventually revisit a policy. If $\pi_{k+1} = \pi_j$ for some $j \leq k$, then $V_{\pi_{k+1}} = V_{\pi_j}$, and by monotonicity $V_{\pi_j} = V_{\pi_{j+1}} = \cdots = V_{\pi_{k+1}}$. In particular $V_{\pi_k} = V_{\pi_{k+1}}$.
It remains to show this implies optimality. Since π_{k+1} is greedy with respect to V_{π_k}, Exercise 3.5 gives $\mathcal{B}V_{\pi_k} = \mathcal{B}_{\pi_{k+1}}V_{\pi_k}$. Since $V_{\pi_{k+1}}$ is the fixed point of $\mathcal{B}_{\pi_{k+1}}$ and $V_{\pi_k} = V_{\pi_{k+1}}$, we get $\mathcal{B}_{\pi_{k+1}}V_{\pi_k} = V_{\pi_k}$. Combining: $\mathcal{B}V_{\pi_k} = V_{\pi_k}$, so π_k is optimal by Exercise 4.1.

5 Bellman convergence rate

Exercise 3.3 showed that the iterates $V_n := \mathcal{B}^n V_0$ converge to V^* for any starting $V_0 \in \mathcal{V}$, but it did not tell us *how fast*, nor how to decide when to stop iterating. We derive some convergence rate bounds [Put94, Chapter 6].

Exercise 5.1. Show that for any initial $V_0 \in \mathcal{V}$, the iterates $V_n := \mathcal{B}^n V_0$ satisfy

$$\|V_n - V^*\|_\infty \;\leq\; \gamma^n \|V_0 - V^*\|_\infty.$$

Remark. This bound is not useful as a stopping criterion in practice as it depends on the unknown V^*.

Solution to Exercise 5.1

Since $V^* = \mathcal{B}V^*$ and $\mathcal{B}$ is a γ-contraction (Exercise 3.2),

$$\|V_n - V^*\|_\infty \;=\; \|\mathcal{B}V_{n-1} - \mathcal{B}V^*\|_\infty \;\leq\; \gamma\, \|V_{n-1} - V^*\|_\infty.$$

Iterating n times gives $\|V_n - V^*\|_\infty \leq \gamma^n \|V_0 - V^*\|_\infty$.

Exercise 5.2. Show the **a priori bound**:

$$\|V_n - V^*\|_\infty \;\le\; \frac{\gamma^n}{1-\gamma}\,\|V_1 - V_0\|_\infty.$$

This bound uses only $\|V_1 - V_0\|_\infty$ — a quantity computable after a single iteration, independent of V^*.
Hint: write $V^* - V_n = \sum_{k=n}^{\infty}(V_{k+1} - V_k)$ (valid since $V_k \to V^*$), and bound each term using the contraction of $\mathcal{B}$.

Solution to Exercise 5.2

By the contraction property applied repeatedly,

$$\begin{aligned}\|V_{k+1} - V_k\|_\infty \;&=\; \|\mathcal{B}V_k - \mathcal{B}V_{k-1}\|_\infty \;\le\; \gamma\,\|V_k - V_{k-1}\|_\infty \\ &\le\; \cdots \;\le\; \gamma^k\,\|V_1 - V_0\|_\infty.\end{aligned}$$

Since $V_k \to V^*$ (Exercise 3.3), we have the telescoping identity $V^* - V_n = \sum_{k=n}^{\infty}(V_{k+1} - V_k)$. Applying the triangle inequality and the geometric bound above:

$$\begin{aligned}\|V^* - V_n\|_\infty \;&\le\; \sum_{k=n}^{\infty}\|V_{k+1} - V_k\|_\infty \;\le\; \sum_{k=n}^{\infty}\gamma^k\,\|V_1 - V_0\|_\infty \\ &=\; \frac{\gamma^n}{1-\gamma}\,\|V_1 - V_0\|_\infty.\end{aligned}$$

Exercise 5.3. Show the **a posteriori bound**: for $n \ge 1$,

$$\|V_n - V^*\|_\infty \;\le\; \frac{\gamma}{1-\gamma}\,\|V_n - V_{n-1}\|_\infty.$$

Hint: telescope from n as in Exercise 5.2, but bound $\|V_{k+1} - V_k\|_\infty$ in terms of $\|V_n - V_{n-1}\|_\infty$ instead.

Solution to Exercise 5.3

Step 1: bound each successive difference in terms of $\|V_n - V_{n-1}\|_\infty$. Apply the contraction property of $\mathcal{B}$ once:

$$\|V_{n+1} - V_n\|_\infty \;=\; \|\mathcal{B}V_n - \mathcal{B}V_{n-1}\|_\infty \;\le\; \gamma\,\|V_n - V_{n-1}\|_\infty.$$

Apply it again, chaining with the previous bound:

$$\|V_{n+2} - V_{n+1}\|_\infty \;\le\; \gamma\,\|V_{n+1} - V_n\|_\infty \;\le\; \gamma^2\,\|V_n - V_{n-1}\|_\infty.$$

Continuing this pattern, for every $j \geq 0$:

$$\|V_{n+j+1} - V_{n+j}\|_\infty \;\leq\; \gamma^{j+1}\,\|V_n - V_{n-1}\|_\infty.$$

Step 2: telescope. Since $V_k \to V^*$ (Exercise 3.3), we have the telescoping identity

$$V^* - V_n \;=\; (V_{n+1} - V_n) + (V_{n+2} - V_{n+1}) + (V_{n+3} - V_{n+2}) + \cdots.$$

Applying the triangle inequality and the bounds from Step 1:

$$\begin{aligned}
\|V^* - V_n\|_\infty &\leq \|V_{n+1} - V_n\|_\infty \;+\; \|V_{n+2} - V_{n+1}\|_\infty \\
&\qquad +\; \|V_{n+3} - V_{n+2}\|_\infty \;+\; \cdots \\
&\leq \gamma\,\|V_n - V_{n-1}\|_\infty \;+\; \gamma^2\,\|V_n - V_{n-1}\|_\infty \\
&\qquad +\; \gamma^3\,\|V_n - V_{n-1}\|_\infty \;+\; \cdots \\
&= \left(\gamma + \gamma^2 + \gamma^3 + \cdots\right)\|V_n - V_{n-1}\|_\infty \\
&= \frac{\gamma}{1-\gamma}\,\|V_n - V_{n-1}\|_\infty.
\end{aligned}$$

Exercise 5.4. Show that the a posteriori bound of Exercise 5.3 is always at least as tight as the a priori bound of Exercise 5.2: for all $n \geq 1$,

$$\frac{\gamma}{1-\gamma}\,\|V_n - V_{n-1}\|_\infty \;\leq\; \frac{\gamma^n}{1-\gamma}\,\|V_1 - V_0\|_\infty.$$

Hint: bound $\|V_n - V_{n-1}\|_\infty$ by repeated application of the contraction.

Solution to Exercise 5.4

Apply the contraction property of $\mathcal{B}$ to $V_n = \mathcal{B}V_{n-1}$ and $V_{n-1} = \mathcal{B}V_{n-2}$:

$$\|V_n - V_{n-1}\|_\infty \;\leq\; \gamma\,\|V_{n-1} - V_{n-2}\|_\infty.$$

Continuing this $n-1$ times altogether, we eventually land on $\|V_1 - V_0\|_\infty$:

$$\|V_n - V_{n-1}\|_\infty \;\leq\; \gamma^{n-1}\,\|V_1 - V_0\|_\infty.$$

Multiplying both sides by $\dfrac{\gamma}{1-\gamma} > 0$ preserves the inequality:

$$\frac{\gamma}{1-\gamma}\,\|V_n - V_{n-1}\|_\infty \;\leq\; \frac{\gamma^n}{1-\gamma}\,\|V_1 - V_0\|_\infty.$$

Both bounds from Exercise 5.3 and Exercise 5.2 upper-bound $\|V_n - V^*\|_\infty$, and the a posteriori one is smaller.

Exercise 5.5. Assume $|R(s,a,s')| \le R_{\max}$ for all s, a, s'. Show that if we initialize with $V_0 \equiv 0$, then

$$\|V_n - V^*\|_\infty \;\le\; \frac{\gamma^n}{1-\gamma}\, R_{\max}.$$

This bound depends only on n, γ, and $R_{\max}$, so it can be evaluated before running a single iteration.

Solution to Exercise 5.5

With $V_0 \equiv 0$:

$$V_1(s) \;=\; (\mathcal{B}V_0)(s) \;=\; \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s' \mid s,a)\, R(s,a,s').$$

Each term is a convex combination of rewards, all bounded in absolute value by $R_{\max}$, so $|V_1(s)| \le R_{\max}$ for every s, hence $\|V_1 - V_0\|_\infty = \|V_1\|_\infty \le R_{\max}$. Substituting into the a priori bound of Exercise 5.2:

$$\|V_n - V^*\|_\infty \;\le\; \frac{\gamma^n}{1-\gamma}\, \|V_1 - V_0\|_\infty \;\le\; \frac{\gamma^n}{1-\gamma}\, R_{\max}.$$

Exercise 5.6. Show that the bound of Exercise 5.5 is **tight**: exhibit an MDP in which $\|V_n - V^*\|_\infty = \frac{\gamma^n}{1-\gamma} R_{\max}$ for every $n \ge 0$ (with $V_0 \equiv 0$).

Solution to Exercise 5.6

Take $\mathcal{S} = \{s\}$, $\mathcal{A} = \{a\}$, $T(s \mid s,a) = 1$, and $R(s,a,s) = R_{\max}$. The Bellman optimality operator collapses to

$$(\mathcal{B}V)(s) \;=\; R_{\max} + \gamma\, V(s).$$

Starting from $V_0(s) = 0$ and iterating,

$$V_n(s) \;=\; R_{\max}\bigl(1 + \gamma + \gamma^2 + \cdots + \gamma^{n-1}\bigr) \;=\; R_{\max} \cdot \frac{1-\gamma^n}{1-\gamma}.$$

The fixed point $V^*(s)$ solves $V^*(s) = R_{\max} + \gamma\, V^*(s)$, giving $V^*(s) = R_{\max}/(1-\gamma)$. Therefore

$$\|V_n - V^*\|_\infty \;=\; V^*(s) - V_n(s) \;=\; \frac{R_{\max}}{1-\gamma} - \frac{R_{\max}(1-\gamma^n)}{1-\gamma} \;=\; \frac{\gamma^n}{1-\gamma}\, R_{\max},$$

so the bound of Exercise 5.5 is attained with equality. No bound in terms of only n, γ, and $R_{\max}$ can be sharper.

Remark. Combining Exercises 5.2, 5.3 and 5.5 gives a chain of progressively weaker but increasingly upfront-computable bounds:

Note

Convergence Bounds

$$\|V_n - V^*\|_\infty \;\le\; \underbrace{\tfrac{\gamma}{1-\gamma}\,\|V_n - V_{n-1}\|_\infty}_{\text{a posteriori, tightest}} \;\le\; \underbrace{\tfrac{\gamma^n}{1-\gamma}\,\|V_1 - V_0\|_\infty}_{\text{a priori}}$$
$$\le\; \underbrace{\tfrac{\gamma^n}{1-\gamma}\,R_{\max}}_{\text{upfront, assumes } V_0\equiv 0,\,|R|\le R_{\max}} .$$

Each step weakens the bound by replacing realized information with something coarser: first the most recent step size $\|V_n - V_{n-1}\|_\infty$ with the worst-case geometric decay $\gamma^{n-1}\|V_1 - V_0\|_\infty$ (contraction of $\mathcal{B}$ applied $n-1$ times), then the initial step size $\|V_1 - V_0\|_\infty$ with the crude reward bound $R_{\max}$. In practice the leftmost bound is significantly tighter, since $\|V_n - V_{n-1}\|_\infty$ often decays strictly faster than $\gamma^{n-1}\|V_1 - V_0\|_\infty$. It also yields a concrete stopping rule: to guarantee $\|V_n - V^*\|_\infty \le \varepsilon$, iterate until $\|V_n - V_{n-1}\|_\infty \le \frac{1-\gamma}{\gamma}\,\varepsilon$.

6 Convergence of Q-learning

The policy improvement theorem from Section 4 shows that we can in principle find an optimal policy in MDPs. However, it makes the uncomfortable assumption that we know the environment dynamics via the transition kernel T. In this problem, we prove that Q-learning, which does not make such an assumption, converges to the optimal action-value function, from which one can trivially extract an optimal policy by choosing actions greedily.

Stochastic approximation theorem

We will make use of the following stochastic approximation result, which we state without proof:

Theorem 6.1 (Stochastic Approximation for Contractions [Tsi94, Theorem 3])**.** *Let $\mathcal{X} = \{x : I \to \mathbb{R}\}$ be the space of real-valued functions on a finite set I, equipped with the sup-norm $\|x\|_\infty = \max_i |x(i)|$. Let $F : \mathcal{X} \to \mathcal{X}$ be a contraction mapping with factor $\gamma \in [0,1)$ and unique fixed point x^*.*
*Fix an initial iterate $x_0 \in \mathcal{X}$, and for each $t \ge 0$ let $\alpha_t : I \to [0,1]$ be a **learning rate** function and $w_t : I \to \mathbb{R}$ a random **noise** function (both functions on I, one value per coordinate). Consider the stochastic iteration*

$$x_{t+1}(i) = (1 - \alpha_t(i))\,x_t(i) + \alpha_t(i)\big[F(x_t)(i) + w_t(i)\big] \qquad \text{for all } i \in I,$$

where updates are applied to a single coordinate $i = i_t$ at each step (i.e. $\alpha_t(i) = 0$ for $i \neq i_t$). Suppose:

(a) ***(Learning rate)*** *For each $i \in I$:*

$$\sum_{t=0}^{\infty} \alpha_t(i) = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t(i)^2 < \infty, \qquad \alpha_t(i) \in [0, 1].$$

(b) ***(Noise)*** *For each $i \in I$, conditioned on the history $\mathcal{F}_t = \{x_0, \alpha_0, w_0, \alpha_1, w_1, \ldots, \alpha_{t-1}, w_{t-1}, x_t, \alpha_t\}$, the noise has zero mean and bounded variance:*

$$\mathbb{E}\big[w_t(i) \mid \mathcal{F}_t\big] = 0, \qquad \mathbb{E}\big[w_t(i)^2 \mid \mathcal{F}_t\big] \leq C\big(1 + \|x_t\|_\infty^2\big)$$

for some constant $C > 0$.

Then $x_t(i) \to x^*(i)$ for all $i \in I$, with probability 1.

What is stochastic? The randomness of the iteration is carried by the noise: each $w_t(i)$ is a real-valued random variable, and w_t is a random element of $\mathbb{R}^I$. Consequently, the iterates $x_1, x_2, \ldots$ are random (they depend on past noise $w_0, \ldots, w_{t-1}$), and the learning rates $\alpha_t(i)$ may be random as well if they depend on the history—e.g. on which coordinate was just visited. The contraction F, its fixed point x^*, the update rule itself, and the initial iterate x_0 are all deterministic. The theorem asserts that despite the noise, the random sequence (x_t) converges pointwise to x^* almost surely.

Remark on "with probability 1". The conclusion $x_t(i) \to x^*(i)$ with probability 1 (also called **almost sure convergence**) means

$$P\Big(\lim_{t\to\infty} x_t(i) = x^*(i)\Big) = 1 \quad \text{for all } i \in I.$$

In other words, the probability that a run of the stochastic iteration fails to converge to x^* is zero.

Why should we believe this? It helps to build up the result in three layers of increasing complexity.

Layer 1: no noise, all coordinates at once. If $w_t = 0$ and $\alpha_t(i) = 1$ for all i and t, the iteration reduces to $x_{t+1} = F(x_t)$. This converges to x^* by Theorem 3.1.

Layer 2: noise, but all coordinates at once. Now suppose we observe $F(x_t)+w_t$ instead of $F(x_t)$, and use a decreasing step size:

$$x_{t+1} = (1 - \alpha_t)\, x_t + \alpha_t\big[F(x_t) + w_t\big] = x_t + \alpha_t\big[\underbrace{F(x_t) - x_t}_{\text{signal}} + \underbrace{w_t}_{\text{noise}}\big].$$

The signal $F(x_t) - x_t$ always points toward x^* (this is what the contraction property buys us). The noise w_t is mean-zero, so it pushes us in random directions that partially cancel over time. The two conditions on the step size mediate between signal and noise:

- $\sum \alpha_t = \infty$ ensures the total step size is large enough to reach x^* from any starting point.
- $\sum \alpha_t^2 < \infty$ ensures the noise averages out. Each noise term w_t enters the iteration scaled by α_t, contributing a random displacement of size $\alpha_t w_t$. Although these displacements are not independent (each depends on the current iterate x_t), they are mean-zero conditioned on the past. For such sequences, the variance of the cumulative sum is the sum of the individual variances, which is proportional to $\sum \alpha_t^2$. Since this is finite, the total random drift converges and cannot overwhelm the steady pull of the signal toward x^*.

The signal accumulates coherently (always pulling toward x^*), while the noise averages out.

Layer 3: one coordinate at a time. The iteration updates only one coordinate $i = i_t$ per step, while the others stay frozen. This makes the analysis more difficult since the target $F(x_t)(i)$ depends on all coordinates, including stale ones. The condition $\sum_t \alpha_t(i) = \infty$ for each i ensures no coordinate is permanently neglected, and the contraction property of F provides enough global coupling to prevent coordinates from drifting apart. Making this rigorous is the main technical content of the proof.

The problem

Consider the following generalization of the Q-learning update from the ARENA materials, now with a time-varying learning rate $\alpha_t > 0$: given a current estimate $Q_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and an observed transition $(s_t, a_t, r_{t+1}, s_{t+1})$ where $s_{t+1} \sim T(\cdot \mid s_t, a_t)$ and $r_{t+1} = R(s_t, a_t, s_{t+1})$, the update is

Note

Q-Learning Update

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t\big(r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(s_{t+1}, a') - Q_t(s_t, a_t)\big),$$

with $Q_{t+1}(s, a) = Q_t(s, a)$ for all $(s, a) \neq (s_t, a_t)$. We will show that, under appropriate conditions, Q-learning converges to the optimal action-value function Q^*.

Let $\mathcal{Q} = \{Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ denote the space of action-value functions, equipped with the sup-norm $\|Q\|_\infty = \max_{s,a} |Q(s, a)|$.

Exercise 6.1. Overloading the notation from Definition 3.2, define the **Bellman optimality operator** on action-value functions, $\mathcal{B} : \mathcal{Q} \to \mathcal{Q}$, by

Note

Bellman Q-Operator

$$(\mathcal{B}Q)(s,a) \coloneqq \sum_{s' \in \mathcal{S}} T(s' \mid s,a)\big[R(s,a,s') + \gamma \max_{a' \in \mathcal{A}} Q(s',a')\big].$$

(Here $\mathcal{B}$ takes a Q-function to a Q-function, whereas the earlier $\mathcal{B}$ in Definition 3.2 takes a V-function to a V-function — whether $\mathcal{B}$ denotes the V- or Q-version is always clear from its argument.) Show that $\mathcal{B}$ is a γ-contraction on $(\mathcal{Q}, \|\cdot\|_\infty)$. *Hint:* this is similar to the proof of Exercise 3.2.

Solution to Exercise 6.1

Fix $Q, Q' \in \mathcal{Q}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. By the inequality from Exercise 3.1 (applied at each s') and the triangle inequality:

$$\begin{aligned}
&\big|(\mathcal{B}Q)(s,a) - (\mathcal{B}Q')(s,a)\big| \\
&\quad= \left|\sum_{s'} T(s' \mid s,a)\,\gamma\Big(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a')\Big)\right| \\
&\quad\le \sum_{s'} T(s' \mid s,a)\,\gamma \max_{a'}\big|Q(s',a') - Q'(s',a')\big| \\
&\quad\le \sum_{s'} T(s' \mid s,a)\,\gamma \max_{a',s''}\big|Q(s'',a') - Q'(s'',a')\big| \\
&\quad\le \gamma\,\|Q - Q'\|_\infty.
\end{aligned}$$

Taking the maximum over (s,a) gives $\|\mathcal{B}Q - \mathcal{B}Q'\|_\infty \le \gamma\,\|Q - Q'\|_\infty$.

Exercise 6.2. Define $Q^* \coloneqq Q_{\pi^*}$, the action-value function of the optimal policy π^* from Section 3. We now show Bellman optimality equations and their consequences.

(a) Show that $Q^*(s,a) = \sum_{s' \in \mathcal{S}} T(s' \mid s,a)\big[R(s,a,s') + \gamma\, V^*(s')\big]$.

(b) Show that $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a)$.

(c) Conclude that Q^* is the unique fixed point of $\mathcal{B}$.

Solution to Exercise 6.2

(a) By Exercise 1.2(b) applied to π^*:

$$Q_{\pi^*}(s,a) = \sum_{s' \in \mathcal{S}} T(s' \mid s,a)\big[R(s,a,s') + \gamma\, V_{\pi^*}(s')\big].$$

By Exercise 3.6, $V_{\pi^*} = V^*$. Substituting gives the result.

(b) By Exercise 1.2(a), $V_{\pi^*}(s) = \sum_a \pi^*(a \mid s)\, Q_{\pi^*}(s,a)$. Since π^* is the greedy policy with respect to V^* (Exercises 3.5 and 3.6), it is deterministic and selects $\pi^*(s) \in \arg\max_a \sum_{s'} T(s' \mid s,a)[R(s,a,s') + \gamma\, V^*(s')]$. By Exercise 6.2(a), the inner expression equals $Q^*(s,a)$, so $\pi^*(s) \in \arg\max_a Q^*(s,a)$. Since π^* is deterministic, $V^*(s) = V_{\pi^*}(s) = Q_{\pi^*}(s,\pi^*(s)) = Q^*(s,\pi^*(s)) = \max_a Q^*(s,a)$.

(c) Substituting Exercise 6.2(b) into Exercise 6.2(a):

$$Q^*(s,a) = \sum_{s'} T(s' \mid s,a)\big[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')\big] = (\mathcal{B}Q^*)(s,a).$$

So Q^* is a fixed point of $\mathcal{B}$. Uniqueness follows from Exercise 6.1 and Theorem 3.1.

Exercise 6.3. Rewrite the Q-learning update in the form of Theorem 6.1:

$$x_{t+1}(i) = (1 - \alpha_t(i))\, x_t(i) + \alpha_t(i)\big[F(x_t)(i) + w_t(i)\big].$$

Specifically, identify the index set I, the mapping F, the iterates x_t, and the learning rates $\alpha_t(i)$. Then give an explicit expression for the noise w_t.

Solution to Exercise 6.3

The identifications are:

- Index set: $I = \mathcal{S} \times \mathcal{A}$.
- Iterates: $x_t = Q_t$.
- Contraction: $F = \mathcal{B}$, with fixed point $x^* = Q^*$ (Exercise 6.2).
- Learning rates: $\alpha_t(s,a) = \alpha_t \cdot [\![(s,a) = (s_t,a_t)]\!]$.

With these identifications, the Q-learning update becomes

$$Q_{t+1}(s,a) = (1 - \alpha_t(s,a))\, Q_t(s,a) + \alpha_t(s,a)\big[(\mathcal{B}Q_t)(s,a) + w_t(s,a)\big],$$

where the noise at the updated coordinate is

$$w_t(s_t,a_t) = R(s_t,a_t,s_{t+1}) + \gamma \max_{a'} Q_t(s_{t+1},a') - (\mathcal{B}Q_t)(s_t,a_t),$$

i.e. the difference between the sampled target (using the single transition s_{t+1}) and its expectation over all possible transitions. For $(s,a) \neq (s_t,a_t)$, we set $w_t(s,a) = 0$.

Exercise 6.4. Show that the noise satisfies $\mathbb{E}[w_t(s,a) \mid \mathcal{F}_t] = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $\mathcal{F}_t$ denotes the history up to time t as in Theorem 6.1. *Hint:* conditioned on $\mathcal{F}_t$, the quantities Q_t, s_t, and a_t are all determined. What is the only remaining source of randomness?

Solution to Exercise 6.4

Conditioned on $\mathcal{F}_t$, the quantities Q_t, s_t, a_t are all determined. The only remaining randomness is in $s_{t+1} \sim T(\cdot \mid s_t, a_t)$. Therefore:

$$\begin{aligned}
&\mathbb{E}[w_t(s_t,a_t) \mid \mathcal{F}_t] \\
&\quad = \mathbb{E}\big[R(s_t,a_t,s_{t+1}) + \gamma \max_{a'} Q_t(s_{t+1},a') \mid \mathcal{F}_t\big] - (\mathcal{B}Q_t)(s_t,a_t) \\
&\quad = \sum_{s'} T(s' \mid s_t,a_t)\big[R(s_t,a_t,s') + \gamma \max_{a'} Q_t(s',a')\big] - (\mathcal{B}Q_t)(s_t,a_t) \\
&\quad = (\mathcal{B}Q_t)(s_t,a_t) - (\mathcal{B}Q_t)(s_t,a_t) = 0.
\end{aligned}$$

For $(s,a) \neq (s_t,a_t)$, we have $w_t(s,a) = 0$ by definition, so $\mathbb{E}[w_t(s,a) \mid \mathcal{F}_t] = 0$ trivially.

Exercise 6.5. Assume $|R(s,a,s')| \le R_{\max}$ for all s,a,s'. Show that there exists a constant $C > 0$ (depending only on $R_{\max}$ and γ) such that

$$\mathbb{E}\big[w_t(s,a)^2 \mid \mathcal{F}_t\big] \le C\big(1 + \|Q_t\|_\infty^2\big).$$

Hint: first show that $|w_t(s_t,a_t)| \le 2(R_{\max} + \gamma\,\|Q_t\|_\infty)$.

Solution to Exercise 6.5

Step 1: a pointwise bound on w_t. Using $|R(s_t,a_t,s_{t+1})| \le R_{\max}$ and $|\max_{a'} Q_t(s_{t+1},a')| \le \|Q_t\|_\infty$, the sampled target satisfies

$$\big|R(s_t,a_t,s_{t+1}) + \gamma \max_{a'} Q_t(s_{t+1},a')\big| \le R_{\max} + \gamma\,\|Q_t\|_\infty.$$

The same bound holds for $|(\mathcal{B}Q_t)(s_t,a_t)|$, since it is a convex combination (over s') of terms each bounded by $R_{\max} + \gamma\,\|Q_t\|_\infty$. By the triangle inequality:

$$|w_t(s_t,a_t)| \;\le\; 2\big(R_{\max} + \gamma\,\|Q_t\|_\infty\big).$$

This bound holds for *every* realization of s_{t+1}, not just in expectation.

Step 2: passing to the conditional second moment. Squaring the pointwise bound,

$$w_t(s_t, a_t)^2 \;\le\; 4\big(R_{\max} + \gamma \,\|Q_t\|_\infty\big)^2 \quad \text{a.s.}$$

Taking conditional expectation of both sides given $\mathcal{F}_t$ preserves the inequality (monotonicity of conditional expectation):

$$\mathbb{E}\big[w_t(s_t, a_t)^2 \mid \mathcal{F}_t\big] \;\le\; \mathbb{E}\big[\,4(R_{\max} + \gamma \,\|Q_t\|_\infty)^2 \mid \mathcal{F}_t\big]\,.$$

Now we simplify the right-hand side. It is a function of Q_t (and the constants $R_{\max}, \gamma$), and Q_t is determined once we condition on $\mathcal{F}_t$ (recall $\mathcal{F}_t$ records $Q_0, Q_1, \ldots, Q_t$). So the right-hand side is $\mathcal{F}_t$**-measurable**: given the history, it is a fixed real number, not random. Conditional expectation acts as the identity on $\mathcal{F}_t$-measurable quantities — i.e. $\mathbb{E}[Y \mid \mathcal{F}_t] = Y$ when Y is $\mathcal{F}_t$-measurable — so

$$\mathbb{E}\big[\,4(R_{\max} + \gamma \,\|Q_t\|_\infty)^2 \mid \mathcal{F}_t\big] \;=\; 4\big(R_{\max} + \gamma \,\|Q_t\|_\infty\big)^2.$$

Combining,

$$\mathbb{E}\big[w_t(s_t, a_t)^2 \mid \mathcal{F}_t\big] \;\le\; 4\big(R_{\max} + \gamma \,\|Q_t\|_\infty\big)^2.$$

Step 3: putting it in the required form. We want a constant C such that the right-hand side is bounded by $C(1 + \|Q_t\|_\infty^2)$. Applying $(a+b)^2 \le 2(a^2+b^2)$ with $a = R_{\max}$, $b = \gamma\,\|Q_t\|_\infty$:

$$\begin{aligned} 4\big(R_{\max} + \gamma \,\|Q_t\|_\infty\big)^2 &\le 8\big(R_{\max}^2 + \gamma^2 \,\|Q_t\|_\infty^2\big) \\ &\le 8\max\big(R_{\max}^2,\, \gamma^2\big)\big(1 + \|Q_t\|_\infty^2\big). \end{aligned}$$

So $C = 8\max(R_{\max}^2, \gamma^2)$ works. For $(s, a) \neq (s_t, a_t)$, we have $w_t(s, a) = 0$, so the bound holds trivially with the same C.

Exercise 6.6. Conclude: state the conditions on the learning rate schedule and the exploration policy under which Q-learning converges, i.e. $Q_t \to Q^*$ with probability 1. *Remark.* The ARENA implementation uses a constant learning rate α, which does not satisfy $\sum_t \alpha_t(s,a)^2 < \infty$. In practice, constant-rate Q-learning does not converge exactly but oscillates in a neighbourhood of Q^* whose size shrinks with α.

Solution to Exercise 6.6

By Exercises 6.1 to 6.5, the Q-learning iteration satisfies all the hypotheses of Theorem 6.1, provided the learning rate conditions hold: for each $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\sum_{t=0}^{\infty} \alpha_t(s,a) = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t(s,a)^2 < \infty, \qquad \alpha_t(s,a) \in [0,1].$$

Under these conditions, Theorem 6.1 gives $Q_t \to Q^*$ with probability 1.
Note that the first condition already implies that each state-action pair (s, a) must be visited infinitely often. A concrete way to achieve all three conditions is to use an exploration policy that visits every (s, a) infinitely often together with a per-coordinate learning rate $\alpha_t(s, a) = 1/n_t(s, a)$, where $n_t(s, a)$ counts the number of visits to (s, a) up to time t. This satisfies $\sum \alpha_t(s, a) = \sum_{k=1}^{\infty} 1/k = \infty$ and $\sum \alpha_t(s, a)^2 = \sum_{k=1}^{\infty} 1/k^2 < \infty$.

7 Exact policy evaluation

In Exercise 3.4, we showed that iterating $\mathcal{B}_\pi$ converges to V_π; this is the numerical policy evaluation scheme used in the ARENA materials. When the state space is finite and the dynamics (T, R) are known, there is also an *exact* (closed-form) solution: the Bellman equation becomes a linear system that can be solved directly. In this problem we derive that closed form for deterministic policies.

Definition 7.1 (Matrix-vector notation for a deterministic policy)**.** Fix a deterministic policy π. Treating the value function as a vector in $\mathbb{R}^{|\mathcal{S}|}$, define:

- $\mathbf{v}^\pi \in \mathbb{R}^{|\mathcal{S}|}$ with $\mathbf{v}^\pi_s := V_\pi(s)$,
- $\mathbf{P}^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ with $\mathbf{P}^\pi_{s,s'} := T(s' \mid s, \pi(s))$ (the **transition matrix** under π),
- $\mathbf{R}^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ with $\mathbf{R}^\pi_{s,s'} := R(s, \pi(s), s')$ (the **reward matrix**),
- $\mathbf{r}^\pi \in \mathbb{R}^{|\mathcal{S}|}$ with $\mathbf{r}^\pi_s := \sum_{s'} \mathbf{P}^\pi_{s,s'} \mathbf{R}^\pi_{s,s'} = \sum_{s'} T(s' \mid s, \pi(s))\, R(s, \pi(s), s')$ (the **expected immediate reward**).

Fact 7.2 (Neumann series)**.** *Equip $\mathbb{R}^{n\times n}$ with the induced ∞-norm*

$$\|\mathbf{A}\|_\infty := \max_i \sum_j |\mathbf{A}_{ij}| \qquad \textit{(maximum absolute row sum).}$$

If $\|\mathbf{A}\|_\infty < 1$, then $\mathbf{I} - \mathbf{A}$ is invertible, and

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k.$$

Exercise 7.1. Starting from the Bellman equation (Exercise 1.1) applied to a deterministic policy π, show that

Note

Exact Policy Evaluation

$$\mathbf{v}^{\pi} \;=\; (\mathbf{I} - \gamma\,\mathbf{P}^{\pi})^{-1}\,\mathbf{r}^{\pi},$$

assuming $\mathbf{I} - \gamma\,\mathbf{P}^{\pi}$ is invertible (which we prove in Exercise 7.2).
Hint: specialize the Bellman equation to a deterministic policy and write the resulting system of $|\mathcal{S}|$ equations in matrix-vector form.

Solution to Exercise 7.1

Since π is deterministic, $\pi(a \mid s) = [\![a = \pi(s)]\!]$, and the Bellman equation from Exercise 1.1 collapses the sum over actions:

$$V_{\pi}(s) \;=\; \sum_{s' \in \mathcal{S}} T(s' \mid s, \pi(s))\big(R(s, \pi(s), s') + \gamma\, V_{\pi}(s')\big).$$

Splitting the reward and value terms and rewriting with the matrix notation of Definition 7.1:

$$\mathbf{v}^{\pi}_{s} = \sum_{s'} \mathbf{P}^{\pi}_{s,s'}\,\mathbf{R}^{\pi}_{s,s'} + \gamma \sum_{s'} \mathbf{P}^{\pi}_{s,s'}\,\mathbf{v}^{\pi}_{s'} = \mathbf{r}^{\pi}_{s} + \gamma\,(\mathbf{P}^{\pi}\mathbf{v}^{\pi})_{s}.$$

This holds for every s, so in vector form

$$\mathbf{v}^{\pi} \;=\; \mathbf{r}^{\pi} + \gamma\,\mathbf{P}^{\pi}\mathbf{v}^{\pi} \quad\Longleftrightarrow\quad (\mathbf{I} - \gamma\,\mathbf{P}^{\pi})\,\mathbf{v}^{\pi} \;=\; \mathbf{r}^{\pi}.$$

Assuming $\mathbf{I} - \gamma\,\mathbf{P}^{\pi}$ is invertible, left-multiplying by its inverse gives $\mathbf{v}^{\pi} = (\mathbf{I} - \gamma\,\mathbf{P}^{\pi})^{-1}\,\mathbf{r}^{\pi}$.

Exercise 7.2. Prove that $\mathbf{I} - \gamma\,\mathbf{P}^{\pi}$ is invertible for any deterministic policy π and any $\gamma \in (0, 1)$.
Hint: apply Fact 7.2.

Solution to Exercise 7.2

The matrix $\mathbf{P}^{\pi}$ is **row-stochastic**: $\mathbf{P}^{\pi}_{s,s'} \geq 0$ and $\sum_{s'} \mathbf{P}^{\pi}_{s,s'} = \sum_{s'} T(s' \mid s, \pi(s)) = 1$ for every s. Therefore

$$\|\mathbf{P}^{\pi}\|_{\infty} \;=\; \max_{s} \sum_{s'} |\mathbf{P}^{\pi}_{s,s'}| \;=\; \max_{s} \sum_{s'} \mathbf{P}^{\pi}_{s,s'} \;=\; 1,$$

and so $\|\gamma \mathbf{P}^\pi\|_\infty = \gamma < 1$. By the Neumann series (Fact 7.2), $\mathbf{I} - \gamma \mathbf{P}^\pi$ is invertible, with

$$(\mathbf{I} - \gamma \mathbf{P}^\pi)^{-1} = \sum_{k=0}^{\infty} (\gamma \mathbf{P}^\pi)^k = \sum_{k=0}^{\infty} \gamma^k (\mathbf{P}^\pi)^k.$$

Combining with Exercise 7.1,

$$\mathbf{v}^\pi = \sum_{k=0}^{\infty} \gamma^k (\mathbf{P}^\pi)^k \mathbf{r}^\pi,$$

which has a direct interpretation: $(\mathbf{P}^\pi)^k \mathbf{r}^\pi$ is the vector of expected rewards exactly k steps into the future under π, and $\mathbf{v}^\pi$ is their discounted sum.

Remark. The same derivation extends to stochastic policies by defining $\mathbf{P}^\pi_{s,s'} = \sum_a \pi(a \mid s)\, T(s' \mid s, a)$ and $\mathbf{r}^\pi_s = \sum_a \pi(a \mid s) \sum_{s'} T(s' \mid s, a)\, R(s, a, s')$. The matrix $\mathbf{P}^\pi$ remains row-stochastic, so the argument above applies unchanged.

References

[Put94] M. L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. Wiley, New York, NY, 1994.

[Tsi94] John N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.