

Reinforcement Learning: Exercises

Leon Lang
Iliad

David Quarel
ARENA

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 $t = 0, 1, 2, \dots$. 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
- $\llbracket P \rrbracket$ — **Iverson bracket**: 1 if P is true, 0 if false

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

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

^aNoting 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** $\pi : \mathcal{S} \rightarrow \Delta \mathcal{A}$ 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** $(s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, \dots)$ is generated by

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

for $k = 0, 1, 2, \dots$. 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:

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

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 :

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

The expectation averages the return over all future trajectories $(a_t, s_{t+1}, a_{t+1}, s_{t+2}, \dots)$ 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

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

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

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:

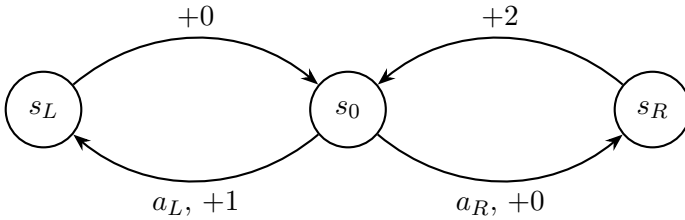
$$Q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid s_t = s, a_t = a].$$

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.

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

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

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 $\gamma = \frac{1}{2}$.

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 $F : X \rightarrow X$ 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 $\lim_{n \rightarrow \infty} d(F^n(x_0), x^*) = 0$ 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 $\mathcal{V} = \{V : \mathcal{S} \rightarrow \mathbb{R}\}$ 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** $\mathcal{B} : \mathcal{V} \rightarrow \mathcal{V}$ is defined by

$$(\mathcal{B}V)(s) := \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s' | s, a) (R(s, a, s') + \gamma V(s')).$$

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

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

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

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

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 $\mathcal{B}^n V_0 \rightarrow V^*$ as $n \rightarrow \infty$.

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

Note

Bellman Policy Operator

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

is also a contraction with factor γ . Conclude that V_π is its unique fixed point and that $\mathcal{B}_\pi^n V_0 \rightarrow V_\pi$ for any initial $V_0 \in \mathcal{V}$ as $n \rightarrow \infty$.

Remark

This shows that the numerical policy evaluation algorithm from the ARENA materials converges to the correct solution.

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.

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

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

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^*} .

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.

4 Policy improvement theorem

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

Note

Policy Improvement

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

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

Exercise 4.2. Show that $V_{\pi'}(s) \geq V_\pi(s)$ for all $s \in \mathcal{S}$.

Hint: Show that $\mathcal{B}_{\pi'} V_\pi \geq V_\pi$ pointwise using Exercise 3.5, then iterate $\mathcal{B}_{\pi'}$ and take the limit.

Exercise 4.3. The **policy iteration** algorithm generates a sequence of policies $\pi_0, \pi_1, \pi_2, \dots$ 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?

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^* .

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

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

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 $V_k \rightarrow V^*$), and bound each term using the contraction of $\mathcal{B}$.

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

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

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.

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.

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

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

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

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 $n \geq 0$ (with $V_0 \equiv 0$).

Remark

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

Convergence Bounds

$$\begin{aligned} \|V_n - V^*\|_\infty &\leq \underbrace{\frac{\gamma}{1-\gamma} \|V_n - V_{n-1}\|_\infty}_{\text{a posteriori, tightest}} \leq \underbrace{\frac{\gamma^n}{1-\gamma} \|V_1 - V_0\|_\infty}_{\text{a priori}} \\ &\leq \underbrace{\frac{\gamma^n}{1-\gamma} R_{\max}}_{\text{upfront, assumes } V_0 \equiv 0, |R| \leq R_{\max}}. \end{aligned}$$

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 $\|V_n - V^*\|_\infty \leq \varepsilon$, iterate until $\|V_n - V_{n-1}\|_\infty \leq \frac{1-\gamma}{\gamma} \varepsilon$.

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 $\mathcal{X} = \{x : I \rightarrow \mathbb{R}\}$ be the space of real-valued functions on a finite set I , equipped with the sup-norm $\|x\|_\infty = \max_i |x(i)|$. Let $F : \mathcal{X} \rightarrow \mathcal{X}$ 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 $t \geq 0$ let $\alpha_t : I \rightarrow [0, 1]$ be a **learning rate** function and $w_t : I \rightarrow \mathbb{R}$ a random **noise** function (both functions on I , one value per coordinate). Consider the stochastic iteration*

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

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 $i \in I$:

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

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

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

for some constant $C > 0$.

Then $x_t(i) \rightarrow x^*(i)$ 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 $x_1, x_2, \dots$ are random (they depend on past noise $w_0, \dots, w_{t-1}$), 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 $x_t(i) \rightarrow x^*(i)$ with probability 1 (also called **almost sure convergence**) means

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

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:

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

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 $Q_t : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ 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

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

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 $\mathcal{Q} = \{Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}\}$ 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, $\mathcal{B} : \mathcal{Q} \rightarrow \mathcal{Q}$, by

Note

Bellman Q-Operator

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

(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.

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

- (a) Show that $Q^*(s, a) = \sum_{s' \in \mathcal{S}} T(s' | s, a) [R(s, a, s') + \gamma V^*(s')]$.
- (b) Show that $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$.
- (c) Conclude that Q^* is the unique fixed point of $\mathcal{B}$.

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

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

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 .

Exercise 6.4. Show that the noise satisfies $\mathbb{E}[w_t(s, a) | \mathcal{F}_t] = 0$ 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?

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

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

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

Exercise 6.6. Conclude: state the conditions on the learning rate schedule and the exploration policy under which Q-learning converges, i.e. $Q_t \rightarrow Q^*$ 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 .

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

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

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

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.

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.

Remark

The same derivation extends to stochastic policies by defining $\mathbf{P}_{s,s'}^\pi = \sum_a \pi(a | s) T(s' | s, a)$ and $\mathbf{r}_s^\pi = \sum_a \pi(a | s) \sum_{s'} T(s' | s, a) R(s, a, s')$. 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.