Reinforcement Learning Study Notes

Pei-Yuan Wu

July 8, 2024

- $[\![m, n]\!]$ indicates the collection of intervals between m and n.
- We slightly abuse notation and let $[K]$ denote the set $\{0, 1, 2, ..., K-1\}$ for an integer K.

Chapter 1

Discounted Markov Decision Process

1.1 Markov Decision Process Basics

In reinforcement learning, the interactions between the agent and the environment are often described by a discounted Markov Decision Process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, specified by:

- A state space $\mathcal{S}$. For mathematical convenience, we assume that $\mathcal{S}$ is finite or countably infinite.
- An action space $\mathcal{A}$. For mathematical convenience, we assume that $\mathcal{A}$ is finite. [1]
- A transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(S)$, where $\Delta(S)$ is the space of probability distributions over $\mathcal{S}$ (i.e., the probability simplex). $P(s'|s, a)$ is the probability of transitioning into state s' upon taking action a in state s. We use $P_{s,a}$ to denote the vector $P(\cdot|s, a)$.
- A reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. $r(s, a)$ is the immediate reward associated with taking action a in state s.
- A discount factor $\gamma \in [0, 1)$, which defines a horizon for the problem.
- An initial state distribution $\mu \in \Delta(S)$, which specifies how the initial state s_0 is generated.

1.1.1 The objectives, policies, and values

Policies. In a given MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, the agent interacts with the environment according to the following protocol: the agent starts at some state $s_0 \sim \mu$; at each time step $t = 0, 1, 2, ...$, the agent takes an action $a_t \in \mathcal{A}$, obtains the immediate reward $r_t = r(s_t, a_t)$, and observes the next state s_{t+1} sampled according to $s_{t+1} \sim P(\cdot|s_t, a_t)$. The interaction record at time t,

$$\tau_t = (s_0, a_0, r_0, s_1, \cdots, s_t)$$

is called a *trajectory* (up to time t), which includes the observed state at time t. We often denote $\tau = \tau_\infty$ unless otherwise specified.

In the most general setting, a policy specifies a decision-making strategy in which the agent chooses actions adaptively based on the history of observations; precisely, a policy is a (possibly randomized) mapping from a trajectory to an action, i.e., $\pi : \mathcal{H} \to \Delta(\mathcal{A})$ where $\mathcal{H}$ is the set of all possible trajectories (of all lengths) and $\Delta(\mathcal{A})$ is the space of probability distributions over $\mathcal{A}$. A *stationary*

[1] Some content in this manuscript may be applicable to the more general case where $\mathcal{A}$ is countable infinite, which will be specified in blue-colored text.

policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies a decision-making strategy in which the agent chooses actions based only on the current state, i.e., $a_t \sim \pi(\cdot|s_t)$. A deterministic, stationary policy is of the form $\pi : \mathcal{S} \to \mathcal{A}$.

Values. We now define values for (general) policies. For a fixed policy and a starting state $s_0 = s$, we define the value function $V_M^\pi : \mathcal{S} \to \mathbb{R}$ as the discounted sum of future rewards

$$V_M^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, s_0 = s\right],$$

where expectation is with respect to the randomness of the trajectory, that is, the randomness in state transitions and the stochasticity of π. Here, since $r(s,a)$ is bounded between 0 and 1, we have $0 \le V_M^\pi(s) \le 1/(1-\gamma)$.

Similarly, the action-value (or Q-value) function $Q_M^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as

$$Q_M^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, s_0 = s, a_0 = a\right],$$

and $Q_M^\pi(s,a)$ is also bounded by $1/(1-\gamma)$.

Goal: Given a state s, the goal of the agent is to find a policy π that maximizes the value, i.e., the optimization problem the agent seeks to solve is

$$\max_\pi V_M^\pi(s) \tag{1.1.1}$$

where the max is over all (possibly non-stationary and randomized) policies.

We drop the dependence on M and write V^π when it is clear from context.

1.1.2 Bellman consistency equations for stationary policies

Stationary policies satisfy the following consistency conditions:

Lemma 1.1.1. *Suppose that π is a stationary policy. Then V^π and Q^π satisfy the following Bellman consistency equations: for all $s \in \mathcal{S}$, $a \in \mathcal{A}$,*

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s,a)]$$
$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')].$$

Proof. Trivial. □

It is helpful to view V^π as vector of length $|S|$ and Q^π and r as vectors of length $|\mathcal{S}| \cdot |\mathcal{A}|$. We overload notation and let P also refer to a matrix of size $(|\mathcal{S}| \cdot |\mathcal{A}|) \times |\mathcal{S}|$ where the entry $P_{(s,a),s'}$ is equal to $P(s'|s,a)$. [2] We also will define P^π to be the transition matrix on state-action pairs induced by a stationary policy π, specifically:

$$P^\pi_{(s,a),(s',a')} := P(s'|s,a)\pi(a'|s').$$

[2]If either $|S|$ or $|\mathcal{A}|$ is countably infinite, a more rigorous statement is to regard V^π as a vector in $\ell^\infty(\mathcal{S})$, and Q^π and r as vectors in $\ell^\infty(\mathcal{S} \times \mathcal{A})$, and $P \in \mathcal{B}(\ell^\infty(\mathcal{S}), \ell^\infty(\mathcal{S} \times \mathcal{A}))$ for which $[Px]_{(s,a)} = \mathbb{E}_{s' \sim P(\cdot|s,a)}[[x]_{s'}]$.

In particular, for deterministic policies we have

$$P^{\pi}_{(s,a),(s',a')} = \begin{cases} P(s'|s,a) & \text{if } a' = \pi(s') \\ 0 & \text{if } a' \neq \pi(s') \end{cases}$$

With this notation, it is straightforward to verify:

$$Q^{\pi} = r + \gamma P V^{\pi}$$
$$Q^{\pi} = r + \gamma P^{\pi} Q^{\pi}.$$

Definition 1.1.2. *The Bellman operator for a policy π is denoted as $\mathbb{B}^{\pi} : \ell^{\infty}(\mathcal{S} \times \mathcal{A}) \to \ell^{\infty}(\mathcal{S} \times \mathcal{A})$,*

$$\mathbb{B}^{\pi} Q := r + \gamma P^{\pi} Q.$$

Corollary 1.1.3. *We have that*

$$Q^{\pi} = (I - \gamma P^{\pi})^{-1} r = \lim_{n\to\infty} \sum_{k=0}^{n} \gamma^k (P^{\pi})^k r \tag{1.1.2}$$

where I is the identity matrix.

Proof. Note that $P^{\pi} \in \mathcal{B}(X)$ for which $X = \ell^{\infty}(\mathcal{S} \times \mathcal{A})$ is a Banach space, and that $\|P^{\pi}\| \leq 1$. Define $\Lambda_n = \sum_{k=0}^{n} \gamma^k (P^{\pi})^k$ for $n \in \mathbb{N}$, then $(\Lambda_n)_{n=1}^{\infty}$ is a Cauchy sequence in Banach space $\mathcal{B}(X)$, so there exists $\Lambda = \lim_{n\to\infty} \Lambda_n \in \mathcal{B}(X)$. Note that for each $x \in X$,

$$\Lambda(I - \gamma P^{\pi})x = \lim_{n\to\infty} \Lambda_n (I - \gamma P^{\pi})x = \lim_{n\to\infty} (I - \gamma^{n+1} (P^{\pi})^{n+1})x = x$$
$$(I - \gamma P^{\pi})\Lambda x = (I - \gamma P^{\pi}) \lim_{n\to\infty} \Lambda_n x = \lim_{n\to\infty} (I - \gamma P^{\pi})\Lambda_n x = \lim_{n\to\infty} (I - \gamma^{n+1} (P^{\pi})^{n+1})x = x$$

Hence $(I - \gamma P^{\pi})^{-1}$ exists and equals Λ. $\square$

Corollary 1.1.4. *We have that*

$$[(1-\gamma)(I - \gamma P^{\pi})^{-1}]_{(s,a),(s',a')} = (1-\gamma) \sum_{k=0}^{\infty} \gamma^k \mathbb{P}[s_k = s', a_k = a' | \pi, s_0 = s, a_0 = a]$$

so we can view the (s,a)-th row of this matrix as an induced distribution over states and actions when following π after starting with $s_0 = s$ and $a_0 = a$.

Proof. Following the discussion in Corollary.1.1.3, define $e_{(s',a')} \in X$ and $e^*_{(s,a)} \in X^*$ as

$$e_{(s',a')}(\hat{s}, \hat{a}) = \begin{cases} 1 & \text{if } \hat{s} = s', \hat{a} = a' \\ 0 & \text{otherwise} \end{cases}, \quad e^*_{(s,a)} : x \mapsto [x]_{(s,a)}$$

The result then follows by

$$[(I - \gamma P^{\pi})^{-1}]_{(s,a),(s',a')} = \langle \Lambda e_{(s',a')}, e^*_{(s,a)} \rangle = \lim_{n\to\infty} \langle \Lambda_n e_{(s',a')}, e^*_{(s,a)} \rangle = \lim_{n\to\infty} \sum_{k=0}^{n} \gamma^k \langle (P^{\pi})^k e_{(s',a')}, e^*_{(s,a)} \rangle$$
$$= \lim_{n\to\infty} \sum_{k=0}^{n} \gamma^k \mathbb{P}[s_k = s', a_k = a' | \pi, s_0 = s, a_0 = a].$$

$\square$

1.1.3 Bellman optimality equations

A remarkable and convenient property of MDPs is that there exists a stationary and deterministic policy that simultaneously maximizes $V^\pi(s)$ for all $s \in \mathcal{S}$. This is formalized in the following theorem:

Theorem 1.1.5. *Let Π be the set of all non-stationary and randomized policies. Define:*

$$V^*(s) := \sup_{\pi \in \Pi} V^\pi(s)$$
$$Q^*(s,a) := \sup_{\pi \in \Pi} Q^\pi(s,a)$$

which is finite since $V^\pi(s)$ and $Q^\pi(s,a)$ are bounded between 0 and $1/(1-\gamma)$.

There exists a stationary and deterministic policy π such that for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$V^\pi(s) = V^*(s)$$
$$Q^\pi(s,a) = Q^*(s,a).$$

We refer to such a π as an optimal policy.

Proof. First, let us show that conditioned on $(s_0, a_0, r_0, s_1) = (s, a, r, s')$, the maximum future discounted value, from time 1 onwards, is not a function of s, a, r. Specifically,

$$\sup_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, (s_0, a_0, r_0, s_1) = (s, a, r, s') \right] = \gamma V^*(s')$$

For any policy π, define an "offset" policy $\pi_{(s,a,r)}$, which is the policy that chooses actions on a trajectory τ according to the same distribution that π chooses actions on the trajectory (s, a, r, τ). By the Markov property, we have that

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, (s_0, a_0, r_0, s_1) = (s, a, r, s') \right] = \gamma \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi_{(s,a,r)}, s_0 = s' \right] = \gamma V^{\pi_{(s,a,r)}}(s').$$

Hence, due to $\{\pi_{(s,a,r)} : \pi \in \Pi\} = \Pi$, we have

$$\sup_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, (s_0, a_0, r_0, s_1) = (s, a, r, s') \right] = \gamma \sup_{\pi \in \Pi} V^{\pi_{(s,a,r)}}(s') = \gamma \sup_{\pi \in \Pi} V^{\pi}(s') = \gamma V^*(s')$$

as desired. We now show the deterministic and stationary policy

$$\pi^*(s) \in \arg\max_{a \in \mathcal{A}} \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')] \right) \tag{1.1.3}$$

is an optimal strategy. To see this, note that

$$\begin{aligned}
V^*(s_0) &= \sup_{\pi \in \Pi} \mathbb{E}\left[r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi \right] \\
&= \sup_{\pi \in \Pi} \mathbb{E}\left[r(s_0, a_0) + \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, (s_0, a_0, r_0, s_1) \right] \middle| \pi \right] \\
&\le \sup_{\pi \in \Pi} \mathbb{E}\left[r(s_0, a_0) + \sup_{\pi' \in \Pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi', (s_0, a_0, r_0, s_1) \right] \middle| \pi \right] \\
&= \sup_{\pi \in \Pi} \mathbb{E}\left[r(s_0, a_0) + \gamma V^*(s_1) \middle| \pi \right] = \mathbb{E}\left[r(s_0, a_0) + \gamma V^*(s_1) \middle| \pi^* \right]
\end{aligned}$$

where the last equality follows from the definition of π^*. Now, by recursion,

$$V^*(s_0) \leq \mathbb{E}\left[r(s_0,a_0) + \gamma V^*(s_1)\middle|\,\pi^*\right] \leq \mathbb{E}\left[r(s_0,a_0) + \gamma r(s_1,a_1) + \gamma^2 V^*(s_2)\middle|\,\pi^*\right] \leq \cdots \leq V^{\pi^*}(s_0)$$

where the last inequality follows by applying the dominated convergence theorem to the following fact

$$\lim_{n\to\infty} \sum_{t=0}^{T-1} \gamma^t r(s_t,a_t) + \gamma^T V^*(s_T) = \sum_{t=0}^{\infty} \gamma^t r(s_t,a_t).$$

Since $V^{\pi^*}(s) \leq \sup_{\pi\in\Pi} V^\pi(s) = V^*(s)$, we have that $V^{\pi^*} = V^*$ as desired. Analogously, note that

$$\begin{aligned}
Q^*(s_0,a_0) &= \sup_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)\middle|\, \pi, (s_0,a_0)\right] \\
&= r(s_0,a_0) + \sup_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t,a_t)\middle|\, \pi, (s_0,a_0)\right] \\
&= r(s_0,a_0) + \sup_{\pi\in\Pi} \mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t,a_t)\middle|\, \pi, (s_0,a_0,r_0,s_1)\right]\right] \\
&\leq r(s_0,a_0) + \mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[\sup_{\pi\in\Pi}\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t,a_t)\middle|\, \pi, (s_0,a_0,r_0,s_1)\right]\right] \\
&= r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[V^*(s_1)\right] = r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[V^{\pi^*}(s_1)\right] = Q^{\pi^*}(s_0,a_0).
\end{aligned}$$

Since $Q^{\pi^*}(s,a) \leq \sup_{\pi\in\Pi} Q^\pi(s,a) = Q^*(s,a)$, we have that $Q^{\pi^*} = Q^*$ as desired. $\square$

Let us say a vector $Q \in \ell^\infty(\mathcal{S}\times\mathcal{A})$ satisfies the *Bellman optimality equations* if

$$Q(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q(s',a')\right].$$

This shows that we may restrict ourselves to using stationary and deterministic policies without any loss in performance. Theorem.1.1.8, also due to [Bellman, 1956], gives a precise characterization of the optimal value function.

Definition 1.1.6. *The Bellman optimality operator* $\mathbb{T} : \ell^\infty(\mathcal{S}\times\mathcal{A}) \to \ell^\infty(\mathcal{S}\times\mathcal{A})$ *is defined as*

$$\mathbb{T}Q := r + \gamma P V_Q \quad \textit{where} \quad V_Q(s) := \max_{a\in\mathcal{A}} Q(s,a).$$

This allows us to rewrite the Bellman optimality equation in the concise form $Q = \mathbb{T}Q$, namely Q is a fixed point of the operator $\mathbb{T}$.

Notation 1.1.7. *We denote* π_Q *as the greedy policy with respect to a vector* $Q \in \ell^\infty(\mathcal{S}\times\mathcal{A})$*, i.e.,*

$$\pi_Q(s) := \arg\max_{a\in\mathcal{A}} Q(s,a)$$

where ties are broken in some arbitrary (and deterministic) manner.

Theorem 1.1.8 (Bellman Optimality Equations)**.** *For any $Q \in \ell^\infty(\mathcal{S} \times \mathcal{A})$, we have that $Q = Q^*$ iff Q satisfies the Bellman optimality equations. Furthermore, the stationary and deterministic policy π_{Q^*} is an optimal policy.*

Proof. We first show sufficiency, i.e., that Q^* (the state-action value of an optimal policy) satisfies $Q^* = \mathbb{T}Q^*$. Let π^* be a stationary and deterministic policy defined by (1.1.3). It follows by Theorem.1.1.5 that π^* is optimal, so

$$Q^*(s,a) = Q^{\pi^*}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^{\pi^*}(s')] = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

Hence $\pi^*(s) \in \arg\max_{a \in \mathcal{A}} Q^*(s,a)$. This leads to

$$V^*(s) = V^{\pi^*}(s) = Q^{\pi^*}(s, \pi^*(s)) = Q^*(s, \pi^*(s)) = \max_{a \in \mathcal{A}} Q^*(s,a)$$

Combining the previous two equations yields $Q^* = \mathbb{T}Q^*$ and proves sufficiency.

For the converse, suppose $Q = \mathbb{T}Q$ for some $Q \in \ell^\infty(\mathcal{S} \times \mathcal{A})$. Let $\pi = \pi_Q$, since $Q = r + \gamma P^\pi Q$, so

$$Q = (I - \gamma P^\pi)^{-1} r = Q^\pi$$

For any other deterministic and stationary policy π':

$$\begin{aligned} Q - Q^{\pi'} &= Q^\pi - Q^{\pi'} = Q^\pi - (I - \gamma P^{\pi'})^{-1} r = (I - \gamma P^{\pi'})^{-1}((I - \gamma P^{\pi'}) - (I - \gamma P^\pi))Q^\pi \\ &= \gamma (I - \gamma P^{\pi'})^{-1}(P^\pi - P^{\pi'})Q^\pi. \end{aligned}$$

Recall that $(I - \gamma P^{\pi'})^{-1}$ is a matrix with all non-negative entries (cf. Corollary.1.1.4), so it suffices to show $[(P^\pi - P^{\pi'})Q^\pi]_{(s,a)} \geq 0$, which follows by

$$[(P^\pi - P^{\pi'})Q^\pi]_{(s,a)} = \mathbb{E}_{s' \sim P(\cdot|s,a)}[Q^\pi(s', \pi(s')) - Q^\pi(s', \pi'(s'))] \geq 0$$

where the last step uses that π is the greedy policy with respect to $Q = Q^\pi$. Thus we have that $Q \geq Q^{\pi'}$ for all deterministic and stationary policy π', which shows $Q = Q^*$ following Theorem.1.1.5. This completes the proof. ☐

1.1.4 Adventages and The Performance Difference Lemma

Throughout, we will overload notation where, for a distribution μ over $\mathcal{S}$, we write:

$$V^\pi(\mu) = \mathbb{E}_{s \sim \mu}[V^\pi(s)].$$

The *adventage* $A^\pi(s,a)$ of a policy π is defined as

$$A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s).$$

Note that

$$A^*(s,a) := A^{\pi^*}(s,a) \leq 0$$

for all state-action pairs.

Define the discounted state visitation distribution $d^\pi_{s_0}$ as:

$$d^\pi_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s|\pi, s_0)$$

where $\mathbb{P}(s_t = s|\pi, s_0)$ is the state visitation probability, under π starting at state s_0. We also write

$$d^\pi_\mu(s) = \mathbb{E}_{s_0\sim\mu}[d^\pi_{s_0}(s)]$$

for a distribution μ over $\mathcal{S}$.

Lemma 1.1.9 (The performance difference lemma)**.** *For all stationary policies π, π' and distributions μ over $\mathcal{S}$,*

$$V^\pi(\mu) - V^{\pi'}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi_\mu}\mathbb{E}_{a\sim\pi(\cdot|s)}[A^{\pi'}(s,a)]$$

Proof. Let $\mathbb{P}(\tau|\pi, s_0)$ denote the probability of observing a trajectory τ when starting in state s_0 and following the policy π. Then

$$\begin{aligned}
V^\pi(s_0) - V^{\pi'}(s_0) &= \mathbb{E}_{\tau\sim\mathbb{P}(\cdot|\pi,s_0)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] - V^{\pi'}(s_0) \\
&= \mathbb{E}_{\tau\sim\mathbb{P}(\cdot|\pi,s_0)}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t,a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\right)\right] \\
&= \mathbb{E}_{\tau\sim\mathbb{P}(\cdot|\pi,s_0)}\left[\sum_{t=0}^{\infty}\gamma^t \left(Q^{\pi'}(s_t,a_t) - V^{\pi'}(s_t)\right)\right] \\
&= \mathbb{E}_{\tau\sim\mathbb{P}(\cdot|\pi,s_0)}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi'}(s_t,a_t)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi_{s_0}}\mathbb{E}_{a\sim\pi(\cdot|s)}[A^{\pi'}(s,a)].
\end{aligned}$$

□

1.2 Iterative Methods

Planning referes to the problem of computing π^*_M given the MDP specification $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. This section reviews classical planning algorithms that compute Q^*.

1.2.1 Value Iteration

A simple algorithm is to iteratively apply the fixed point mapping: stating at some Q, we iteratively apply the Bellman optimality operator $\mathbb{T}$:

$$Q \leftarrow \mathbb{T}Q,$$

This algorithm is referred to as *Q-value iteration.*

Lemma 1.2.1 (contraction)**.** *For any two vectors $Q, Q' \in \ell^\infty(\mathcal{S} \times \mathcal{A})$,*

$$\|\mathbb{T}Q - \mathbb{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof. First, let us show that for all $s \in \mathcal{S}$, $|V_Q(s) - V_{Q'}(s)| \leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|$. Assume $V_Q(s) \geq V_{Q'}(s)$ (the other direction is symmetric), and let a be the greedy action for Q at s. Then

$$|V_Q(s) - V_{Q'}(s)| = Q(s,a) - \max_{a' \in \mathcal{A}} Q'(s,a') \leq Q(s,a) - Q'(s,a) \leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|.$$

The claim then follows by

$$\|\mathbb{T}Q - \mathbb{T}Q'\|_\infty = \gamma \|P(V_Q - V_{Q'})\|_\infty \leq \gamma \|V_Q - V_{Q'}\|_\infty \leq \gamma \|Q - Q'\|_\infty.$$

□

The following result bounds the sub-optimality of the greedy policy itself, based on the error in Q-value function.

Lemma 1.2.2 (Q-Error Amplification)**.** *For any vector $Q \in \ell^\infty(\mathcal{S} \times \mathcal{A})$,*

$$V^{\pi_Q} \geq V^* - \frac{2\|Q - Q^*\|_\infty}{1 - \gamma} \mathbf{1}$$

where $\mathbf{1}$ denotes the vector of all ones.

Proof. Fix state s and let $a = \pi_Q(s)$. Let $\pi^* = \pi_{Q^*}$ be an optimal policy (cf. Theorem.1.1.3), we have:

$$\begin{aligned} V^*(s) - V^{\pi_Q}(s) =& Q^*(s, \pi^*(s)) - Q^{\pi_Q}(s,a) = (Q^*(s,\pi^*(s)) - Q^*(s,a)) + (Q^*(s,a) - Q^{\pi_Q}(s,a)) \\ =& (Q^*(s,\pi^*(s)) - Q(s,\pi^*(s))) + (Q(s,\pi^*(s)) - Q(s,a)) + (Q(s,a) - Q^*(s,a)) \\ &+ \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s') - V^{\pi_Q}(s')] \\ \leq& 2\|Q - Q^*\|_\infty + \gamma \|V^* - V^{\pi_Q}\|_\infty \end{aligned}$$

The claim then follows by

$$\|V^* - V^{\pi_Q}\|_\infty \leq 2\|Q - Q^*\|_\infty + \gamma \|V^* - V^{\pi_Q}\|_\infty.$$

□

Theorem 1.2.3 (Q-value iteration convergence)**.** *Set* $\mathbf{0} \leq Q^{(0)} \leq \mathbf{1}/(1-\gamma)$*. For* $k = 0, 1, ...,$ *suppose:*

$$Q^{(k+1)} = \mathbb{T}Q^{(k)}$$

Let $\pi^{(k)} = \pi_{Q^{(k)}}$*. For* $k \geq \frac{\log \frac{2}{(1-\gamma)^2\epsilon}}{\log(1/\gamma)} \leq \frac{1}{1-\gamma} \log \frac{2}{(1-\gamma)^2\epsilon}$*,*

$$V^{\pi^{(k)}} \geq V^* - \epsilon \mathbf{1}.$$

Proof. Since Q^* is a fixed point of $\mathbb{T}$, Lemma.1.2.1 gives

$$\|Q^{(k)} - Q^*\|_\infty = \|\mathbb{T}^k Q^{(0)} - \mathbb{T}^k Q^*\|_\infty \leq \gamma^k \|Q^{(0)} - Q^*\|_\infty \leq \frac{\gamma^k}{1-\gamma}.$$

The proof is completed with our choice of k and using Lemma.1.2.2. □

1.2.2 Policy Iteration

The policy iteration algorithm starts from an arbitrary policy π_0, and repeat the following iterative procedure: for $k = 0, 1, 2, ...$

1. *Policy evaluation.* Compute Q^{π_k}.
2. *Policy improvement.* Update the policy:

$$\pi_{k+1} = \pi_{Q^{\pi_k}}$$

In each iteration, we compute the Q-value function of π_k, using the analytical form given in (1.1.2), and update the policy to be greedy with respect to this new Q-value. The first step is often called *policy evaluation*, and the second step is often called *policy improvement*.

Lemma 1.2.4. *We have that:*

1. $Q^{\pi_{k+1}} \geq \mathbb{T}Q^{\pi_k} \geq Q^{\pi_k}$.
2. $\|Q^{\pi_{k+1}} - Q^*\|_\infty \leq \gamma \|Q^{\pi_k} - Q^*\|_\infty$

Proof. First we show that $\mathbb{T}Q^{\pi_k} \geq Q^{\pi_k}$ as follows:

$$\begin{aligned}\mathbb{T}Q^{\pi_k}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q^{\pi_k}(s', a') \right] \\ &\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\mathbb{E}_{a' \sim \pi_k(\cdot|s')} \left[Q^{\pi_k}(s', a') \right] \right] = Q^{\pi_k}(s,a).\end{aligned}$$

Next let us prove that $Q^{\pi_{k+1}} \geq \mathbb{T}Q^{\pi_k}$. First, let us see that $Q^{\pi_{k+1}} \geq Q^{\pi_k}$:

$$Q^{\pi_k} = r + \gamma P^{\pi_k} Q^{\pi_k} \leq r + \gamma P^{\pi_{k+1}} Q^{\pi_k} \leq \cdots \leq \sum_{t=0}^{\infty} \gamma^t (P^{\pi_{k+1}})^t r = Q^{\pi_{k+1}},$$

where we have used that π_{k+1} is the greedy policy with respect to Q^{π_k} in the first inequality and recursion in the second inequality. Using this,

$$Q^{\pi_{k+1}} = r + \gamma P^{\pi_{k+1}} Q^{\pi_{k+1}} \geq r + \gamma P^{\pi_{k+1}} Q^{\pi_k} = r + \gamma P V_{Q^{\pi_k}} = \mathbb{T}Q^{\pi_k}$$