Markov Decision Processes

Garrett Thomas

April 6, 2020

[redacted]

[redacted]

2 Markov Decision Processes

Markov decision processes (**MDPs**) provide a mathematical framework in which to study discrete-time[1] decision-making problems. Formally, a Markov decision process is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mu_0, T, r, \gamma, H)$, where

1. $\mathcal{S}$ is the **state space**, which contains all possible states the system may be in.
2. $\mathcal{A}$ is the **action space**, which contains all possible actions the agent may take when interacting with the system.
3. $\mu_0 \in \Delta(\mathcal{S})$ is the **initial state distribution**, a probability distribution over states in which the system will be initialized.
4. $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the **transition dynamics**. For each state s and action a, $T(s, a)$ yields a probability distribution over states that the system may transition into when taking action a from state s. We sometimes write the conditional density/mass of transitioning into state s' as $T(s' \mid s, a)$.
5. $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the **reward function**. The value $r(s, a, s')$ gives the amount of "reward" associated transitioning into state s' when taking action a from state s.
6. $\gamma \in [0, 1]$ is the **discount factor**, which determines how much future rewards should be "discounted" when making decisions. A value of $\gamma = 0$ means that we don't care about future rewards at all, while a value of $\gamma = 1$ indicates that rewards in the distant future should count just as much as the reward at the next time-step.
7. H is the **horizon**, the maximum possible number of time-steps in each episode. It may be a positive integer (the **finite-horizon** case) or ∞ (the **infinite-horizon** case).

[1] One can consider continuous-time MDPs as well, but this is rare (or rather, usually falls under the domain of control theory), so we will assume that all MDPs are discrete-time.

Usually $\gamma = 1$ for finite-horizon problems, but for infinite-horizon problems we require $\gamma < 1$ for the purposes of analysis.

An MDP is said to be **finite** if $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$. A finite MDP is usually described as **tabular** if $\mathcal{S}$ and $\mathcal{A}$ are small enough that we can store vectors of dimension $|\mathcal{S} \times \mathcal{A}|$ in a table.

A **policy** π determines what action(s) to take at each state.[2] A **stochastic policy** $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ assigns a distribution over actions to each state. As with the dynamics, we write the action distribution that π assigns to state s as $\pi(s)$, e.g. $a \sim \pi(s)$, but the conditional probability mass/density of action a in state s when executing policy π as $\pi(a \mid s)$. A **deterministic policy** is a function $\pi : \mathcal{S} \to \mathcal{A}$ that directly maps a state to an action to take when in that state. One can (and we generally will) consider deterministic policies as a special case of stochastic policies where every distribution produced by the policy is a point mass (i.e. a Dirac measure) on the prescribed action.[3] If we write $a = \pi(s)$ instead of $a \sim \pi(s)$, we are referring to a deterministic policy.

We consider the common **episodic** setting, in which the system is periodically reset to some (possibly random) initial state. Each episode is assumed to proceed as follows:

(1) Sample the initial state $s_0 \sim \mu_0$.

(2) For $t = 0, 1, 2, \ldots, H$

   (i) Sample an action $a_t \sim \pi(s_t)$ from the policy

   (ii) Sample the next state $s_{t+1} \sim T(s_t, a_t)$ according to the dynamics

   (iii) Observe a reward $r_t = r(s_t, a_t, s_{t+1})$

Note that the distribution of the next state s_{t+1} depends only on the current state s_t and action a_t, rather than on the whole sequence of states and actions encountered up to time t:

$$\forall t, \qquad p(s_{t+1} \mid s_t, a_t) = p(s_{t+1} \mid s_1, a_1, \ldots, s_t, a_t)$$

This is the **Markov property**, hence the "M" in "MDP".

The collection of all states and actions encountered during an episode can be collected into a **trajectory** $\tau = (s_0, a_0, s_1, a_1, \ldots, a_{H-1}, s_H)$. The **return** associated with a trajectory τ is the sum of all discounted rewards

$$R(\tau) := \sum_{t=0}^{H} \gamma^t r_t$$

The expected return of a policy π is denoted

$$\eta(\pi) := \mathbb{E}^\pi[R(\tau)]$$

where $\mathbb{E}^\pi$ means that all actions are sampled according to the policy π. There is also stochasticity coming from the initial state and the dynamics, but this is typically not reflected in the notation unless there is ambiguity.

The standard objective in reinforcement learning is to find a policy which maximizes the expected return:

$$\pi^\star \in \operatorname*{argmax}_{\pi \in \Pi} \eta(\pi)$$

where Π is some family of policies under consideration.

[2] More generally, the policy can also depend on the time t. Policies which do not depend on time (which is what we consider here) are described as **stationary**. It turns out that in infinite-horizon MDPs, it is sufficient to consider stationary policies, but in finite-horizon MDPs stationary policies may not be optimal

[3] These are, of course, not literally the same mathematical objects, but they are effectively the same for our purposes.

3 Value functions

It is often useful to estimate how "good" a given state (and action) is. The standard tools for this are the **Q-function** (which considers a given state and action) and the **value function** (which considers a state only). These are defined (assuming $H = \infty$) as

$$Q^\pi(s,a) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^t r_t \;\middle|\; s_0 = s, a_0 = a\right]$$

$$V^\pi(s) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^t r_t \;\middle|\; s_0 = s\right]$$

In words, $Q^\pi(s,a)$ gives the expected return if you start in state s, take action a, and then act according to π thereafter. Similarly, $V^\pi(s)$ gives the expected return if you start in state s and act according to π.

There is a lot to say about these functions. First, we note the following identities which follow directly from the law of total expectation:

$$\eta(\pi) = \mathop{\mathbb{E}}_{s_0\sim\mu_0}[V^\pi(s_0)]$$

and

$$V^\pi(s) = \mathop{\mathbb{E}}_{a\sim\pi(s)}[Q^\pi(s,a)]$$

Another very useful relationship is the following:

Proposition 1.

$$Q^\pi(s,a) = \mathop{\mathbb{E}}_{s'\sim T(s,a)}[r(s,a,s') + \gamma V^\pi(s')]$$

Proof. The key is to split the infinite sum into the immediate next reward plus the remaining terms. A factor of γ can be pulled out and the index t shifted by one:

$$\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s_{t+1}) \;\middle|\; s_0 = s, a_0 = a\right] \\
&= \mathbb{E}^\pi\left[r(s_0,a_0,s_1) + \gamma\sum_{t=1}^{\infty}\gamma^{t-1} r(s_t,a_t,s_{t+1}) \;\middle|\; s_0 = s, a_0 = a\right] \\
&= \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[r(s,a,s') + \gamma\mathbb{E}^\pi\left[\sum_{t=1}^{\infty}\gamma^{t-1} r(s_t,a_t,s_{t+1}) \;\middle|\; s_1 = s'\right]\right] \\
&= \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[r(s,a,s') + \gamma\mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^{t} r(s_t,a_t,s_{t+1}) \;\middle|\; s_0 = s'\right]\right] \\
&= \mathop{\mathbb{E}}_{s'\sim T(s,a)}[r(s,a,s') + \gamma V^\pi(s')]
\end{aligned}$$

as claimed. □

In fact, the same technique can be used to show a more general statement:

$$Q^\pi(s,a) = \mathbb{E}^\pi\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^{k-1} r_{t+k-1} + \gamma^k V^\pi(s_{t+k}) \;\middle|\; s_t = s, a_t = a\right]$$

3.1 Optimal value functions

Recall that our goal is to find an optimal policy $\pi^\star \in \operatorname{argmax}_\pi \eta(\pi)$. We define the optimal value functions

$$Q^\star(s,a) = \max_\pi Q^\pi(s,a)$$
$$V^\star(s) = \max_\pi V^\pi(s)$$

The optimal value functions are useful for studying the optimality of policies, and for deriving optimal policies.

Proposition 2. *The following are equivalent:*

1. $\pi \in \operatorname{argmax}_\pi \eta(\pi)$
2. *If* $\mathcal{A}^\star(s) = \operatorname{argmax}_a Q^\star(s,a)$ *is the set of optimal actions at state* s*, then* $\operatorname{supp} \pi(s) \subseteq \mathcal{A}^\star(s)$ *for all* $s \in \mathcal{S}$*.*
3. $Q^\pi = Q^\star$
4. $V^\pi = V^\star$

In particular, if we define the **greedy policy** with respect to $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ by

$$\pi_Q(s) = \operatorname*{argmax}_a Q(s,a)$$

then $\pi_{Q^\star}$ is an optimal policy.

We also have the following relationships between $V^\star$ and $Q^\star$.

Proposition 3.

$$\begin{aligned} V^\star(s) &= \max_a Q^\star(s,a) \\ Q^\star(s,a) &= \mathop{\mathbb{E}}_{s'\sim T(s,a)} \left[r(s,a,s') + \gamma V^\star(s')\right] \end{aligned}$$

Proof. For the first identity, we have

$$\begin{aligned} V^\star(s) &= \max_\pi V^\pi(s) \\ &= \max_\pi \mathop{\mathbb{E}}_{a\sim\pi(s)} [Q^\pi(s,a)] \\ &\leq \max_\pi \mathop{\mathbb{E}}_{a\sim\pi(s)} [Q^\star(s,a)] \\ &= \max_a Q^\star(s,a) \end{aligned}$$

In fact, the inequality holds with equality because we can choose π so that simultaneously $\pi \in \operatorname{argmax}_\pi Q^\pi(s,a)$ and $\pi(s) = \operatorname{argmax}_a Q^\star(s,a)$.

For the second identity,

$$\begin{aligned} Q^\star(s,a) &= \max_\pi Q^\pi(s,a) \\ &= \max_\pi \mathop{\mathbb{E}}_{s'\sim T(s,a)} [r(s,a,s') + \gamma V^\pi(s')] \\ &= \mathop{\mathbb{E}}_{s'\sim T(s,a)} \left[r(s,a,s') + \gamma \max_\pi V^\pi(s')\right] \\ &= \mathop{\mathbb{E}}_{s'\sim T(s,a)} \left[r(s,a,s') + \gamma V^\star(s')\right] \end{aligned}$$

□

3.2 Bellman equations and operators

Combining previously shown relationships between the value functions, we can derive the following recurrences, known as **Bellman equations** (after Richard Bellman):

$$\begin{aligned}
Q^\pi(s,a) &= \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[r(s,a,s') + \gamma \mathop{\mathbb{E}}_{a'\sim\pi(s')} Q^\pi(s',a')\right] \\
Q^\star(s,a) &= \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[r(s,a,s') + \gamma \max_{a'} Q^\star(s',a')\right] \\
V^\pi(s) &= \mathop{\mathbb{E}}_{a\sim\pi(s)}\left[\mathop{\mathbb{E}}_{s'\sim T(s,a)}[r(s,a,s') + \gamma V^\pi(s')]\right] \\
V^\star(s) &= \max_a\left[\mathop{\mathbb{E}}_{s'\sim T(s,a)}[r(s,a,s') + \gamma V^\star(s')]\right]
\end{aligned}$$

These observations motivate the following definitions: the **Bellman operators** $\mathcal{B}_q^\pi$, $\mathcal{B}_q^\star$, $\mathcal{B}_v^\pi$, and $\mathcal{B}_v^\star$ are given by

$$\begin{aligned}
\mathcal{B}_q^\pi Q(s,a) &:= \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[r(s,a,s') + \gamma \mathop{\mathbb{E}}_{a'\sim\pi(s')} Q(s',a')\right] \\
\mathcal{B}_q^\star Q(s,a) &:= \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[r(s,a,s') + \gamma \max_{a'} Q(s',a')\right] \\
\mathcal{B}_v^\pi V(s) &:= \mathop{\mathbb{E}}_{a\sim\pi(s)}\left[\mathop{\mathbb{E}}_{s'\sim T(s,a)}[r(s,a,s') + \gamma V(s')]\right] \\
\mathcal{B}_v^\star V(s) &:= \max_a\left[\mathop{\mathbb{E}}_{s'\sim T(s,a)}[r(s,a,s') + \gamma V(s')]\right]
\end{aligned}$$

Note that the Bellman operators map *functions to functions*; specifically, $\mathcal{B}_q^\pi, \mathcal{B}_q^\star : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\mathcal{B}_v^\pi, \mathcal{B}_v^\star : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$. A key point about these operators is that (by definition) the value functions are fixed-points, i.e.

$$Q^\pi = \mathcal{B}_q^\pi Q^\pi \qquad Q^\star = \mathcal{B}_q^\star Q^\star \qquad V^\pi = \mathcal{B}_v^\pi V^\pi \qquad V^\star = \mathcal{B}_v^\star V^\star$$

Another important fact about the Bellman operators is that they are γ-contractions in the sup-norm (also called the ∞-norm), which for a space $\mathbb{R}^{\mathcal{X}}$ is defined as

$$\|f\|_\infty := \sup_{x\in\mathcal{X}} |f(x)|$$

Recall that given a normed space $(N, \|\cdot\|)$, an operator $f : N \to N$ is said to be a κ-contraction if $\kappa \in (0,1)$ and

$$\forall x, y \in N, \qquad \|f(x) - f(y)\| \le \kappa\|x - y\|$$

The **Banach fixed-point theorem** guarantees that if f is a contraction[4], then f has a unique fixed-point $x^* \in N$ satisfying

$$x^* = f(x^*) = \lim_{k\to\infty} f^{\circ k}(x) \qquad \forall x \in N$$

where $f^{\circ k}$ denotes composition of f with itself, i.e. $f^{\circ 0} = \mathrm{id}$, $f^{k+1} = f \circ f^{\circ k}$ for $k > 0$.

[4] and N is a non-empty, complete metric space

Proposition 4. *$\mathcal{B}_q^\pi$ and $\mathcal{B}_q^\star$ are γ-contractions on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. $\mathcal{B}_v^\pi$ and $\mathcal{B}_v^\star$ are γ-contractions on $\mathbb{R}^{\mathcal{S}}$.*

Proof. Suppose $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $V, V' \in \mathbb{R}^{\mathcal{S}}$.

We consider the $\mathcal{B}^\pi$ versions first. When we write out $(\mathcal{B}_q^\pi Q - \mathcal{B}_q^\pi Q')(s,a) = \mathcal{B}_q^\pi Q(s,a) - \mathcal{B}_q^\pi Q'(s,a)$, the expected reward cancels and, by linearity, we are left with

$$(\mathcal{B}_q^\pi Q - \mathcal{B}_q^\pi Q')(s,a) = \gamma \mathop{\mathbb{E}}_{s'\sim T(s,a)}\left[\mathop{\mathbb{E}}_{a'\sim\pi(s')}[(Q-Q')(s',a')]\right]$$

Then

$$\begin{aligned}\|\mathcal{B}_q^\pi Q - \mathcal{B}_q^\pi Q'\|_\infty &= \sup_{s,a} |(\mathcal{B}_q^\pi Q - \mathcal{B}_q^\pi Q')(s,a)| \\ &= \gamma \sup_{s,a} \left|\mathop{\mathbb{E}}_{s',a'}[(Q-Q')(s',a')]\right| \\ &\leq \gamma \sup_{s,a} \mathop{\mathbb{E}}_{s',a'} |(Q-Q')(s',a')| \\ &\leq \gamma \underbrace{\sup_{s',a'} |(Q-Q')(s',a')|}_{\|Q-Q'\|_\infty}\end{aligned}$$

as claimed. Similarly,

$$(\mathcal{B}_v^\pi V - \mathcal{B}_v^\pi V')(s) = \gamma \mathop{\mathbb{E}}_{a\sim\pi(s)}\left[\mathop{\mathbb{E}}_{s'\sim T(s,a)}[(V-V')(s')]\right]$$

so

$$\|\mathcal{B}_v^\pi V - \mathcal{B}_v^\pi V'\|_\infty \leq \gamma \sup_s \mathop{\mathbb{E}}_{a,s'} |(V-V')(s')| \leq \gamma \underbrace{\sup_{s'} |(V-V')(s')|}_{\|V-V'\|_\infty}$$

The $\mathcal{B}^\star$ versions are slightly more involved because the max makes things nonlinear. We use the following lemma:

$$|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$$

To see this, suppose $\max_x f(x) > \max_x g(x)$ (the other case is symmetric) and let $\tilde{x} = \operatorname{argmax}_x f(x)$. Then

$$|\max_x f(x) - \max_x g(x)| = f(\tilde{x}) - \max_x g(x) \leq f(\tilde{x}) - g(\tilde{x}) \leq \max_x |f(x) - g(x)|$$

With this established, we see that

$$\begin{aligned}\|\mathcal{B}_q^\star Q - \mathcal{B}_q^\star Q'\|_\infty &= \sup_{s,a} |(\mathcal{B}_q^\star Q - \mathcal{B}_q^\star Q')(s,a)| \\ &= \gamma \sup_{s,a} \left|\mathop{\mathbb{E}}_{s'}\left[\max_{a'} Q(s',a') - \max_{a''} Q'(s',a'')\right]\right| \\ &\leq \gamma \sup_{s,a} \mathop{\mathbb{E}}_{s'} \left|\max_{a'} Q(s',a') - \max_{a''} Q'(s',a'')\right| \\ &\leq \gamma \sup_{s,a} \mathop{\mathbb{E}}_{s'} \max_{a'} |Q(s',a') - Q'(s',a')| \qquad \text{(lemma)} \\ &\leq \gamma \underbrace{\sup_{s',a'} |(Q-Q')(s',a')|}_{\|Q-Q'\|_\infty}\end{aligned}$$

Finally,

$$
\begin{aligned}
\|\mathcal{B}_v^\star V - \mathcal{B}_v^\star V'\|_\infty &= \sup_s |\mathcal{B}_v^\star V(s) - \mathcal{B}_v^\star V'(s)| \\
&= \sup_s \left| \max_a \mathbb{E}_{s'}[r(s,a,s') + \gamma V(s')] - \max_a \mathbb{E}_{s'}[r(s,a,s') + \gamma V'(s')] \right| \\
&\le \sup_s \max_a \left| \big(\mathbb{E}_{s'}[r(s,a,s') + \gamma V(s')]\big) - \big(\mathbb{E}_{s'}[r(s,a,s') + \gamma V'(s')]\big) \right| \qquad \text{(lemma)} \\
&= \gamma \sup_s \max_a \left| \mathbb{E}_{s'}[(V - V')(s')] \right| \\
&\le \gamma \sup_s \max_a \mathbb{E}_{s'} |(V - V')(s')| \\
&\le \gamma \underbrace{\sup_{s'} |(V - V')(s')|}_{\|V - V'\|_\infty}
\end{aligned}
$$

as was to be shown. □

As a consequence, the value functions can be obtained by iterating their corresponding Bellman equations: for any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,

$$\lim_{k\to\infty} (\mathcal{B}_q^\pi)^{\circ k} Q = Q^\pi \qquad\qquad \lim_{k\to\infty} (\mathcal{B}_q^\star)^{\circ k} Q = Q^\star$$

and for any $V \in \mathbb{R}^{\mathcal{S}}$,

$$\lim_{k\to\infty} (\mathcal{B}_v^\pi)^{\circ k} V = V^\pi \qquad\qquad \lim_{k\to\infty} (\mathcal{B}_v^\star)^{\circ k} V = V^\star$$

Algorithm 1 Q-Value Iteration

```
Initialize $Q_0$ arbitrarily, e.g. $Q_0 = 0$
for all $k = 0, 1, 2, \dots$ do
    for all $s \in \mathcal{S}, a \in \mathcal{A}$ do
        $Q_{k+1}(s, a) = \sum_{s'} T(s' \mid s, a)\big(r(s, a, s') + \gamma \max_{a'} Q_k(s', a')\big)$
    end for
end for
```

Algorithm 2 Value Iteration

```
Initialize $V_0$ arbitrarily, e.g. $V_0 = 0$
for all $k = 0, 1, 2, \dots$ do
    for all $s \in \mathcal{S}$ do
        $V_{k+1}(s) = \max_a \sum_{s'} T(s' \mid s, a)\big(r(s, a, s') + \gamma V_k(s')\big)$
    end for
end for
```

4 Dynamic programming for solving tabular MDPs

Now we discuss fundamental algorithms for solving tabular MDPs when the reward function and transition probabilities are known. In this setting, the Bellman operators can be computed exactly.

4.1 (Q-)Value iteration

Q-value iteration (Algorithm 1) is the algorithm defined by iteratively applying $\mathcal{B}_q^\star$:

$$Q_{k+1} = \mathcal{B}_q^\star Q_k$$

where Q_0 is an arbitrary initialization. As previously mentioned, we have $\lim_{k\to\infty} Q_k = Q^\star$. Then an optimal policy can be derived from $Q^\star$.

Value iteration (Algorithm 2) is the algorithm obtained by iteratively applying $\mathcal{B}_v^\star$:

$$V_{k+1} = \mathcal{B}_v^\star V_k$$

where V_0 is an arbitrary initialization. As previously mentioned, we have $\lim_{k\to\infty} V_k = V^\star$. Then an optimal policy can be derived by

$$\pi_{Q^\star}(s) = \operatorname*{argmax}_a Q^\star(s, a) = \operatorname*{argmax}_a \sum_{s'} T(s' \mid s, a)\big(r(s, a, s') + \gamma V^\star(s')\big)$$

For both algorithms, the computational complexity of each iteration is $O(|\mathcal{S}|^2|\mathcal{A}|)$.

However, we cannot not run $k \to \infty$ in practice, so it is of interest to understand the performance of the greedy policy with respect to an imperfect value function.

Proposition 5. *For any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

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

For proof, see [1].

Corollary 1 (Convergence of Q-value iteration)**.** *Suppose the rewards are bounded such that $|r| \leq r_{\max}$, and let $\epsilon > 0$. Then if we initialize $Q_0 = 0$, then for $k \geq \frac{1}{1-\gamma} \log \frac{2r_{\max}}{\epsilon(1-\gamma)^2}$,*[5] *the greedy policy $\pi_k := \pi_{Q_k}$ satisfies*

$$V^{\pi_k} \geq V^\star - \epsilon$$

[5] Note that the bound is slightly more general than in [1] because they assume $r_{\max} = 1$. Also our version has an extra $1 - \gamma$ factor in the denominator inside the log because we do not normalize the value functions.

Algorithm 3 Policy Iteration

```
Initialize $\pi_0$ arbitrarily
for all $k = 0, 1, 2, \dots$ do
    Compute $Q^{\pi_k}$
    for all $s$ do
        $\pi_{k+1}(s) = \operatorname{argmax}_a Q^{\pi_k}(s, a)$
    end for
end for
```

Proof. First note that since the rewards are bounded by $r_{\max}$,

$$\|Q^\star\|_\infty \le \sum_{t=0}^{\infty} \gamma^t |r_t| \le \frac{r_{\max}}{1-\gamma}$$

Applying the contraction property and the fact that $Q^\star = \mathcal{B}_q^\star Q^\star$, we obtain

$$\|Q_k - Q^\star\|_\infty = \|(\mathcal{B}_q^\star)^{\circ k} Q_0 - (\mathcal{B}_q^\star)^{\circ k} Q^\star\|_\infty \le \gamma^k \|Q_0 - Q^\star\|_\infty = \gamma^k \|Q^\star\|_\infty \le \frac{\gamma^k r_{\max}}{1-\gamma}$$

Since $1 - x \le e^{-x}$ for any x, we have

$$\gamma^k = (1 - (1-\gamma))^k \le (e^{-(1-\gamma)})^k \le e^{-(1-\gamma)k}$$

Combining this with the previous bound yields

$$V^{\pi_{Q_k}} \ge V^\star - 2\frac{\|Q_k - Q^\star\|_\infty}{1-\gamma} \ge V^\star - \frac{2r_{\max}}{(1-\gamma)^2} e^{-(1-\gamma)k}$$

Thus it suffices to have

$$\frac{2r_{\max}}{(1-\gamma)^2} e^{-(1-\gamma)k} \le \epsilon$$

Solving for k gives the stated result. □

A similar bound can be obtained for value iteration.

4.2 Policy iteration

The previously discussed algorithms aim to approximate the optimal value functions directly. **Policy iteration** (Algorithm 3) is a different approach based on two alternating steps:

1. **Policy evaluation**: compute Q^{π_k}, the value of the current policy π_k
2. **Policy improvement**: update the policy as $\pi_{k+1} = \pi_{Q^{\pi_k}}$, the greedy policy with respect to the Q function of the previous policy

4.2.1 Policy evaluation

The policy evaluation step can be done in more than one way. As previously mentioned, Q^π is the unique fixed-point of $\mathcal{B}_q^\pi$, so one can approximate Q^π by iterating $\mathcal{B}_q^\pi$, analogous to Q-value iteration.

Alternatively, we can find Q^π exactly by solving a linear system. In a tabular MDP, one can view Q^π as a vector $\mathbf{q}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Define a vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ whose (s, a)th element is $\mathbb{E}_{s'} r(s, a, s')$, and a matrix $\mathbf{T}^\pi$ with entries

$$T^\pi_{(s,a),(s',a')} = \begin{cases} T(s' \mid s, a) & \text{if } a' = \pi(s') \\ 0 & \text{otherwise} \end{cases}$$

Then we have

$$\mathbf{q}^\pi = \mathbf{r} + \gamma \mathbf{T}^\pi \mathbf{q}^\pi$$

so

$$\mathbf{q}^\pi = (\mathbf{I} - \gamma \mathbf{T}^\pi)^{-1} \mathbf{r}$$

We are guaranteed that this matrix is invertible because for any $\mathbf{x} \neq \mathbf{0}$,

$$\begin{aligned} \|(\mathbf{I} - \gamma \mathbf{T}^\pi)\mathbf{x}\|_\infty &= \|\mathbf{x} - \gamma \mathbf{T}^\pi \mathbf{x}\|_\infty \\ &\geq \|\mathbf{x}\|_\infty - \gamma \|\mathbf{T}^\pi \mathbf{x}\|_\infty \\ &\geq \|\mathbf{x}\|_\infty - \gamma \|\mathbf{x}\|_\infty \\ &> 0 \end{aligned}$$

which implies $(\mathbf{I} - \gamma \mathbf{T}^\pi)\mathbf{x} \neq \mathbf{0}$.

Yet another way is to compute V^π (using $\mathcal{B}_v^\pi$ or an analogous linear system) and then derive Q^π from that.

4.2.2 Policy improvement

The name "policy improvement" is justified by the following facts.

Proposition 6 (Policy improvement)**.** *For all k,*

1. *The value function is non-decreasing: $V^{\pi_{k+1}} \geq V^{\pi_k}$*
2. *The Q-function is non-decreasing: $Q^{\pi_{k+1}} \geq \mathcal{B}_q^\star Q^{\pi_k} \geq Q^{\pi_k}$*
3. *The Q-function makes progress towards $Q^\star$: $\|Q^{\pi_{k+1}} - Q^\star\|_\infty \leq \gamma \|Q^{\pi_k} - Q^\star\|_\infty$*
4. *If $\pi_{k+1} = \pi_k$ for some k, then π_k is optimal.*

Proof. We begin with the first claim. For any state s:

$$\begin{aligned} V^{\pi_k}(s) &= Q^{\pi_k}(s, \pi_k(s)) \\ &\leq \max_a Q^{\pi_k}(s, a) \\ &= Q^{\pi_k}(s, \pi_{k+1}(s)) \\ &= \mathop{\mathbb{E}}_{s' \sim T(s, \pi_{k+1}(s))} [r(s, \pi_{k+1}(s), s') + \gamma V^{\pi_k}(s')] \\ &\leq \mathop{\mathbb{E}}_{s' \sim T(s, \pi_{k+1}(s))} \left[r(s, \pi_{k+1}(s), s') + \gamma \mathop{\mathbb{E}}_{s'' \sim T(s', \pi_{k+1}(s'))} [r(s', \pi_{k+1}(s'), s'') + \gamma V^{\pi_k}(s'')] \right] \\ &\vdots \quad \text{(repeat this process until every action is sampled from } \pi_{k+1}) \\ &\leq V^{\pi_{k+1}}(s) \end{aligned}$$

Having established the first claim, the second claim follows readily:

$$\begin{aligned} Q^{\pi_{k+1}}(s, a) &= \mathop{\mathbb{E}}_{s' \sim T(s,a)} \left[r(s, a, s') + V^{\pi_{k+1}}(s')) \right] \\ &\geq \mathop{\mathbb{E}}_{s' \sim T(s,a)} \left[r(s, a, s') + \max_{a'} Q^{\pi_k}(s', a') \right] = \mathcal{B}_q^\star Q^{\pi_k}(s, a) \\ &\geq \mathop{\mathbb{E}}_{s' \sim T(s,a)} \left[r(s, a, s') + Q^{\pi_k}(s', \pi_k(s')) \right] \\ &= Q^{\pi_k}(s, a) \end{aligned}$$

For the third claim,

$$\begin{aligned}
\|Q^{\pi_{k+1}} - Q^\star\|_\infty &\leq \|\mathcal{B}_q^\star Q^{\pi_k} - Q^\star\|_\infty && (\mathcal{B}_q^\star Q^{\pi_k} \leq Q^{\pi_{k+1}} \leq Q^\star) \\
&= \|\mathcal{B}_q^\star Q^{\pi_k} - \mathcal{B}_q^\star Q^\star\|_\infty && (Q^\star = \mathcal{B}_q^\star Q^\star) \\
&\leq \gamma \|Q^{\pi_k} - Q^\star\|_\infty && (\mathcal{B}_q^\star \text{ is } \gamma\text{-contraction})
\end{aligned}$$

For the final claim, observe that if $\pi_{k+1} = \pi_k$, then $Q^{\pi_{k+1}} = Q^{\pi_k}$, so we must have $\mathcal{B}_q^\star Q^{\pi_k} = Q^{\pi_k}$, which implies $Q^{\pi_k} = Q^\star$. □

Since there are only $|\mathcal{A}|^{|\mathcal{S}|}$ distinct deterministic policies, policy iteration will converge in at most this many iterations.

References

[1] Alekh Agarwal, Nan Jiang, Sham M. Kakade. *Reinforcement Learning: Theory and Algorithms.*