Markov Decision Process

Sunmook Choi felixchoi@korea.ac.kr

January 5, 2024

1 Markov Decision Processes

Definition 1 (Markov Decision Process, MDP)**.** *MDP is a tuple* $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ *where*

- $\mathcal{S}$ *is the state space,*
- $\mathcal{A}$ *is the action space,*
- P *is the transition probability,*
  - *the transition probability from the state* $s \in \mathcal{S}$ *to the state* $s' \in \mathcal{S}$ *given an action* $a \in \mathcal{A}$*.*
  - $P^a_{ss'} := P(S_{t+1} = s' | S_t = s, A_t = a)$
  - $P(S_{t+1} = s' | S_t = s) = P(S_{t+1} = s' | S_t = s, S_{t-1} = s_{t-1}, \ldots, S_1 = s_1, S_0 = s_0)$
  - *We only consider stationary Markov Processes, that is,* $P^a_{ss'}$ *does not depend on time step* t*.*
- $\mathcal{R}$ *is the reward function, and*
- $\gamma \in [0, 1]$ *is the discount factor.*

Example 1 (Grid World)**.** Let's consider an example, the Grid World, illustrated in Fig. 1.1. We call a robot agent which is moving in the grid world. The thing it interacts with, comprising

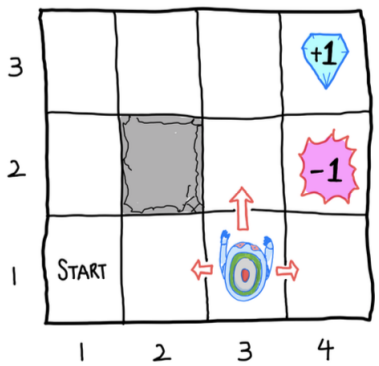

Figure 1.1: An example of Grid world Environment

everything outside the agent, is called environment. There are total 11 places that an agent can stay

in, $(1,1),(2,1),\ldots,(4,3)$, except for $(2,2)$. Such places are called *states* of the grid world. The agent starts the game from the state $(1,1)$, and the game ends when it reaches one of the final states, $(4,2)$ or $(4,3)$. The agent can move either North, South, East, or West. The agent should take an *action* in order to move; the actions are North, South, East, and West. However, the grid world is *slippery* so that the next state is determined stochastically. The transition occurs corresponding to the transition probability $P^a_{ss'}$, which has the Markov Property, and an example is as follows:

$$P(s_{t+1} = (3,2)|s_t = (3,1), a_t = North) = 0.8$$
$$P(s_{t+1} = (2,1)|s_t = (3,1), a_t = North) = 0.1$$
$$P(s_{t+1} = (4,1)|s_t = (3,1), a_t = North) = 0.1.$$

Here, if the state transition is impossible due to the wall of the grid world, the agent stays in the current state. For each step, the agent gets reward from the environment. The agent gets small negative reward c (*e.g.*, $c = -0.1$) at each step until it reaches the final state. This encourages the agent to reach the final state as soon as possible. The agent gets $+1$ reward when it reaches the state $(4,3)$, and gets -1 reward when it reaches $(4,2)$. The description above defines the state space, the action space, the transition probability, and the reward function. With a hyperparameter γ, the grid world is a Markov Decision Process.

The general interaction between an agent and an environment can be illustrated as the figure below.

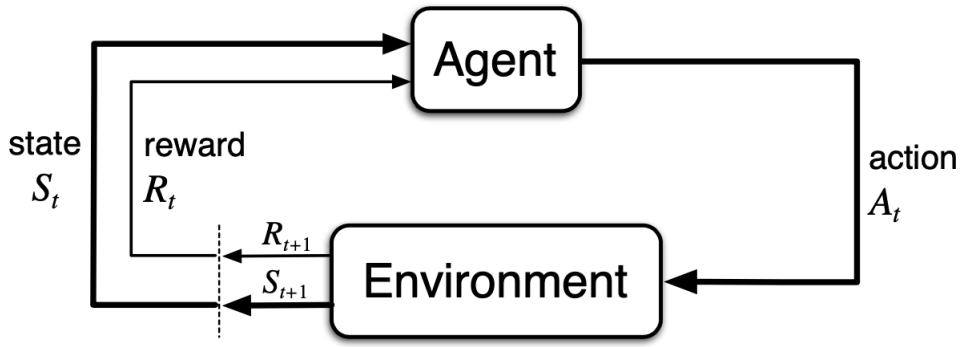

Figure 1.2: The agent-environment interaction in a Markov Decision Process.

Based on environments, the rewards can also be randomly given according to the current state or the chosen action. In this sense, we generalize the transition probabilities into *dynamics*, which we denote by $p(s', r|s, a)$.

The agent and environment interact at each of a sequence of discrete time steps, $t = 0, 1, 2, \ldots$. At each time step t, the agent receives its state $S_t \in \mathcal{S}$ and takes action $A_t \in \mathcal{A}$. In part as a consequence of its action, the agent receives a reward $R_{t+1} \in \mathcal{R}$, and finds itself in a new state S_{t+1}. The MDP and agent together give rise to a stochastic process or trajectory as follows:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$

Based on the game, the episode can be either finite or infinite. If the state space of an MDP has a final state, then the episode terminates at the final state. In such cases, we usually write the trajectory as follows:

$$S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$$

We also call the whole sequence an episode of the game/environment if it terminates at some time step. Notice that the states, actions, and rewards are written as random variables due to the dynamics and policy (we discuss about policy right in the next section). In this paper, we will assume that the state space and the action space is finite, that is, $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$.

2 Goal of the agent in an environment

The goal of an agent is to maximize the cumulative sum of rewards during the game. In this sense, given an MDP, the agent aims to **find a rule that chooses an action at each state.** We call such a rule the policy, which can be defined mathematically as a function π from the state space $\mathcal{S}$ to the action space $\mathcal{A}$.

The policy $\pi \colon \mathcal{S} \to \mathcal{A}$ can be either stochastic or deterministic. For example, in the grid world, a policy at the state (1,1) can be defined as follows:

$$\pi(a|s) = P(A_t = a|S_t = s) = \begin{cases} 0.6 & (a = North) \\ 0.4 & (a = East) \\ 0 & otherwise \end{cases} \tag{2.1}$$

The stochastic policy is indeed a conditional probability function. One should be able to discriminate two probabilities, the **transition probability** and the **policy**. Given a state $S_t = s$, the agent first chooses an action a according to the policy $\pi(\cdot|s)$. After it chooses an action, the state transition occurs according to the transition probability.

The agent aims to find an optimal policy that maximizes the cumulative sum of rewards in an episode. This idea can be stated informally as the reward hypothesis.

Remark 1 (Reward Hypothesis)**.** *That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of rewards.*

We must stay tuned to the idea that *the agent is not trying to maximize immediate reward* **but the cumulative sum of rewards in the long run**. We define the cumulative sum of rewards formally as the return or the discounted return.

Definition 2 (Return)**.** *At each time step t, the return G_t is defined to be*

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots . \tag{2.2}$$

More generally, with the discount factor $\gamma \in [0, 1]$ given in an MDP, the discounted return is defined to be

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2.3}$$

Note that the return is also a random variable. The discount rate determines the present value of future rewards: a reward received k time steps in the future is worth only γ^{k-1} times what it would be worth if it were received immediately. We usually use the discounted return for the following reasons.

1. It is mathematically convenient. If the rewards are bounded, then the discounted return converges.
2. The future is uncertain. That is, the agents do not know how many steps they should go to earn a specific future reward.
3. For agents, immediate rewards may earn more interest than delayed rewards.

Now we formally state the goal of the agents. The agent finds an *optimal policy* that maximizes the expected return $\mathbb{E}_{\tau\sim\pi}[G_0(\tau)] = \int G_0(\tau)\, \pi(\tau)\, d\tau$, where $G_0(\tau)$ is the total discounted return of a random terminating trajectory $\tau = (s_0, a_0, r_1, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, r_T, s_T)$ and

$$\pi(\tau) = p(s_0, a_0, r_1, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, r_T, s_T).$$

Because of the Markov property, $\pi(\tau)$ can be expressed as follows:

$$\pi(\tau) = p(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t)\, p(s_{t+1}|s_t, a_t) \tag{2.4}$$

To summary, the agents seeks an optimal policy π^* which is equal to

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}[G_0(\tau)]. \tag{2.5}$$

3 Value Functions

Instead of maximizing the expected discounted return directly, we consider a method that assigns some scalar value for each state or each state-action pair. The value is determined by the expected return that an agent receives in the environment starting from the state (and taking an action). We will use these scalar values, called value functions, to find an optimal policy.

Definition 3 (Value Function)**.** *Given an MDP and a policy π of an agent, the state-value function $v_\pi(s)$ of a state $s \in \mathcal{S}$ is defined to be the expected return when starting in s and following π thereafter. Formally,*

$$v_\pi(s) = \mathbb{E}_\pi\big[G_t|S_t = s\big]. \tag{3.1}$$

Here, $\mathbb{E}_\pi[\cdot]$ denotes the expected value of a random variable given that the agent follows policy π, and t is any time step. The action-value function $q_\pi(s, a)$ of taking action $a \in \mathcal{A}$ in a state $s \in \mathcal{S}$

is defined to be the expected return starting from s, taking the action a, and thereafter following π. Formally,

$$q_\pi(s,a) = \mathbb{E}_\pi\big[G_t | S_t = s, A_t = a\big] \tag{3.2}$$

Note that, by the law of total probability, we have

$$v_\pi(s) = \mathbb{E}_\pi\big[G_t | S_t = s\big] = \sum_{a\in\mathcal{A}} \pi(a|s)\, \mathbb{E}_\pi\big[G_t | S_t = s, A_t = a\big] = \sum_{a\in\mathcal{A}} \pi(a|s)\, q_\pi(s,a). \tag{3.3}$$

If an initial state s_0 is fixed in a given environment, then Eq. (2.5) can also be expressed as

$$\pi^* = \arg\max_\pi v_\pi(s_0). \tag{3.4}$$

A fundamental property of value functions that is used throughout reinforcement learning and dynamic programming is that they satisfy recursive relationships. The recursive formulae are called the Bellman expectation equations.

Theorem 4 (Bellman Expectation Equations)**.** *The value functions can be decomposed into the immediate reward R_{t+1} and the discounted successor value function, $\gamma v_\pi(S_{t+1})$ or $\gamma q_\pi(S_{t+1}, A_{t+1})$.*

$$v_\pi(s) = \mathbb{E}_\pi\big[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\big] \tag{3.5}$$

$$q_\pi(s,a) = \mathbb{E}_\pi\big[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\big] \tag{3.6}$$

Proof. First we recall the definitions of value functions in Eq. 3.1 and Eq. 3.2.

$$v_\pi(s) = \mathbb{E}_\pi\big[G_t \mid S_t = s\big] \tag{3.1}$$

$$q_\pi(s,a) = \mathbb{E}_\pi\big[G_t \mid S_t = s, A_t = a\big] \tag{3.2}$$

Starting from their definitions, we will derive the Bellman expectation equations. We first show the

Bellman expectation equation for state-value function.

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi\big[G_t \mid S_t = s\big] && (3.1)\\
&= \sum_{a\in\mathcal{A}} \pi(a|s)\, q_\pi(s,a) && (3.3)\\
&= \sum_{a\in\mathcal{A}} \pi(a|s)\, \mathbb{E}_\pi\big[G_t | S_t = s, A_t = a\big] && (3.7)\\
&= \sum_{a\in\mathcal{A}} \pi(a|s)\, \mathbb{E}_\pi\big[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a\big] && (3.8)\\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s',r} \mathbb{E}_\pi\big[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a, R_{t+1} = r, S_{t+1} = s'\big] \\
&\qquad\qquad \times P(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) && (3.9)\\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s',r} p(s', r \mid s, a)\Big[r + \gamma \mathbb{E}_\pi\big[G_{t+1} \mid S_{t+1} = s'\big]\Big] && (3.10)\\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')] && (3.11)\\
&= \sum_{a\in\mathcal{A}} \sum_{s',r} p(a, s', r \mid s)[r + \gamma v_\pi(s')] && (3.12)\\
&= \sum_{s',r} p(s', r \mid s)[r + \gamma v_\pi(s')] && (3.13)\\
&= \mathbb{E}_\pi\big[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\big] && (3.5)
\end{aligned}
$$

Notice that Eq. 3.10 is obtained by Markov property. Next we show the Bellman expectation equation for the action-value function.

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi\big[G_t \mid S_t = s, A_t = a\big] && (3.2)\\
&= \mathbb{E}_\pi\big[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\big] && (3.14)\\
&= \sum_{r,s',a'} p(r, s', a' \mid s, a) \cdot \mathbb{E}_\pi\big[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a, R_{t+1} = r, S_{t+1} = s', A_{t+1} = a'\big] \\
& && (3.15)\\
&= \sum_{r,s',a'} p(r, s', a' \mid s, a) \cdot \Big[r + \gamma \mathbb{E}_\pi\big[G_{t+1} \mid S_{t+1} = s', A_{t+1} = a'\big]\Big] && (3.16)\\
&= \sum_{r,s',a'} p(r, s', a' \mid s, a) \cdot \Big[r + \gamma q_\pi(s', a')\Big] && (3.17)\\
&= \mathbb{E}_\pi\big[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\big] && (3.6)
\end{aligned}
$$

This completes the proof. □

We would like to note that, from Eq. 3.3 to Eq. 3.11, we have

$$q_\pi(s,a) = \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')]. \tag{3.18}$$

This equation tells you that the action-value function at (s,a) is the expectation of sum of the immediate reward r and the discounted state-value function at a successor state s'. Here, the expectation is computed over the dynamics of the environment.

4 Optimal Policies and Optimal Value Functions

For finite MDPs, we can precisely define an optimal policy in the following way. Value functions define a partial ordering over policies. If we say that $\pi \geq \pi'$ if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in \mathcal{S}$, then the relation $\geq$ on the set of policies becomes a partial order. We also say that π_* is an optimal policy if and only if $\pi_* \geq \pi$ for any policy π. We will show that *there is always at least one optimal policy* in Theorem 8. First, let's see a property that optimal policies have.

Proposition 5. *If π_* is an optimal policy, then for all state $s \in \mathcal{S}$,*

$$v_{\pi_*}(s) = \max_{a \in \mathcal{A}} q_{\pi_*}(s, a). \tag{4.1}$$

Furthermore, we obtain that

$$v_{\pi_*}(s) = \max_{a \in \mathcal{A}} \sum_{s', r} p(s', r \mid s, a)[r + \gamma v_{\pi_*}(s')]. \tag{4.2}$$

Proof. Observe that, for all $s \in \mathcal{S}$,

$$v_{\pi_*}(s) = \sum_{a \in \mathcal{A}} \pi_*(a|s)\, q_{\pi_*}(s, a) \leq \max_{a \in \mathcal{A}} q_{\pi_*}(s, a). \tag{4.3}$$

For contradiction, assume that there is a state s^* such that

$$v_{\pi_*}(s^*) < \max_{a \in \mathcal{A}} q_{\pi_*}(s^*, a) = q_{\pi_*}(s^*, a^*) \tag{4.4}$$

where $a^* = \arg\max_{a \in \mathcal{A}} q_{\pi_*}(s^*, a)$. Consider the policy π' given by

$$\pi'(a|s) = \begin{cases} \pi_*(a|s) & \text{if } s \neq s^* \\ \begin{cases} 1 & \text{if } a = a^* \\ 0 & \text{otherwise} \end{cases} & \text{if } s = s^* \end{cases}. \tag{4.5}$$

This implies that $q_{\pi_*}(s^*, a^*) = \sum_{a \in \mathcal{A}} \pi'(a|s^*)\, q_{\pi_*}(s^*, a)$. From Eq. 4.4, we have the following:

$$\begin{aligned} v_{\pi_*}(s^*) &< \sum_{a \in \mathcal{A}} \pi'(a|s^*)\, q_{\pi_*}(s^*, a) \\ &= \sum_{a \in \mathcal{A}} \pi'(a|s^*) \sum_{s', r} p(s', r|s^*, a)[r + \gamma v_{\pi_*}(s')] \\ &= \mathbb{E}_{\pi'}\big[R_{t+1} + \gamma v_{\pi_*}(S_{t+1}) \mid S_t = s^*\big] \end{aligned} \tag{4.6}$$

Here, the last equality is obtained by Eq. 3.18. Next, we consider the term $v_{\pi_*}(S_{t+1})$. Since the state random variable S_{t+1} may be the state s^*, we obtain the following inequality:

$$v_{\pi_*}(S_{t+1}) = \sum_{a \in \mathcal{A}} \pi_*(a|S_{t+1})\, q_{\pi_*}(S_{t+1}, a) \leq \sum_{a \in \mathcal{A}} \pi'(a|S_{t+1})\, q_{\pi_*}(S_{t+1}, a). \tag{4.7}$$

Similar to the previous process in Eq. 4.6, we also have

$$v_{\pi_*}(S_{t+1}) \leq \sum_{a \in \mathcal{A}} \pi'(a|S_{t+1})\, q_{\pi_*}(S_{t+1}, a) = \mathbb{E}_{\pi'}\big[R_{t+2} + \gamma v_{\pi_*}(S_{t+2}) \mid S_{t+1}\big]. \tag{4.8}$$

Therefore, we have the following:

$$v_{\pi_*}(s^*) < \mathbb{E}_{\pi'}\big[R_{t+1} + \gamma v_{\pi_*}(S_{t+1}) \mid S_t = s^*\big] \tag{4.6}$$

$$\leq \mathbb{E}_{\pi'}\Big[R_{t+1} + \gamma \mathbb{E}_{\pi'}\big[R_{t+2} + \gamma v_{\pi_*}(S_{t+2}) \mid S_{t+1} = s'\big] \mid S_t = s^*\Big] \tag{4.9}$$

$$= \mathbb{E}_{\pi'}\big[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_{\pi_*}(S_{t+2}) \mid S_t = s^*\big] \tag{4.10}$$

By continuing this process, we will obtain the inequality:

$$v_{\pi_*}(s^*) < \mathbb{E}_{\pi'}\big[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s^*\big] = v_{\pi'}(s^*) \tag{4.11}$$

This contradicts to the assumption that π_* is an optimal policy. Therefore, we conclude that, for all state $s \in \mathcal{S}$, $v_{\pi_*}(s) = \max_{a\in\mathcal{A}} q_{\pi_*}(s,a)$. Furthermore, by Eq. 3.18, we also obtain

$$v_{\pi_*}(s) = \max_{a\in\mathcal{A}} \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_{\pi_*}(s')].$$

This completes the proof. □

Proposition 5 is also called the Bellman operator equations. This will be again introduced after we define what optimal value functions are.

Definition 6 (Optimal Value Function)**.** *The optimal state-value function v_* and the optimal action-value function q_* is defined as*

$$v_*(s) = \max_{\pi} v_\pi(s), \quad \forall s \in \mathcal{S}, \tag{4.12}$$

$$q_*(s,a) = \max_{\pi} q_\pi(s,a), \quad \forall s \in \mathcal{S},\ \forall a \in \mathcal{A}. \tag{4.13}$$

Proposition 7. *For any two optimal policies π_1 and π_2, we have $v_{\pi_1}(s) = v_{\pi_2}(s)$ for all $s \in \mathcal{S}$.*

Proof. Since π_1 is optimal, we have $v_{\pi_1}(s) \geq v_{\pi_2}(s)$ for all $s \in \mathcal{S}$. Likewise, since π_2 is optimal, we have $v_{\pi_2}(s) \geq v_{\pi_1}(s)$ for all $s \in \mathcal{S}$. This shows that $v_{\pi_1}(s) = v_{\pi_2}(s)$ for all $s \in \mathcal{S}$. □

By Proposition 7, we can see that all the optimal policies share the same state-value function. According to the definition of optimal policies, for any optimal policy π_*, we can see that

$$v_{\pi_*}(s) = \max_{\pi} v_\pi(s) = v_*(s)$$

for all state $s \in \mathcal{S}$. However, since the maximization in the definition of optimal value functions is done independently of each state, the optimal policies $\arg\max_\pi v_\pi(s)$ and $\arg\max_\pi v_\pi(s')$ may not be the same if $s \neq s'$. In other words, an optimal policy that maximizes the state value function at a state s may not be the same policy that maximizes the state value function at another state s'. With this caution in mind, consider the following theorem.

Theorem 8 (Optimal Policy and Optimal Value Functions)**.** *Any finite MDP satisfies the following.*

1. *There exists an optimal policy $\pi_* \geq \pi$ for all π.*

2. *All optimal policies achieve the optimal state-value function* $v_{\pi_*}(s) = v_*(s)$ *for all* $s \in \mathcal{S}$.

3. *All optimal policies achieve the optimal action-value function* $q_{\pi_*}(s,a) = q_*(s,a)$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Proof. Assuming the existence of optimal policies, the second and the third statements are immediate results by Proposition 7. Fix a state $s \in \mathcal{S}$. Choose an optimal policy π_s such that

$$\pi_s = \arg\max_\pi v_\pi(s).$$

Then, for any optimal policy π_*, we have $v_{\pi_*}(s) = v_{\pi_s}(s)$ by Proposition 7. Hence,

$$v_*(s) = \max_\pi v_\pi(s) = v_{\pi_s}(s) = v_{\pi_*}(s).$$

This shows that the second statement holds. Now we prove the third statement. For an optimal π_*,

$$\begin{aligned}
q_{\pi_*}(s,a) &= \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma v_{\pi_*}(s')\big] && \text{(by Eq. 3.18)} \\
&= \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma v_*(s')\big] \\
&= \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma \max_\pi v_\pi(s')\big] \\
&= \max_\pi \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma v_\pi(s')\big] && (\because |\mathcal{S}| < \infty,\ |\mathcal{A}| < \infty) \\
&= \max_\pi q_\pi(s,a) && \text{(by Eq. 3.18)} \\
&= q_*(s,a)
\end{aligned}$$

This shows the third statement. Now it remains to prove the existence of an optimal policy. We first define the Bellman optimality operator.

Definition 9 (Bellman Optimality Operator)**.** *The Bellman optimality operator* $\mathcal{T} \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ *is an operator that satisfies the following: for any vector* $V = \{v(s) \colon s \in \mathcal{S}\} \in \mathbb{R}^{|\mathcal{S}|}$,

$$(\mathcal{T}V)(s) = \max_a \sum_{s',r} p(s', r | s, a)[r + \gamma\, V(s')] \tag{4.14}$$

Notice that, for an optimal policy π_*, let $V^{\pi_*} = \{v_{\pi_*}(s) \colon s \in \mathcal{S}\} \in \mathbb{R}^{|\mathcal{S}|}$. Then, Proposition 5 implies that $\mathcal{T}V^{\pi_*} = V^{\pi_*}$.

Definition 10 (Contraction)**.** *Let* X *be a metric space, with metric* d*. If* φ *maps* X *into* X *and if there is a number* $c < 1$ *such that*

$$d(\varphi(x), \varphi(y)) \le c\, d(x, y)$$

for all $x, y \in X$*, then* φ *is said to be a contraction of* X *into* X*.*

We will show that the Bellman optimality operator $\mathcal{T}$ is a contraction mapping in infinity norm,

$$\|V\|_\infty = \max_s |v(s)|$$

and we use the Banach fixed-point theorem to show the existence of optimal policy.

Theorem 11 (Banach fixed-point theorem)**.** *Let (X, d) be a non-empty complete metric space with a contraction mapping $\mathcal{T}\colon X \to X$. Then $\mathcal{T}$ admits a unique fixed-point x^* in X, i.e., $T(x^*) = x^*$. Furthermore, x^* can be found as follows: start with an arbitrary element x_0 in M and define a sequence $\{x_n\}$ by $x_n = T(x_{n-1})$ for $n \geq 1$. Then $x_n \to x_*$ exponentially rapidly.*

Before we show that the Bellman optimality operator $\mathcal{T}$ is a contraction, we state a lemma.

Lemma 12. *For any two real-valued functions $f, g\colon \mathcal{A} \to \mathbb{R}$,*

$$\left|\max_a f(a) - \max_a g(a)\right| \leq \max_a |f(a) - g(a)| \tag{4.15}$$

Proof. Without loss of generality, assume that $\max_a f(a) \geq \max_a g(a)$. Let $a^* = \arg\max_a f(a)$. Then,

$$\left|\max_a f(a) - \max_a g(a)\right| = \max_a f(a) - \max_a g(a) = f(a^*) - \max_a g(a) \leq f(a^*) - g(a^*) \leq \max_a |f(a) - g(a)|$$

This proves the lemma. □

Now we show that the Bellman optimality operator $\mathcal{T}$ is a contraction. For any $V, V' \in \mathbb{R}^{|\mathcal{S}|}$,

$$\begin{aligned}
\|\mathcal{T}V - \mathcal{T}V'\|_\infty &= \max_s |\mathcal{T}V(s) - \mathcal{T}V'(s)| && (4.16)\\
&= \max_s \left| \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')] - \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V'(s')] \right| && (4.17)\\
&\leq \max_s \max_a \left| \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')] - \sum_{s',r} p(s', r|s, a)[r + \gamma V'(s')] \right| && (4.18)\\
&\leq \max_s \max_a \left| \sum_{s'} p(s'|s, a)\, \gamma\big(V(s') - V'(s')\big) \right| && (4.19)\\
&\leq \gamma \max_{s'} \big|V(s') - V'(s')\big| && (4.20)\\
&= \gamma \|V - V'\|_\infty && (4.21)
\end{aligned}$$

Banach-fixed theorem implies that, there is a unique fixed-point V^*, that is, $\mathcal{T}V^* = V^*$. Let $V^* = \{v^*(s)\colon s \in \mathcal{S}\}$. It still remains to show that v^* is a value function for some policy π. If we show that $v^* = v_\pi$ for some π, then □

To summarize, the optimal value functions v_* and q_* are value functions corresponding to any optimal policy π_*. Therefore, the uniqueness of the optimal value functions implies that the Bellman expectation equation for an optimal policy can be restated in a special form without reference to any specific policy. We call the equation *Bellman optimality equation*.

Theorem 13 (Bellman Optimality Equation)**.**

$$v_*(s) = \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')] \quad \Big(= \max_a q_*(s, a) \Big) \tag{4.22}$$

$$q_*(s, a) = \sum_{s',r} p(s', r|s, a)[r + \gamma \max_{a'} q_*(s', a')] \quad \Big(= \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')] \Big) \tag{4.23}$$

References

[1] Andrew Barto and Richard S. Sutton, *Reinforcement Learning: An Introduction* (2nd ed.), The MIT Press, 2018.

[2] Alireza Modirshanechi, *Why does the optimal policy exist?* Medium, towards data science blog, `https://towardsdatascience.com/why-does-the-optimal-policy-exist-29f30fd51f8c`