Algorithmic and Theoretical Foundations of RL

MDP and Bellman Optimality

Table of Contents

Definition 1 (Markov Chain)
Let $\mathcal{S} = \{s_1, \cdots, s_n\}$ be a finite state space. The discrete-time dynamic system $(s_t)_{t\in\mathbb{N}} \in \mathcal{S}$ is a Markov chain if it satisfies the Markov property:

$$\mathbb{P}\left(s_{t+1} = s \mid s_t, s_{t-1}, \ldots, s_0\right) = \mathbb{P}\left(s_{t+1} = s \mid s_t\right).$$

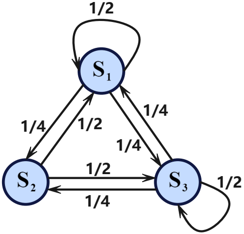
Transition Matrix:

$$P = \begin{bmatrix} p_{s_1 s_1} & p_{s_1 s_2} & \cdots & p_{s_1 s_n} \\ p_{s_2 s_1} & p_{s_2 s_2} & \cdots & p_{s_2 s_n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s_n s_1} & p_{s_n s_2} & \cdots & p_{s_n s_n} \end{bmatrix}, \quad \text{where } p_{s_i s_j} = \mathbb{P}\left(s_{t+1} = s_i | s_t = s_j\right).$$

- Under some mild conditions, there exits a stationary distribution $x \in \Delta(\mathcal{S})$ such that $Px = x$.

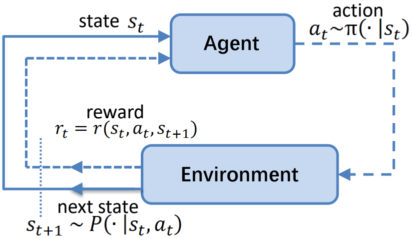
$\Delta(\mathcal{S})$ means the probability simplex on $\mathcal{S}$.

Illustrative Example



$$P = \begin{bmatrix} 1/2 & 1/2 & 1/4 \\ 1/4 & 0 & 1/4 \\ 1/4 & 1/2 & 1/2 \end{bmatrix}, \quad x = \begin{bmatrix} 2/5 \\ 1/5 \\ 2/5 \end{bmatrix}.$$

Markov Decision Process (MDP)

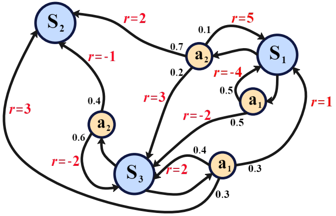


Markov chain augmented with decision and reward: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$

- $\mathcal{S}$: state space (状态空间)
- $P(\cdot|s, a)$: state transition model (状态转移模型)
- $\gamma \in [0, 1]$: discount factor (折扣因子)
- $\mathcal{A}$: action space (动作空间)
- $r(s, a, s')$: immediate reward (即时奖励), $r(s, a) = \sum_{s'} P(s'|s, a) r(s, a, s')$
- $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ (策略)

Without further specification, we assume $|\mathcal{S}| < \infty$, $|\mathcal{A}| < \infty$, and bounded immediate reward in this lecture for ease of discussion.

Illustrative Example



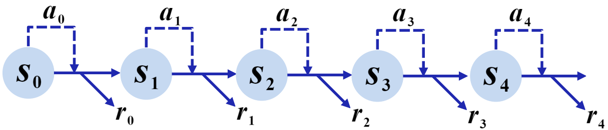
- three states: $\mathcal{S} = \{s_1, s_1, s_3\}$
- two actions: $\mathcal{A} = \{a_1, a_2\}$

Each edge is associated with a transition probability and a reward.

For instance, we can observe that:

- $p(s_3|s_3, a_2) = 0.6,\ p(s_2|s_3, a_2) = 0.4,$
- $r(s_3, a_2, s_3) = -2,\ r(s_3, a_2, s_2) = -1,$
- $r(s_3, a_2) = p(s_3|s_3, a_2)r(s_3, a_2, s_3) + p(s_2|s_3, a_2)r(s_3, a_2, s_2) = -1.6.$

State Value and Action Value



- Trajectory (轨迹):

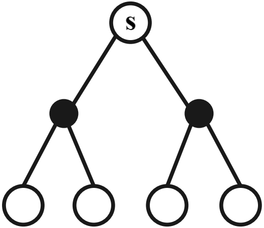
$$s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, s_3, \cdots, \quad r_t = r(s_t, a_t, s_{t+1})$$

- Infinite horizon discounted return (折扣回报):

$$r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots = \sum_{t=0}^{\infty} \gamma^t r_t$$

Here we consider infinite horizon discounted return which enable us to focus on the stationary policy. In finite horizon problems, it may be beneficial to select a different action depending on the remaining time steps which has the form $\pi(s) = (\pi_0(s), \pi_1(s), \cdots)$.

State Value and Action Value

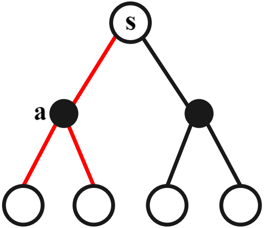


- State value (状态价值函数):

$$v_\pi(s) = \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s \right], \forall s \in \mathcal{S}$$

The expectation is indeed taken with respect to all possible random trajectories whose distribution is determined by π and P. Later, we often simplify the expression for the expectation to $\mathbb{E}_\pi$.

State Value and Action Value



- Action value (Q) (动作价值函数):

$$q_\pi(s, a) = \mathop{\mathbb{E}}_{\substack{s_{t+1} \sim P(\cdot|s_t, a_t), \\ a_{t+1} \sim \pi(\cdot|s_{t+1})}} \left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a \right], \forall (s, a) \in \mathcal{S} \times \mathcal{A}$$

State Value and Action Value

- The relation between the state value and the action value is given by

$$v_\pi(s) = \sum_a \pi(a|s)q(s, a).$$

- Computing the expectation seems not easy. However, the MDP structure enables us to compute the values by finding the solutions to linear systems (i.e., Bellman equations).

Theorem 1 (Bellman Equations)
Given an MDP, for any policy π, the values satisfy the following expectation equations:

$$v_\pi(s) = \sum_a \pi(a|s) \underbrace{\sum_{s'} P(s' \mid s, a) \left(r(s, a, s') + \gamma v_\pi(s')\right)}_{q_\pi(s,a)},$$

$$q_\pi(s, a) = \sum_{s'} P(s' \mid s, a)\Bigg(r(s, a, s') + \gamma \underbrace{\sum_{a' \in \mathcal{A}} \pi(a' \mid s') q_\pi(s', a')}_{v_\pi(s')}\Bigg).$$

- For any $v \in \mathbb{R}^{|\mathcal{S}|}$, define **Bellman Operator**

$$[\mathcal{T}_\pi v](s) = \sum_a \pi(a|s) \sum_{s'} P(s' \mid s, a) \left(r(s, a, s') + \gamma v(s')\right).$$

The Bellman equation can be rewritten as

$$v_\pi = \mathcal{T}_\pi v_\pi.$$

Matrix Form of Bellman Equation

The linear matrix-vector equation for the bellman equation for state value is given by:

$$v_\pi = \underbrace{r_\pi + \gamma P^\pi v_\pi}_{\mathcal{T}_\pi v_\pi},$$

where

$$\begin{bmatrix} v_\pi(s_1) \\ \vdots \\ v_\pi(s_n) \end{bmatrix} = \begin{bmatrix} r_\pi(s_1) \\ \vdots \\ r_\pi(s_n) \end{bmatrix} + \gamma \begin{bmatrix} p^\pi_{s_1 s_1} & \dots & p^\pi_{s_1 s_n} \\ \vdots & \ddots & \vdots \\ p^\pi_{s_n s_1} & \dots & p^\pi_{s_n s_n} \end{bmatrix} \begin{bmatrix} v_\pi(s_1) \\ \vdots \\ v_\pi(s_n) \end{bmatrix},$$

and the entries of r_π and P^π are

$$r_\pi(s) = \sum_a \pi(a|s) \sum_{s'} P(s'|s,a) r(s,a,s') \quad \text{and} \quad p^\pi_{ss'} = \sum_a \pi(a|s) P(s'|s,a).$$

Only consider the matrix form of the bellman equation for state value.

Matrix Form of Bellman Equation

Properties:

- $P^\pi \mathbf{1} = \mathbf{1}$, $|\lambda(P^\pi)| \leq 1$ for any eigenvalue of P^π
- $(I - \gamma P^\pi)$ is invertible
- $(I - \gamma P^\pi)^{-1} \geq I$
- if a vector $r \geq 0$, then $(I - \gamma P^\pi)^{-1} r \geq r \geq 0$

Solutions:

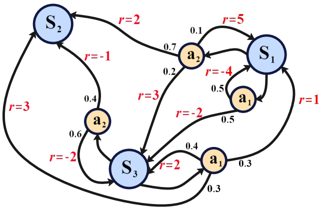
- Direct solution:

$$v_\pi = (\mathrm{I} - \gamma P^\pi)^{-1} r_\pi$$

- Iterative solution:

$$\begin{aligned} v_{k+1} &= r_\pi + \gamma P_\pi v_k = \mathcal{T}_\pi v_k, \\ v_k &\to v_\pi = (\mathrm{I} - \gamma P^\pi)^{-1} r_\pi \text{ as } k \to \infty \end{aligned}$$

Illustrative Example



Consider the policy $\pi(a|s) = 0.5$ for each state s and each action a and $\gamma = 0.9$:

$$P^\pi = \begin{bmatrix} 0.3 & 0.35 & 0.35 \\ 0 & 1 & 0 \\ 0.15 & 0.35 & 0.5 \end{bmatrix},$$

$$r_\pi = [-0.25, 0, 0.2]^\top,$$

$$v_\pi = [-0.21, 0, 0.31]^\top.$$

We can also verify the correctness of v_π. Taking the state s_0 as an example, it is not hard to show that

$$\begin{aligned} v_\pi(s_3) =& \sum_a \pi(a|s_3) \sum_{s'} p(s'|s_3, a) \left(r(s_3, a, s') + \gamma v_\pi(s') \right) \\ =& 0.5 \left(-1.6 + 0.9 \times 0.6 \times 0.31 \right) + 0.5 \left(2 + 0.9(0.4 \times 0.31 - 0.3 \times 0.21) \right) \\ =& 0.31. \end{aligned}$$

Assume s_2 will always transfers to s_2 with reward 0 no matter what action is taken.

Table of Contents

Optimal Values and Optimal Policy

Definition:

- Optimal state value: $v^*(s) = \max_\pi v_\pi(s), \forall s \in \mathcal{S}$
- Optimal action value: $q^*(s,a) = \max_\pi q_\pi(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}$

Theorem 2 (Existence of optimal policy)

For an MDP, there exists a deterministic optimal policy π^* such that

$$v_{\pi*}(s) = v^*(s), \quad q_{\pi^*}(s,a) = q^*(s,a).$$

For conciseness, we only consider stationary policies.

Proof of Theorem 2

Given the optimal values $v^*(s)$, $s \in \mathcal{S}$, define the following deterministic policy

$$\pi^*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_a \mathbb{E}_{s'}\left[r(s,a,s') + \gamma v^*(s')\right] \\ 0 & \text{otherwise.} \end{cases}$$

We are going to show that $v^*(s) = v^{\pi^*}(s)$.

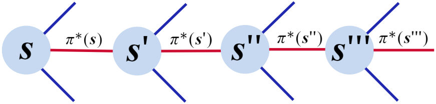
By the Bellman equation we have

$$\begin{aligned} v^*(s) = \max_\pi v_\pi(s) &= \max_\pi \left\{\mathbb{E}_\pi\left[r(s,a,s') + \gamma v_\pi(s')\right]\right\} \\ &\le \max_\pi \left\{\mathbb{E}_\pi\left[r(s,a,s') + \gamma v^*(s')\right]\right\} \\ &= \max_a \left\{r(s,a,s') + \gamma v^*(s')\right\} \\ &= \mathbb{E}_{s'}\left[r(s,\pi^*(s),s') + \gamma v^*(s')\right], \end{aligned}$$

where with a slight abuse of notation, we use $a^*(s)$ to denote the action that π^* selects at s.

We use $\pi^*(s)$ to denote the action π^* chooses.

Proof of Theorem 2 (Cont'd)



Iterating this procedure yields

$$\begin{aligned}
v^*(s) &\leq \mathbb{E}_{s'}\left[r(s,\pi^*(s),s') + \gamma v^*(s')\right] \\
&\leq \mathbb{E}_{s'}\left[r(s,\pi^*(s),s') + \gamma \mathbb{E}_{s''}\left[r(s',\pi^*(s'),s'') + \gamma v^*(s'')\right]\right] \\
&\leq \cdots\cdots \\
&\leq \mathbb{E}\left[r(s,\pi^*(s),s') + \gamma r(s',\pi^*(s'),s'') + \gamma^2 r(s'',\pi^*(s''),s''') + \cdots\right\} \\
&= v_{\pi^*}(s).
\end{aligned}$$

By the definition of v^*, we conclude that $v^*(s) = v_{\pi^*}(s)$. Moreover, a similar argument can show that $q_{\pi^*}(s,a) = q^*(s,a)$ for the same policy π^*.

Proof of Theorem 2 (More Compact Form)

Define (in an elementwise way)

$$\pi^* = \arg\max_\pi (r_\pi + P^\pi v^*).$$

Then, $\forall \pi$, one has

$$v_\pi \le r_{\pi^*} + P^{\pi^*} v^* \quad \Rightarrow \quad v_* \le r_{\pi^*} + P^{\pi^*} v^*.$$

Let $v_0 = v^*$ and define

$$v_{k+1} = r_{\pi^*} + P^{\pi^*} v_k, \ k = 0, 1, \cdots$$

Then $v_{k+1} \to v_{\pi^*}$. Moreover, if $v_k \ge v_*$, it can be easily verified $v_{k+1} \ge v_*$. By taking a limit, we have $v_{\pi^*} \ge v^*$. By the definition of v^*, there must hold $v^*(s) = v_{\pi^*}(s)$.

Remark

It is not hard to see that π^* can be rewritten as

$$\pi^*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_a q^*(s,a) \\ 0 & \text{otherwise.} \end{cases}$$

Thus, if we know $q^*(s,a)$, we immediately have the optimal policy. This observation forms the foundation of Q-learning.

- Finding the optimal policy looks challenging since there are at least as many as $|\mathcal{A}|^{|\mathcal{S}|}$ deterministic policies to test. However, we can leverage the MDP structure to transfer this problem into a dynamic programming problem. The key is hidden in Bellman optimality equations.

Bellman Optimality Equations

Theorem 3

The optimal values satisfy the following Bellman optimality equations:

$$v^*(s) = \max_a \sum_{s'} p\left(s'|s,a\right)\left(r(s,a,s') + \gamma v^*\left(s'\right)\right),$$

$$q^*(s,a) = \sum_{s'} p\left(s'|s,a\right)\left(r(s,a,s') + \gamma \max_{a'\in\mathcal{A}} q^*(s',a')\right).$$

Proof: Since $v^*(s) = v_{\pi^*}(s)$, by Bellman equation for $v_{\pi^*}(s)$, we have

$$\begin{aligned} v^*(s) = v_{\pi^*}(s) &= \mathbb{E}_{\pi^*}\left[r(s,a,s') + \gamma v^{\pi^*}(s')\right] \\ &= \mathbb{E}_{\pi^*}\left[r(s,a,s') + \gamma v^*(s')\right] \\ &= \max_a \mathbb{E}_{s'}\left[r(s,a,s') + \gamma v^*(s')\right]. \end{aligned}$$

In the remaining part, we restrict our discussion on Bellman optimality equation for optimal state value. The one for action value can be similarly discussed.

Fixed Point Theorem

Definition 2 (Contraction mapping)
Let (X, d) be a complete metric space. Then a map $T : X \to X$ is called a contraction mapping on X if there exists $\rho \in [0, 1)$ such that $d(T(x), T(y)) \leq \rho \cdot d(x, y)$ for all $x, y \in X$.

Theorem 4 (Fixed point theorem)
Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \to X$. Then T admits a unique fixed point x^* in X (i.e. $T(x^*) = x^*$). Furthermore, x^* can be obtained as follows: start with an arbitrary element $x_0 \in X$ and define a sequence $(x_k)_{k \in \mathbb{N}}$ by $x_k = T(x_{k-1})$ for $k \geq 1$. Then $\lim_{k \to \infty} x_k = x^*$.

Contraction Property of Bellman Optimality Operator

Bellman optimality operator of state value:

$$[\mathcal{T}v](s) = \max_{a} \sum_{s'} p\left(s'|s,a\right)\left(r(s,a,s') + \gamma v\left(s'\right)\right),$$

It is straightforward to see that $\mathcal{T}$ is monotone, that is $Tv_1 \leq Tv_2$ if $v_1 \leq v_2$.

Theorem 5
The Bellman optimality operator of state value is a contraction with respect to infinity norm,

$$\|\mathcal{T}v_1 - \mathcal{T}v_2\|_\infty \leq \gamma \|v_1 - v_2\|_\infty.$$

It follows that there exists a unique solution for Bellman optimality equation of state value.

Proof: The proof is based directly on the following observation:

$$|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|.$$

Questions?