强化学习的数学原理

Xuhang Ye
cquleaf@yexuhang.com

创建于: 2025 年 11 月 11 日
更新于: 2025 年 11 月 21 日

目录

1 Basic Concepts

1.1 Basic concepts in reinforcement learning

1. ***State:*** The status of the agent with respect to the environment.
2. ***State space:*** The set of all states $\mathcal{S} = \{s_i\}_{i=1}^n$.
3. ***Action:*** For each state, there are several possible actions: $a_1, \dots, a_n$.
4. ***Action space of a state:*** The set of all possible actions of a state $\mathcal{A}(s_i) = \{a_i\}_{i=1}^n$.
5. ***State transition:*** When taking an action, the agent may move from one state to another.
6. ***State transition probability:*** Use (conditional) probability to describe state transition!
7. ***Policy:*** It tells the agent what actions to take at a state.
8. ***Reward:*** A real number we get after taking an action. (Usually, positive reward represents encouragement and negative reward represents punishment)
9. ***Trajectory:*** A state-action-reward chain.
10. ***Return:*** The return of one trajectory is the sum of all the rewards collected along the trajectory.
11. ***Discounted rate:*** $\gamma \in [0, 1)$, its roles are making the sum become finite and balancing the far and near future rewards: $\gamma \to 0$ represents the near future and $\gamma \to 1$ represents the far future.
12. ***Episode:*** When interacting with the environment following a policy, the agent may stop at some *terminal states*. The resulting trajectory is called an episode (or a trial). An episode is usually assumed to be a finite trajectory. Tasks with episodes are called *episode tasks*.

1.2 Markov decision process (MDP)

1. ***Sets:***
   - ***State:*** The set of state $\mathcal{S}$
   - ***Action:*** The set of actions $\mathcal{A}(s)$ is associated for state $s \in \mathcal{S}$.
   - ***Reward:*** The set of rewards $\mathcal{R}(s,a)$.
2. ***Probability distribution:***
   - ***State transition probability:*** At state s, taking action a, the probability to transit to state s' is $p(s' \mid s,a)$.
   - ***Reward probability:*** At state s, taking action a, the probability to get reward r is $p(r \mid s,a)$.
3. ***Policy:*** At state s, the probability to choose action a is $\pi(a \mid s)$.
4. ***Markov property:*** memoryless property

$$\begin{aligned} p\left(s_{t+1} \mid a_{t+1}, s_t, \ldots, a_1, s_0\right) &= p\left(s_{t+1} \mid a_{t+1}, s_t\right), \\ p\left(r_{t+1} \mid a_{t+1}, s_t, \ldots, a_1, s_0\right) &= p\left(r_{t+1} \mid a_{t+1}, s_t\right). \end{aligned} \tag{1}$$

Markov decision process becomes Markov process once the policy is given!

2 Bellman Equation

2.1 Motivating examples

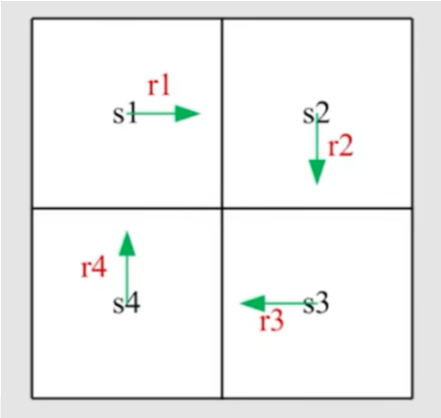

图 1: The example to introduce bellman equation

How to calculate return?

Method 1: by definition

Let v_i denote the return obtained starting from s_i $(i = 1, 2, 3, 4)$

$$\begin{aligned} v_1 &= r_1 + \gamma r_2 + \gamma^2 r_3 + \dots \\ v_2 &= r_2 + \gamma r_3 + \gamma^2 r_4 + \dots \\ v_3 &= r_3 + \gamma r_4 + \gamma^2 r_1 + \dots \\ v_4 &= r_4 + \gamma r_1 + \gamma^2 r_2 + \dots \end{aligned} \tag{2}$$

Method 2: bootstrapping

$$\begin{aligned} v_1 &= r_1 + \gamma\left(r_2 + \gamma r_3 + \dots\right) = r_1 + \gamma v_2 \\ v_2 &= r_2 + \gamma\left(r_3 + \gamma r_4 + \dots\right) = r_2 + \gamma v_3 \\ v_3 &= r_3 + \gamma\left(r_4 + \gamma r_1 + \dots\right) = r_3 + \gamma v_4 \\ v_4 &= r_4 + \gamma\left(r_1 + \gamma r_2 + \dots\right) = r_4 + \gamma v_1 \end{aligned} \tag{3}$$

How to solve the above equations?

Write in the following matrix-vector form:

$$\underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{\mathbf{v}} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \begin{bmatrix} \gamma v_2 \\ \gamma v_3 \\ \gamma v_4 \\ \gamma v_1 \end{bmatrix} = \underbrace{\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}}_{\mathbf{r}} + \gamma \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{\mathbf{v}} \tag{4}$$

which can be rewritten as

$$\mathbf{v} = \mathbf{r} + \gamma \mathbf{P}\mathbf{v} \tag{5}$$

This is **Bellman equation** (for this specific deterministic problem)

1. Though simple, it demonstrates the core idea: the value of one state relies on the values of other states.
2. A matrix-vector form is more clear to see how to solve the state values.

2.2 State value

Consider the following single-step process:

$$S_t \xrightarrow{A_t} R_{t+1}, S_{t+1} \tag{6}$$

1. $t, t+1$: discrete time instances
2. S_t: **state** at time t
3. A_t: the **action** taken at state S_t
4. R_{t+1}: the **reward** obtained after taking A_t
5. S_{t+1}: the **state transited** to after taking A_t

Note that S_t, A_t, R_{t+1} are all *random variables*.
This step is governed by the following probability distributions:

1. $S_t \to A_t$ is governed by $\pi(A_t = a \mid S_t = s)$
2. $S_t, A_t \to R_{t+1}$ is governed by $p(R_{t+1} = r \mid S_t = s, A_t = a)$
3. $S_t, A_t \to S_{t+1}$ is governed by $p(S_{t+1} = s' \mid S_t = s, A_t = a)$

At this moment, we assume we know the model (i.e., the probability distributions)! Consider the following multi-step trajectory:

$$S_t \xrightarrow{A_t} R_{t+1}, S_{t+1} \xrightarrow{A_{t+1}} R_{t+2}, S_{t+2} \xrightarrow{A_{t+2}} R_{t+3}, \dots \tag{7}$$

The discounted return is

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \tag{8}$$

Where:

- $\gamma \in [0, 1)$ is a discounted rate.
- G_t is also a random variable since $R_{t+1}, R_{t+2}, \dots$ are random variables.

The expectation (or called expected value or mean) of G_t is defined as the *state-value function* or simply *state value*:

$$v_\pi(s) = \mathbb{E}[G_t \mid S_t = s] \tag{9}$$

Remarks:

1. It is a function of s. It is a conditional expectation with the condition that the state starts from s.
2. It is based on the policy π. For a different policy, the state value may be different.
3. It represents the "value" of a state. If the state value is greater, then the policy is better because greater cumulative rewards can be obtained.

Q: What is the relationship between **return** and **state value**?

A: The state value is the mean of all possible returns that can be obtained starting from a state. If everything —$\pi(a \mid s)$, $p(r \mid s, a)$, $p(s' \mid s, a)$ —is deterministic, then state value is the same as return.

2.3 Bellman equation: Derivation

Consider a random trajectory:

$$S_t \xrightarrow{A_t} R_{t+1}, S_{t+1} \xrightarrow{A_{t+1}} R_{t+2}, S_{t+2} \xrightarrow{A_{t+2}} R_{t+3}, \ldots \tag{10}$$

The return G_t can be written as

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots, \\ &= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \ldots \right), \\ &= R_{t+1} + \gamma G_{t+1}, \end{aligned} \tag{11}$$

Then, it follows from the definition of the state value that

$$\begin{aligned} v_\pi(s) &= \mathbb{E}\left[G_t \mid S_t = s \right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s \right] \\ &= \mathbb{E}\left[R_{t+1} \mid S_t = s \right] + \gamma \mathbb{E}\left[G_{t+1} \mid S_t = s \right] \end{aligned} \tag{12}$$

Next, calculate the two terms, respectively.
First, calculate the first term $\mathbb{E}\left[R_{t+1} \mid S_t = s \right]$:

$$\begin{aligned} \mathbb{E}\left[R_{t+1} \mid S_t = s \right] &= \sum_a \pi(a \mid s) \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a \right] \\ &= \sum_a \pi(a \mid s) \sum_r p(r \mid s, a) r \end{aligned} \tag{13}$$

Note: This is the mean of *immediate rewards*
Second, calculate the second term $\mathbb{E}\left[G_{t+1} \mid S_t = s \right]$

$$\begin{aligned} \mathbb{E}\left[G_{t+1} \mid S_t = s \right] &= \sum_{s'} \mathbb{E}\left[G_{t+1} \mid S_t = s, S_{t+1} = s' \right] p\left(s' \mid s \right) \\ &= \sum_{s'} \mathbb{E}\left[G_{t+1} \mid S_{t+1} = s' \right] p\left(s' \mid s \right) \\ &= \sum_{s'} v_\pi\left(s' \right) p\left(s' \mid s \right) \\ &= \sum_{s'} v_\pi\left(s' \right) \sum_a p\left(s' \mid s, a \right) \pi(a \mid s) \end{aligned} \tag{14}$$

Note: This is the mean of *future rewards*

Theorem 2.1: Bellman equation

Therefore, we have:

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}\left[R_{t+1} \mid S_t = s\right] + \gamma \mathbb{E}\left[G_{t+1} \mid S_t = s\right], \\
&= \underbrace{\sum_a \pi(a \mid s) \sum_r p(r \mid s,a) r}_{\text{mean of immediate rewards}} + \gamma \underbrace{\sum_a \pi(a \mid s) \sum_{s'} p\left(s' \mid s,a\right) v_\pi\left(s'\right)}_{\text{mean of future rewards}}, \\
&= \sum_a \pi(a \mid s) \left[\sum_r p(r \mid s,a) r + \gamma \sum_{s'} p\left(s' \mid s,a\right) v_\pi\left(s'\right) \right], \quad \forall s \in \mathcal{S}.
\end{aligned} \tag{15}
$$

Highlights:

1. The above equation is called **the Bellman equation**, which characterizes the relationship among the state-value functions of different states.
2. It consists of two terms: the immediate reward term and the future reward term.
3. A set of equations: every state has an equation like this.(important!!!)
4. $v_\pi(s)$ and $v_\pi\left(s'\right)$ are state values to be calculated. Using **bootstrapping** to solve it!
5. $\pi(a \mid s)$ is a given policy. Solving the equation is called policy evaluation.
6. $p(r \mid s,a)$ and $p\left(s' \mid s,a\right)$ represent the dynamic model.

2.4 Bellman equation: Matrix-vector form

3 插图示例

图 2: 示例图片

4 彩色信息框示例

Definition 4.1: 具身马尔可夫决策过程（EMDP）

个具身马尔可夫决策过程定义为五元组 $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$，其中：

- $\mathcal{S} = \mathbb{R}^2 \times [0, 2\pi)$ 为状态空间（位置 + 朝向）；
- $\mathcal{A} = \{\text{前进}, \text{左转}, \text{右转}\}$ 为离散动作空间；
- $T(s' \mid s, a)$ 为状态转移函数；
- $R(s, a) = -\|s_{\text{pos}} - s_{\text{goal}}\|$ 为稀疏奖励（负欧氏距离）；
- $\gamma \in (0, 1)$ 为折扣因子。

Assumption 4.1: 可观测性

能体在每一步都能精确观测当前状态 $s_t \in \mathcal{S}$。

Assumption 4.2: 确定性动力学

态转移是确定性的，即 $s_{t+1} = f(s_t, a_t)$，无环境噪声。

该算法的正确性由以下定理保证。

Lemma 4.1: 贝尔曼最优性

函数 V^* 是贝尔曼最优算子 $\mathcal{T}$ 的唯一不动点，即 $V^* = \mathcal{T} V^*$，其中

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

Theorem 4.1: 值迭代收敛性

???? 下，值迭代算法生成的序列 $\{V_k\}$ 以指数速率收敛到 V^*，即

$$\|V_k - V^*\|_\infty \leq \frac{2\gamma^k}{1 - \gamma} \|V_1 - V_0\|_\infty.$$

证明. 由于状态空间 $\mathcal{S}$ 在假设下为有限（或可离散化），且 $\gamma < 1$，贝尔曼算子 $\mathcal{T}$ 是 γ-压缩映射。由巴拿赫不动点定理，迭代 $V_{k+1} = \mathcal{T} V_k$ 收敛到唯一不动点 V^*，且误差界如上所述。 □

Example 4.1: 2D 网格世界导航

考虑一个 5×5 的网格世界，机器人从左下角 $(0,0)$ 出发，目标为右上角 $(4,4)$。状态 $s = (x, y, \theta)$，其中 $\theta \in \{0, \pi/2, \pi, 3\pi/2\}$。动作 $\mathcal{A} = \{\text{前进}, \text{左转}, \text{右转}\}$ 控制朝向与位置。在 **??** 下，值迭代可精确计算到达目标的最短路径。