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 12 月 2 日

Contents

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

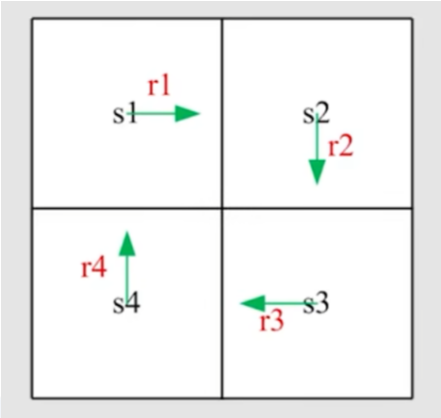

Figure 1: The example to introduce bellman equation

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

Definition 2.1: A random trajectory

Consider a random trajectory:

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

Proof.

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

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

Why consider the matrix-vector form?

1. For Bellman equation, one unkown relies on another unkown.
2. The above *elementwise form* is valid for every state $s \in \mathcal{S}$. That means there are $|\mathcal{S}|$ equations like this.
3. If we put all the equations together, we have a set of linear equations, which can be concisely written in a *matrix-vector form*.
4. The matrix-vector form is very elegant and important.

Recall that:

$$v_\pi(s) = \sum_a \pi(a \mid s) \left[\sum_r p(r \mid s,a) r + \gamma \sum_{s'} p\left(s' \mid s,a\right) v_\pi\left(s'\right) \right] \tag{16}$$

Rewrite the Bellman equation as

$$v_\pi(s) = r_\pi(s) + \gamma \sum_{s'} p_\pi\left(s' \mid s\right) v_\pi\left(s'\right) \tag{17}$$

Where

$$r_\pi(s) \triangleq \sum_a \pi(a \mid s) \sum_r p(r \mid s,a) r, \quad p_\pi\left(s' \mid s\right) \triangleq \sum_a \pi(a \mid s) p\left(s' \mid s,a\right) \tag{18}$$

Suppose the states could be indexed as $s_i (i = 1, \ldots, n)$.
For state s_i, the Bellman equation is

$$v_\pi\left(s_i\right) = r_\pi\left(s_i\right) + \gamma \sum_{s_j} p_\pi\left(s_j \mid s_i\right) v_\pi\left(s_j\right) \tag{19}$$

Put all these equations for all the states together and rewrite to a matrix-vector form

$$v_\pi = r_\pi + \gamma P_\pi v_\pi \tag{20}$$

where

1. $v_\pi = \left[v_\pi\left(s_1\right), \ldots, v_\pi\left(s_n\right)\right]^T \in \mathbb{R}^n$
2. $r_\pi = \left[r_\pi\left(s_1\right), \ldots, r_\pi\left(s_n\right)\right]^T \in \mathbb{R}^n$
3. $P_\pi \in \mathbb{R}^{n \times n}$, where $\left[P_\pi\right]_{ij} = p_\pi\left(s_j \mid s_i\right)$, is the *state transition matrix*.

If there are four states, $v_\pi = r_\pi + \gamma P_\pi v_\pi$ can be written out as

$$\underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} = \underbrace{\begin{bmatrix} r_\pi(s_1) \\ r_\pi(s_2) \\ r_\pi(s_3) \\ r_\pi(s_4) \end{bmatrix}}_{r_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(s_1|s_1) & p_\pi(s_2|s_1) & p_\pi(s_3|s_1) & p_\pi(s_4|s_1) \\ p_\pi(s_1|s_2) & p_\pi(s_2|s_2) & p_\pi(s_3|s_2) & p_\pi(s_4|s_2) \\ p_\pi(s_1|s_3) & p_\pi(s_2|s_3) & p_\pi(s_3|s_3) & p_\pi(s_4|s_3) \\ p_\pi(s_1|s_4) & p_\pi(s_2|s_4) & p_\pi(s_3|s_4) & p_\pi(s_4|s_4) \end{bmatrix}}_{P_\pi} \underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} \tag{21}$$

For this specific example:

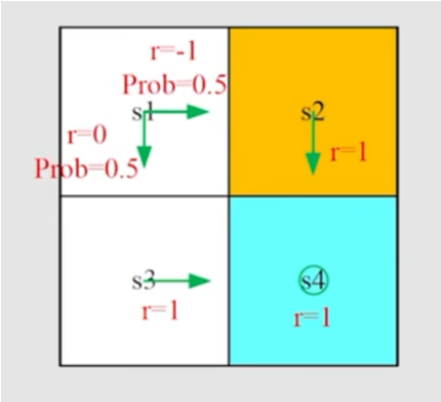

Figure 2: The example to sovle matrix-vector bellman equation

$$\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix} = \begin{bmatrix} 0.5(0) + 0.5(-1) \\ 1 \\ 1 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix} \tag{22}$$

2.5 Bellman equation: Solve the state values

Why to solve state values?

Given a policy, finding out the corresponding state values is called **policy evaluation**! It is a fundamental problem in RL. It is the foundation to find better policies.

The Bellman equation in matrix-vector form is

$$v_\pi = r_\pi + \gamma P_\pi v_\pi \tag{23}$$

The closed-form solution is:

$$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi \tag{24}$$

In practice, we still need to use numerical tools to calculate the matrix inverse.

An iterative solution is:

$$v_{k+1} = r_\pi + \gamma P_\pi v_k \tag{25}$$

This algorithm leads to a sequence $\{v_0, v_1, v_2, ...\}$. We can show that

$$v_k \to v_\pi = (I - \gamma P_\pi)^{-1} r_\pi, \quad k \to \infty \tag{26}$$

Proof. Define the error as $\delta_k = v_k - v_\pi$. We only need to show $\delta_k \to 0$. Substituting $v_{k+1} = \delta_{k+1} + v_\pi$ and $v_k = \delta_k + v_\pi$ into $v_{k+1} = r_\pi + \gamma P_\pi v_k$ gives

$$\delta_{k+1} + v_\pi = r_\pi + \gamma P_\pi (\delta_k + v_\pi) \tag{27}$$

which can be rewritten as

$$\delta_{k+1} = -v_\pi + r_\pi + \gamma P_\pi \delta_k + \gamma P_\pi v_\pi = \gamma P_\pi \delta_k \tag{28}$$

As a result,

$$\delta_{k+1} = \gamma P_\pi \delta_k = \gamma^2 P_\pi^2 \delta_{k-1} = \cdots = \gamma^{k+1} P_\pi^{k+1} \delta_0 \tag{29}$$

For P_π^k, P_π is the *state transition probability matrix* under policy π has a sum of 1 for each row (because it is a probability distribution), and all elements are in the interval $[0, 1]$. Thus, each element of P_π^k must also be between $[0, 1]$(because the product and sum of probabilities will not exceed 1).
For γ^k, since $\gamma < 1$, we know $\gamma^k \to 0$
hence $\delta_{k+1} = \gamma^{k+1} P_\pi^{k+1} \delta_0 \to 0$ as $k \to \infty$. □

2.6 Action value

Action value: The average return the agent can get starting from a state and taking an action.

Definition 2.2: Action value

$$q_\pi(s, a) = \mathbb{E}\left[G_t \mid S_{t_s} = s, A_t = a\right] \tag{30}$$

1. $q_\pi(s, a)$ is a function of the state-action pair (s, a)
2. $q_\pi(s, a)$ depends on π

It follows from the properties of conditional expectation that

$$\underbrace{\mathbb{E}\left[G_t \mid S_t = s\right]}_{v_\pi(s)} = \sum_a \underbrace{\mathbb{E}\left[G_t \mid S_t = s, A_t = a\right]}_{q_\pi(s,a)} \pi(a \mid s) \tag{31}$$

Hence,

$$v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s, a) \tag{32}$$

Recall that the state value is given by

$$v_\pi(s) = \sum_a \pi(a \mid s)[\underbrace{\sum_r p(r \mid s,a)r + \gamma \sum_{s'} p\left(s' \mid s,a\right) v_\pi\left(s'\right)}_{q_\pi(s,a)}] \tag{33}$$

By comparing (32) and (33), we have the **action-value function** as

$$q_\pi(s,a) = \sum_r p(r \mid s,a)r + \gamma \sum_{s'} p\left(s' \mid s,a\right) v_\pi\left(s'\right) \tag{34}$$

(32) and (34) are the two sides of the same coin:

- (32) shows how to obtain state values from action values.
- (34) shows how to obtain action values from state values.

Highlights

1. Action value is important since we care about which action to take.
2. We can first calculate all the state values and then calculate the action values.
3. We can also directly calculate the action values with or without models.

3 Bellman Optimality Equation

3.1 Definition of optimal policy

The state value could be used to evaluate if a policy is good or not if $v_{\pi_1}(s) \geq v_{\pi_2}(s)$ for all $s \in \mathcal{S}$ then π_1 is "better" than π_2.

Definition 3.1: Optimal policy

A policy π^* is optimal if $v_{\pi^*}(s) \geq v_{\pi}(s)$ for all s and for any other policy π.

This definition leads to many questions:

1. Does the optimal policy exist?
2. Is the optimal policy unique?
3. Is the optimal policy stochastic or deterministic?
4. How to obtain the optimal policy?

To answer these questions, we should study the *Bellman optimality equation*.

3.2 Introduce and solve the BOE

Theorem 3.1: Bellman Optimality Equation (BOE)

BOE(elementwise form):

$$\begin{aligned} v(s) &= \max_{\pi} \sum_{a} \pi(a \mid s) \left(\sum_{r} p(r \mid s,a) r + \gamma \sum_{s'} p\left(s' \mid s,a\right) v\left(s'\right) \right), \quad \forall s \in \mathcal{S} \\ &= \max_{\pi} \sum_{a} \pi(a \mid s) q(s,a) \quad s \in \mathcal{S} \end{aligned} \tag{35}$$

Remarks:

1. $p(r \mid s,a), p\left(s' \mid s,a\right)$ are known.
2. $v(s), v\left(s'\right)$ are unknown and to be calculated.
3. Is $\pi(s)$ known or unknown?

BOE(matrix-vector form):

$$v = \max_{\pi} \left(r_{\pi} + \gamma P_{\pi} v\right) \tag{36}$$

where the elements corresponding to s or s' are

$$\begin{aligned}[r_\pi]_s &\triangleq \sum_a \pi(a \mid s) \sum_r p(r \mid s,a) r \\ [P_\pi]_{s,s'} &= p\left(s' \mid s\right) \triangleq \sum_a \pi(a \mid s) \sum_{s'} p\left(s' \mid s,a\right)\end{aligned} \tag{37}$$

Here $\max_\pi$ is performed elementwise.

BOE is tricky yet elegant!

1. *Why elegant?* It describes the optimal policy and optimal state value in an elegant way.
2. *Why tricky?* There is a maximization on the right-hand side, which may not be straightforward to see how to compute.

There are also many questions to answer for this theorem

1. Algorithm: how to solve this equation?
2. existence: does this equation have solutions?
3. Uniqueness: is the solution to this equation unique?
4. Optimality: how is it related to optimal policy?

For theorem 3.1, there are two unknowns from one equation. This seems hard to solve.

Example 3.1: How to solve two unknowns from one equation

Consider two variables $x, a \in \mathbb{R}$. Suppose they satisfy

$$x = \max_a \left(2x - 1 - a^2\right) \tag{38}$$

This equation has two unknowns. To solve them, first consider the right hand side. Regardless the value of x, $\max_a \left(2x - 1 - a^2\right) = 2x - 1$ where the maximization is achieved when $a = 0$. Second, when $a = 0$, the equation becomes $x = 2x - 1$, which leads to $x = 1$. Therefore, $a = 0$ and $x = 1$ are the solution of the equation.

We can use the approach from example 3.1 to solve the BOE problem.
Fix $v'(s)$ first and solve π:

$$
\begin{aligned}
v(s) &= \max_{\pi} \sum_{a} \pi(a \mid s) \left(\sum_{r} p(r \mid s,a) r + \gamma \sum_{s'} p\left(s' \mid s,a\right) v\left(s'\right) \right), \quad \forall s \in \mathcal{S} \\
&= \max_{\pi} \sum_{a} \pi(a \mid s) q(s,a)
\end{aligned}
\tag{39}
$$

Example 3.2: How to solve $\max_{\pi} \sum_{a} \pi(a \mid s) q(s,a)$

Suppose $q_1, q_2, q_3 \in \mathbb{R}$ are given. Find c_1^*, c_2^*, c_3^* solving

$$
\max_{c_1, c_2, c_3} c_1 q_1 + c_2 q_2 + c_3 q_3 \tag{40}
$$

where $c_1 + c_2 + c_3 = 1$ and $c_1, c_2, c_3 \geq 0$.
Without loss of generality, suppose $q_3 \geq q_1, q_2$. Then, the optimal solution is $c_3^* = 1$ and $c_1^* = c_2^* = 0$. That is because for any c_1, c_2, c_3
$q_3 = \left(c_1 + c_2 + c_3\right) q_3 = c_1 q_3 + c_2 q_3 + c_3 q_3 \geq c_1 q_1 + c_2 q_2 + c_3 q_3$

Inspired by the above example 3.2, consider that $\sum_{a} \pi(a \mid s) = 1$, we have

$$
\max_{\pi} \sum_{a} \pi(a \mid s) q(s,a) = \max_{a \in \mathcal{A}(s)} q(s,a) \tag{41}
$$

where the optimality is achieved when

$$
\pi(a \mid s) = \begin{cases} 1 & a = a^* \\ 0 & a \neq a^* \end{cases} \tag{42}
$$

where $a^* = \arg\max_{a} q(s,a)$
The BOE is $v = \max_{\pi} \left(r_\pi + \gamma P_\pi v\right)$. Let

$$
f(v) := \max_{\pi} \left(r_\pi + \gamma P_\pi v\right) \tag{43}
$$

Then, the BOE becomes

$$
v = f(v) \tag{44}
$$

where

$$
[f(v)]_s = \max_{\pi} \sum_{a} \pi(a \mid s) q(s,a), \quad s \in \mathcal{S} \tag{45}
$$

Theorem 3.2: Contraction mapping theorem

Fixed point: $x \in X$ is a fixed point of $f : X \to X$ if $f(x) = x$

Contraction mapping(or contractive function): f is a contraction mapping if

$$\|f(x_1) - f(x_2)\| \leq \gamma \|x_1 - x_2\| \tag{46}$$

where $\gamma \in (0, 1)$

1. γ must be strictly less than 1 so that many limits such as $\gamma^k \to 0$ as $k \to 0$
2. Here $\|\cdot\|$ can be any vector norm.

Here are some examples to demonstrate the concepts.

Example 3.3: Fixed point and contraction mapping

For $x = f(x) = 0.5x, x \in \mathbb{R}$

It is easy to verify that $x = 0$ is a fixed point. Moreover, $f(x) = 0.5x$ is a contraction mapping because $\|0.5x_1 - 0.5x_2\| = 0.5\|x_1 - x_2\| \leq \gamma \|x_1 - x_2\|$ for any $\gamma \in [0.5, 1)$.

For $x = f(x) = Ax$**, where** $x \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n}$ **and** $\|A\| \leq \gamma < 1$

It is easy to verify that $x = 0$ is a fixed point. To see the contraction property, $\|Ax_1 - Ax_2\| = \|A(x_1 - x_2)\| \leq \|A\| \|x_1 - x_2\| \leq \gamma \|x_1 - x_2\|$. Therefore, $f(x) = Ax$ is a contraction mapping.

Theorem 3.3: Contraction mapping theorem

For any equation that has the form of $x = f(x)$, if f is a contraction mapping, then

1. ***Existence:*** there exists a fixed point x^* satisfying $f(x^*) = x^*$.
2. ***Uniqueness:*** the fixed point x^* is unique.
3. ***Algorithm:*** consider a sequence $\{x_k\}$ where $x_{k+1} = f(x_k)$, then $x_k \to x^*$ as $k \to \infty$. Moreover, the convergence rate is exponentially fast.

Let's come back to the Bellman optimality equation:

$$v = f(v) = \max_{\pi} (r_\pi + \gamma P_\pi v) \tag{47}$$

Lemma 3.1: Contraction property

$f(v)$ is a contraction mapping satisfying

$$\|f(v_1) - f(v_2)\| \leq \gamma \|v_1 - v_2\| \tag{48}$$

where γ is the discount rate!

Apply the theorem 3.3 gives the following results.

Theorem 3.4: Existence, uniqueness and algorithm

For the BOE (47), there always **exists** a solution v^* and the solution is **unique**. The solution could be solved iteratively by

$$v_{k+1} = f(v_k) = \max_{\pi} (r_\pi + \gamma P_\pi v_k) \tag{49}$$

This sequence $\{v_k\}$ converges to v^* **exponentially fast** given any initial guess v_0. The convergence rate is determined by γ.

Suppose v^* is the solution to the BOE (47). It satisfies

$$v^* = \max_{\pi} (r_\pi + \gamma P_\pi v^*) \tag{50}$$

Suppose

$$\pi^* = \arg\max_{\pi} (r_\pi + \gamma P_\pi v^*) \tag{51}$$

Then

$$v^* = r_{\pi^*} + \gamma P_{\pi^*} v^* \tag{52}$$

Therefore, π^* is a policy and $v^* = v_{\pi^*}$ is the corresponding state value.

Theorem 3.5: Policy optimality

Suppose that v^* is the unique solution to $v = \max_\pi (r_\pi + \gamma P_\pi v)$, and v_π is the state value function satisfying $v_\pi = r_\pi + \gamma P_\pi v_\pi$ for any given policy π, then

$$v^* \geq v_\pi, \quad \forall \pi \tag{53}$$

Therefore the reason why we should study the BOE is that it describes the optimal state value and optimal policy.

Then what does an optimal policy π^* look like?

Theorem 3.6: Greedy optimal policy

For any $s \in \mathcal{S}$, the deterministic greedy policy

$$\pi^*(a \mid s) = \begin{cases} 1 & a = a^*(s) \\ 0 & a \neq a^*(s) \end{cases} \tag{54}$$

is an optimal policy solving the BOE. Here,

$$a^*(s) = \arg\max_a q^*(a, s) \tag{55}$$

where

$$q^*(s, a) := \sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v^*\left(s'\right) \tag{56}$$

3.3 Analyzing optimal policies

What factors determine the optimal policy?

Recall the BOE (35), we can find that there are three factors:

1. Reward design: r
2. System model: $p\left(s' \mid s, a\right), p(r \mid s, a)$
3. Discount rate: γ

$v(s), v\left(s'\right), \pi(a \mid s)$ are unknowns to be calculated and system model is hard to change. So what we can change is only the **reward design** and **discount rate**.

Theorem 3.7: Optimal policy invariance

Consider a MDP with $v^* \in \mathbb{R}^{|\mathcal{S}|}$ as the optimal state value satisfying $v^* = \max_\pi \left(r_\pi + \gamma P_\pi v^*\right)$. If every reward r is changed by an affine transformation to $ar+b$, where $a, b \in \mathbb{R}$ and $a \neq 0$, then the corresponding optimal state value v' is also an affine transformation of v^*

$$v' = av^* + \frac{b}{1-\gamma}\mathbf{1} \tag{57}$$

where $\gamma \in (0, 1)$ is the discount rate and $\mathbf{1} = [1, \ldots, 1]^T$. Consequently, the optimal policies are invariance to the affine transformation of the reward signals.

4 彩色信息框示例