

Lecture 3: Optimal State Values and Bellman Optimality Equation

Wen Fuxi

Review

Key concepts and results:

- ▶ State value: $v_\pi(s) = \mathbb{E}[G_t | S_t = s]$
- ▶ Action value: $q_\pi(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a]$
- ▶ The Bellman equation (elementwise form):

$$\begin{aligned} v_\pi(s) &= \sum_a \pi(a | s) \underbrace{\left[\sum_r p(r | s, a) r + \gamma \sum_{s'} p(s' | s, a) v_\pi(s') \right]}_{q_\pi(s, a)} \\ &= \sum_a \pi(a | s) q_\pi(s, a) \end{aligned}$$

- ▶ The Bellman equation (matrix-vector form):

$$v_\pi = r_\pi + \gamma P_\pi v_\pi$$

- ▶ How to solve the Bellman equation: closed-form solution, iterative solution

Deriving the Bellman equation

Consider a random trajectory:

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

The return G_t can be written as

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

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

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

Next, calculate the two terms, respectively.

Deriving the Bellman equation

First, calculate the first term $\mathbb{E}[R_{t+1} | S_t = s]$:

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

Note that

- ▶ This is the mean of immediate rewards

Deriving the Bellman equation

Second, calculate the second term $\mathbb{E}[G_{t+1} | S_t = s]$:

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

Note that

- This is the mean of future rewards
- $\mathbb{E}[G_{t+1} | S_t = s, S_{t+1} = s'] = \mathbb{E}[G_{t+1} | S_{t+1} = s']$ due to the memoryless Markov property.

Deriving the Bellman equation

Therefore, we have

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

- The above equation is known as the Bellman equation, which characterizes the relationship between the state-value functions of different states.
- It consists of two terms: the immediate reward term and the future reward term.
- $v_{\pi}(s)$ and $v_{\pi}(s')$ are state values to be calculated. **Bootstrapping!**
- $\pi(a \mid s)$ is a given policy. Solving the equation is referred to as **policy evaluation**.
- $p(r \mid s, a)$ and $p(s' \mid s, a)$ represent the dynamic "model".

Action value

Definition:

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

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

It follows from the properties of conditional expectation that

$$\underbrace{\mathbb{E}[G_t | S_t = s]}_{v_{\pi}(s)} = \sum_a \underbrace{\mathbb{E}[G_t | S_t = s, A_t = a] \pi(a | s)}_{q_{\pi}(s, a)}$$

Hence,

$$v_{\pi}(s) = \sum_a \pi(a | s) q_{\pi}(s, a) \quad (2)$$

Action value

Recall that the state value is given by

$$v_{\pi}(s) = \sum_a \pi(a | s) [\underbrace{\sum_r p(r | s, a)r + \gamma \sum_{s'} p(s' | s, a) v_{\pi}(s')}_{q_{\pi}(s, a)}] \quad (3)$$

By comparing (2) and (3), we have the action-value function as

$$q_{\pi}(s, a) = \sum_r p(r | s, a)r + \gamma \sum_{s'} p(s' | s, a) v_{\pi}(s') \quad (4)$$

(2) and (4) are the **two sides of the same coin**:

- (2) shows how to obtain state values from action values.
- (4) shows how to obtain action values from state values.

In this lecture:

- ▶ **Core Concepts:**
Optimal state value and Optimal policy
- ▶ **A Fundamental Tool:**
Bellman optimality equation (BOE)

1. Motivating examples
2. Definition of optimal policy
3. BOE: Introduction
4. BOE: Preliminaries
 - 4.1 BOE: Maximization on the right-hand side
 - 4.2 BOE: Rewrite as $v = f(v)$
 - 4.3 Contraction mapping theorem
5. BOE: Solution
6. BOE: Optimality
7. Analyzing optimal policies

1. Motivating examples

2. Definition of optimal policy

3. BOE: Introduction

4. BOE: Preliminaries

4.1 BOE: Maximization on the right-hand side

4.2 BOE: Rewrite as $v = f(v)$

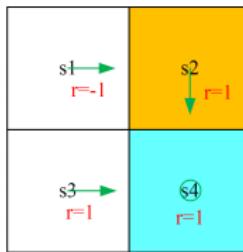
4.3 Contraction mapping theorem

5. BOE: Solution

6. BOE: Optimality

7. Analyzing optimal policies

Motivating examples



Exercise: Write out the Bellman equation and solve the state values (set $\gamma = 0.9$)

Bellman equations:

$$v_{\pi}(s_1) = -1 + \gamma v_{\pi}(s_2),$$

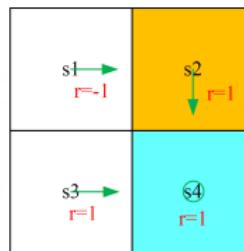
$$v_{\pi}(s_2) = +1 + \gamma v_{\pi}(s_4),$$

$$v_{\pi}(s_3) = +1 + \gamma v_{\pi}(s_4),$$

$$v_{\pi}(s_4) = +1 + \gamma v_{\pi}(s_4).$$

State values: $v_{\pi}(s_4) = v_{\pi}(s_3) = v_{\pi}(s_2) = 10, v_{\pi}(s_1) = 8$

Motivating examples



Exercise: calculate the action values of the five actions for s_1

Action values:

$$q_{\pi}(s_1, a_1) = -1 + \gamma v_{\pi}(s_1) = 6.2,$$

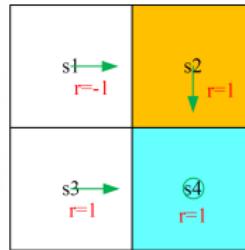
$$q_{\pi}(s_1, a_2) = -1 + \gamma v_{\pi}(s_2) = 8,$$

$$q_{\pi}(s_1, a_3) = 0 + \gamma v_{\pi}(s_3) = 9$$

$$q_{\pi}(s_1, a_4) = -1 + \gamma v_{\pi}(s_1) = 6.2,$$

$$q_{\pi}(s_1, a_5) = 0 + \gamma v_{\pi}(s_1) = 7.2.$$

Motivating examples



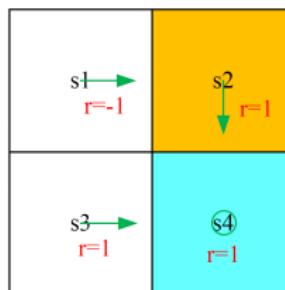
Question: While the policy is not good, how can we improve it?

Answer: We can improve the policy based on action values.

In particular, the current policy $\pi(a | s_1)$ is

$$\pi(a | s_1) = \begin{cases} 1 & a = a_2 \\ 0 & a \neq a_2 \end{cases}$$

Motivating examples



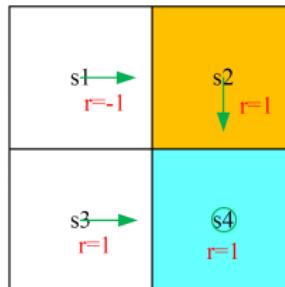
Observe the action values that we obtained just now:

$$\begin{aligned} q_{\pi}(s_1, a_1) &= 6.2, & q_{\pi}(s_1, a_2) &= 8, & q_{\pi}(s_1, a_3) &= 9, \\ q_{\pi}(s_1, a_4) &= 6.2, & q_{\pi}(s_1, a_5) &= 7.2. \end{aligned}$$

What if we select the greatest action value? Then, the new policy is

$$\pi_{\text{new}}(a | s_1) = \begin{cases} 1 & a = a_3 \\ 0 & a \neq a_3 \end{cases}$$

Motivating examples



Question: Why can doing this improve the policy?

- ▶ Intuition: easy! Actions with greater values are better.
- ▶ Math: nontrivial! This will be introduced in this lecture and the next one!

1. Motivating examples

2. Definition of optimal policy

3. BOE: Introduction

4. BOE: Preliminaries

4.1 BOE: Maximization on the right-hand side

4.2 BOE: Rewrite as $v = f(v)$

4.3 Contraction mapping theorem

5. BOE: Solution

6. BOE: Optimality

7. Analyzing optimal policies

Optimal policy

The **state value** could be used to evaluate if a policy is good or not: if

$$v_{\pi_1}(s) \geq v_{\pi_2}(s) \quad \text{for all } s \in \mathcal{S}$$

then π_1 is "**better**" than π_2 .

Definition

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

The definition leads to many questions:

- Does the optimal policy exist?
- Is the optimal policy unique?
- Is the optimal policy stochastic or deterministic?
- How to obtain the optimal policy?

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

1. Motivating examples

2. Definition of optimal policy

3. BOE: Introduction

4. BOE: Preliminaries

4.1 BOE: Maximization on the right-hand side

4.2 BOE: Rewrite as $v = f(v)$

4.3 Contraction mapping theorem

5. BOE: Solution

6. BOE: Optimality

7. Analyzing optimal policies

Bellman optimality equation (BOE)

Bellman optimality equation (elementwise form):

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

Remarks:

- ▶ $p(r | s, a), p(s' | s, a), r, \gamma$ are known.
- ▶ $v(s), v(s')$ are unknown and to be calculated.
- ▶ Is $\pi(s)$ known or unknown?

Bellman optimality equation (BOE)

Bellman optimality equation (matrix-vector form):

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

where the elements corresponding to s or s' are

$$[r_{\pi}]_s \triangleq \sum_a \pi(a | s) \sum_r p(r | s, a) r,$$

$$[P_{\pi}]_{s,s'} = p(s' | s) \triangleq \sum_a \pi(a | s) \sum_{s'} p(s' | s, a)$$

Here \max_{π} is performed elementwise:

$$\max_{\pi} \begin{bmatrix} * \\ \vdots \\ * \end{bmatrix} = \begin{bmatrix} \max_{\pi(s_1)} * \\ \vdots \\ \max_{\pi(s_n)} * \end{bmatrix}$$

Bellman optimality equation (BOE)

Bellman optimality equation (matrix-vector form):

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

BOE is tricky yet elegant!

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

Questions:

- Algorithm: how to solve this equation?
- Existence: Does this equation have solutions?
- Uniqueness: Is the solution to this equation unique?
- Optimality: How is it related to optimal policy?

1. Motivating examples

2. Definition of optimal policy

3. BOE: Introduction

4. BOE: Preliminaries

4.1 BOE: Maximization on the right-hand side

4.2 BOE: Rewrite as $v = f(v)$

4.3 Contraction mapping theorem

5. BOE: Solution

6. BOE: Optimality

7. Analyzing optimal policies

1. Motivating examples
2. Definition of optimal policy
3. BOE: Introduction
4. BOE: Preliminaries
 - 4.1 BOE: Maximization on the right-hand side
 - 4.2 BOE: Rewrite as $v = f(v)$
 - 4.3 Contraction mapping theorem
5. BOE: Solution
6. BOE: Optimality
7. Analyzing optimal policies

Maximization on the right-hand side of BOE

BOE: elementwise form

$$v(s) = \max_{\pi} \sum_a \pi(a | s) \left(\sum_r p(r | s, a) r + \gamma \sum_{s'} p(s' | s, a) v(s') \right), \quad \forall s \in \mathcal{S}$$

BOE: matrix-vector form $v = \max_{\pi} (r_{\pi} + \gamma P_{\pi} v)$

Example (How to solve two unknowns from one equation)

Solve two unknown variables $x, a \in \mathbb{R}$ from the following equation:

$$x = \max_a (2x - 1 - a^2)$$

To solve them, first consider the right-hand side. Regardless the value of x , $\max_a (2x - 1 - a^2) = 2x - 1$ 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 solutions of the equation.

Maximization on the right-hand side of BOE

Fix $v'(s)$ first and solve π :

$$\begin{aligned} v(s) &= \max_{\pi} \sum_a \pi(a | s) \left(\sum_r p(r | s, a) r + \gamma \sum_{s'} p(s' | s, a) v(s') \right), \quad \forall s \in \mathcal{S} \\ &= \max_{\pi} \sum_a \pi(a | s) q(s, a) = \max_{\pi} [\pi(a_1 | s) q(s, a_1) + \dots + \pi(a_5 | s) q(s, a_5)] \\ &\doteq \max_{c_1, \dots, c_5} [c_1 q(s, a_1) + \dots + c_5 q(s, a_5)], \quad c_1 + \dots + c_5 = 1 \end{aligned}$$

Example (How to solve $\max_{\pi} \sum_a \pi(a | s) q(s, a)$)

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

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

where $c_1 + c_2 + c_3 = 1$ and $c_1, c_2, c_3 \geq 0$.

Answer: 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

$$q_3 = (c_1 + c_2 + c_3) q_3 = c_1 q_3 + c_2 q_3 + c_3 q_3 \geq c_1 q_1 + c_2 q_2 + c_3 q_3$$

Maximization on the right-hand side of BOE

Inspired by the above example, considering that $\sum_a \pi(a | s) = 1$, we have

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

where the optimality is achieved when

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

where $a^* = \arg \max_a q(s, a)$.

1. Motivating examples
2. Definition of optimal policy
3. BOE: Introduction
4. BOE: Preliminaries
 - 4.1 BOE: Maximization on the right-hand side
 - 4.2 BOE: Rewrite as $v = f(v)$
 - 4.3 Contraction mapping theorem
5. BOE: Solution
6. BOE: Optimality
7. Analyzing optimal policies

Solve the Bellman optimality equation

The BOE is $v = \max_{\pi} (r_{\pi} + \gamma P_{\pi} v)$. Let

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

Then, the Bellman optimality equation becomes

$$v = f(v)$$

where

$$[f(v)]_s = \max_{\pi} \sum_a \pi(a | s) q(s, a), \quad s \in \mathcal{S}$$

This equation looks very simple. How to solve it?

1. Motivating examples
2. Definition of optimal policy
3. BOE: Introduction
4. BOE: Preliminaries
 - 4.1 BOE: Maximization on the right-hand side
 - 4.2 BOE: Rewrite as $v = f(v)$
 - 4.3 Contraction mapping theorem
5. BOE: Solution
6. BOE: Optimality
7. Analyzing optimal policies

Preliminaries: Contraction mapping theorem

Some concepts:

- **Fixed point:** $x \in X$ is a fixed point of $f: X \rightarrow X$ if

$$f(x) = x$$

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

$$\|f(x_1) - f(x_2)\| \leq \gamma \|x_1 - x_2\|$$

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

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

Preliminaries: Contraction mapping theorem

Examples to demonstrate the concepts.

Example



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

It is easy to verify that $x = 0$ is a fixed point since $0 = 0.5 \times 0$. Moreover, $f(x) = 0.5x$ is a contraction mapping because

$$\|0.5x_1 - 0.5x_2\| = 0.5 \|x_1 - x_2\| \leq \gamma \|x_1 - x_2\| \text{ for any } \gamma \in [0.5, 1).$$



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

It is easy to verify that $x = 0$ is a fixed point since $0 = A0$.

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.

Preliminaries: Contraction mapping theorem

Theorem (Contraction Mapping Theorem)

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

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

Preliminaries: Contraction mapping theorem

Examples:

- ▶ $x = 0.5x$, where $f(x) = 0.5x$ and $x \in \mathbb{R}$
 $x^* = 0$ is the unique fixed point. It can be solved iteratively by

$$x_{k+1} = 0.5x_k$$

- ▶ $x = Ax$, where $f(x) = Ax$ and $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ and $\|A\| < 1$ $x^* = 0$ is the unique fixed point. It can be solved iteratively by

$$x_{k+1} = Ax_k$$

1. Motivating examples
2. Definition of optimal policy
3. BOE: Introduction
4. BOE: Preliminaries
 - 4.1 BOE: Maximization on the right-hand side
 - 4.2 BOE: Rewrite as $v = f(v)$
 - 4.3 Contraction mapping theorem
5. BOE: Solution
6. BOE: Optimality
7. Analyzing optimal policies

Contraction property of BOE

Let's come back to the Bellman optimality equation:

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

Theorem (Contraction Property)

$f(v)$ is a contraction mapping satisfying

$$\|f(v_1) - f(v_2)\| \leq \gamma \|v_1 - v_2\|$$

where γ is the discount rate!

Solve the Bellman optimality equation

Applying the contraction mapping theorem gives the following results.

Theorem (Existence, Uniqueness, and Algorithm)

For the BOE $v = f(v) = \max_{\pi} (r_{\pi} + \gamma P_{\pi} v)$, there always exists a solution v^* and the solution is unique. The solution could be solved iteratively by

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

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

Important: The algorithm in (1) is called the value iteration algorithm. We will analyze it in the following lecture! This lecture focuses more on the fundamental properties.

1. Motivating examples

2. Definition of optimal policy

3. BOE: Introduction

4. BOE: Preliminaries

4.1 BOE: Maximization on the right-hand side

4.2 BOE: Rewrite as $v = f(v)$

4.3 Contraction mapping theorem

5. BOE: Solution

6. BOE: Optimality

7. Analyzing optimal policies

Policy optimality

Suppose v^* is the solution to the Bellman optimality equation. It satisfies

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

Suppose

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

Then

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

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

Is π^* the optimal policy? Is v^* the greatest state value can be achieved?

Policy optimality

Theorem (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

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

Now we understand why we study the BOE. That is because it describes the optimal state value and optimal policy.

Policy optimality

What does an optimal policy π^* look like?

$$\pi^*(s) = \arg \max_{\pi} \sum_a \pi(a | s) \underbrace{\left(\sum_r p(r | s, a) r + \gamma \sum_{s'} p(s' | s, a) v^*(s') \right)}_{q^*(s, a)}$$

Theorem (**Greedy Optimal Policy**)

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

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

is an optimal policy solving the BOE. Here,

$$a^*(s) = \arg \max_a q^*(a, s),$$

where $q^*(s, a) \doteq \sum_r p(r | s, a) r + \gamma \sum_{s'} p(s' | s, a) v^*(s')$.

1. Motivating examples
2. Definition of optimal policy
3. BOE: Introduction
4. BOE: Preliminaries
 - 4.1 BOE: Maximization on the right-hand side
 - 4.2 BOE: Rewrite as $v = f(v)$
 - 4.3 Contraction mapping theorem
5. BOE: Solution
6. BOE: Optimality
7. Analyzing optimal policies

Analyzing optimal policies

What factors determine the optimal state value and optimal policy?

It can be seen from the BOE

$$v(s) = \max_{\pi} \sum_a \pi(a | s) \left(\sum_r p(r | s, a) r + \gamma \sum_{s'} p(s' | s, a) v(s') \right)$$

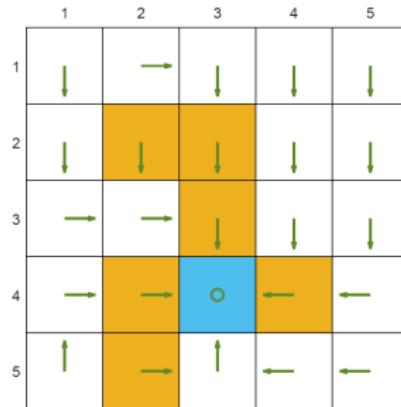
that there are three factors:

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

We next show how r and γ can affect the optimal policy.

Analyzing optimal policies

The optimal policy and the corresponding optimal state value are obtained by solving the BOE.



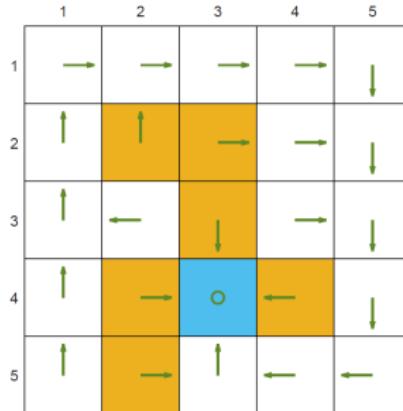
	1	2	3	4	5
1	5.8	5.6	6.2	6.5	5.8
2	6.5	7.2	8.0	7.2	6.5
3	7.2	8.0	10.0	8.0	7.2
4	8.0	10.0	10.0	10.0	8.0
5	7.2	9.0	10.0	9.0	8.1

(a) $r_{\text{boundary}} = r_{\text{forbidden}} = -1, r_{\text{target}} = 1, \gamma = 0.9$

The optimal policy dares to take risks: entering forbidden areas!!

Analyzing optimal policies

If we change $\gamma = 0.9$ to $\gamma = 0.5$



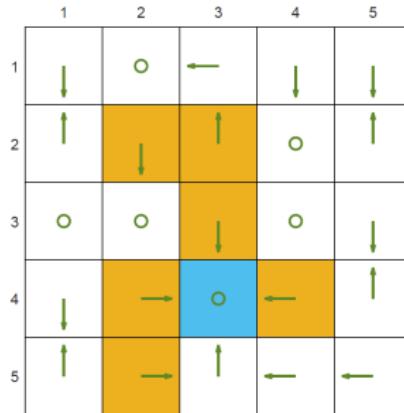
	1	2	3	4	5
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.1
3	0.0	0.0	2.0	0.1	0.1
4	0.0	2.0	2.0	2.0	0.2
5	0.0	1.0	2.0	1.0	0.5

- (b) The discount rate is $\gamma = 0.5$. Others are the same as (a).

The optimal policy becomes short-sighted! Avoid all the forbidden areas!

Analyzing optimal policies

If we change γ to 0



	1	2	3	4	5
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0	0.0
4	0.0	1.0	1.0	1.0	0.0
5	0.0	0.0	1.0	0.0	0.0

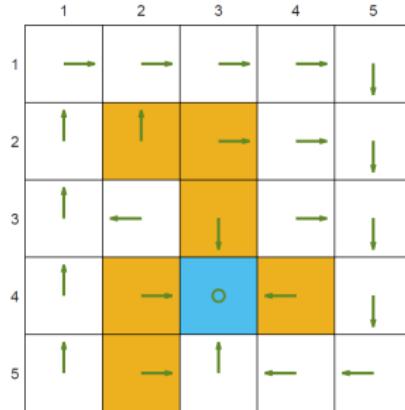
- (c) The discount rate is $\gamma = 0$. Others are the same as (a).

The optimal policy becomes extremely short-sighted! Also, choose the action that has the greatest immediate reward! Cannot reach the target!

Analyzing optimal policies

If we increase the punishment when entering forbidden areas: change

$r_{\text{forbidden}} = -1$ to $r_{\text{forbidden}} = -10$



	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

(d) $r_{\text{forbidden}} = -10$. Others are the same as (a).

The optimal policy would also avoid the forbidden areas.

Analyzing optimal policies

What if we change $r \rightarrow ar + b$?

For example,

$$r_{\text{boundary}} = r_{\text{forbidden}} = -1, \quad r_{\text{target}} = 1, \quad r_{\text{otherstep}} = 0$$

becomes

$$r_{\text{boundary}} = r_{\text{forbidden}} = 0, \quad r_{\text{target}} = 2, \quad r_{\text{otherstep}} = 1$$

The optimal policy remains the same!

What matters is not the absolute reward values! It is their relative values!

Analyzing optimal policies

Theorem (Optimal Policy Invariance)

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

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

where $\gamma \in (0, 1)$ is the discount rate and $\mathbf{1} = [1, \dots, 1]^T$.

Consequently, the optimal policies are invariant to the affine transformation of the reward signals.

Summary

Bellman optimality equation:

- ▶ Elementwise form:

$$v(s) = \max_{\pi} \sum_a \pi(a | s) \underbrace{\left(\sum_r p(r | s, a) r + \gamma \sum_{s'} p(s' | s, a) v(s') \right)}_{q(s, a)}, \quad \forall s \in \mathcal{S}$$

- ▶ Matrix-vector form:

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

Summary

Questions about the Bellman optimality equation:

- ▶ Does this equation have solutions? - Existence
 - Yes, by the contraction mapping theorem
- ▶ Is the solution to this equation unique? - Uniqueness
 - Yes, by the contraction mapping theorem
- ▶ How to solve this equation? - Algorithm
 - Iterative algorithm suggested by the contraction mapping theorem
- ▶ Why we study this equation - Optimality
 - Because its solution corresponds to the optimal state value and optimal policy.