Chapter 3. Bellman Optimality Equation

Outline:

① Core concepts = optimal state value and optimal policy

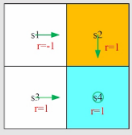
② A fundamental tool = the Bellman optimality equation (BOE)

寻找最优策略.

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

1. Motivating examples.

Motivating examples



π

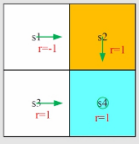
Bellman equation:

$$\begin{cases} v_\pi(s_1) = -1 + \gamma v_\pi(s_2), \\ v_\pi(s_2) = +1 + \gamma v_\pi(s_4), \\ v_\pi(s_3) = +1 + \gamma v_\pi(s_4), \\ v_\pi(s_4) = +1 + \gamma v_\pi(s_4). \end{cases}$$

State value: Let $\gamma = 0.9$. Then, it can be calculated that

$$v_\pi(s_4) = v_\pi(s_3) = v_\pi(s_2) = 10, \quad v_\pi(s_1) = 8.$$

Motivating examples

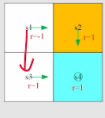


Action value: consider s_1

$$q_\pi(s_1, a_1) = -1 + \gamma v_\pi(s_1) = 6.2,$$
$$q_\pi(s_1, a_2) = -1 + \gamma v_\pi(s_2) = 8,$$
$$q_\pi(s_1, a_3) = 0 + \gamma v_\pi(s_3) = 9,$$
$$q_\pi(s_1, a_4) = -1 + \gamma v_\pi(s_1) = 6.2,$$
$$q_\pi(s_1, a_5) = 0 + \gamma v_\pi(s_1) = 7.2.$$

While the policy is not good, how can we improve it by using action values. ☆

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



Observe the action values that we obtained just now:

$$q_\pi(s_1, a_1) = 6.2, q_\pi(s_1, a_2) = 8, q_\pi(s_1, a_3) = 9,$$
$$q_\pi(s_1, a_4) = 6.2, q_\pi(s_1, a_5) = 7.2.$$

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

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

where $a^* = \arg\max_a q_\pi(s_1, a) = a_3$

当其他action并非最优时
s_1选择可能并非最优
但通过全局取优，遍历迭代
可以得到最优解（证明）

2. Optimal policy

$V_{\pi_1}(s) \geq V_{\pi_2}(s)$ for all $s \in S$

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 π

The definition leads to many questions:

- Does the optimal policy exist? ← 存在?
- Is the optimal policy unique? 唯一?
- Is the optimal policy stochastic or deterministic? 确定性?
- How to obtain the optimal policy? 怎么得到

贝尔曼最优公式可以解答.

Bellman Optimal Equation.

3, BOE: (Introduction):

Bellman optimality equation (elementwise form):

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

$$= \max_{\pi} \sum_{a} \pi(a|s) q(s,a) \quad s \in \mathcal{S}$$

Remarks:

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

矩阵形式

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

4. BOE : Maximization on the right-hand side.

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)

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

$$x = \max_{a}(2x - 1 - a^2).$$

This equation has two unknowns. 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 solution of the equation.

Fix $v'(s)$ first and solve π :

给定一个值.

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

已知

$$= \max_{\pi} \sum_{a} \pi(a|s) q(s,a)$$

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 $c_1, c_2, c_3 \geq 0$ → 概率 π.

Suppose $q_3 \geq q_1, q_2$ Then $c_3^* = 1$, $c_1^* = c_2^* = 0$

∵ considering that: $\sum_a \pi(a|s) = 1$

$$\max_\pi \sum_a \pi(a|s)\, q(s,a) = \max_{a \in A(s)} q(s,a)$$

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

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

最大 $q(s,a)$ 对应 a^*

BOE: $v = \max_\pi (r_\pi + \gamma P_\pi v)$

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

$$v = f(v)$$

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

6. Preliminaries: Contraction mapping theorem.

① Fixed point: $x \in X$ is a fix point of f:

$X \xrightarrow[\text{自映射}]{} X$ if $f(x) = x$

② Contraction mapping (or contractive function)

f is a contraction mapping if

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

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

$\|\cdot\|$ can be any vector norm

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
  $\|0.5x_1 - 0.5x_2\| = 0.5\|x_1 - x_2\| \leq \gamma\|x_1 - x_2\|$ for any $\gamma \in [0.5, 1)$.
- $x = f(x) = Ax$, 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 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.

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 $x_k \to x^*$ as $k \to \infty$. Moreover, the convergence rate is exponentially fast.*

For the proof of this theorem, see the book.

不动点存在，唯一，其可迭代得到

$$V = f(V) = \max_{\pi}(R_\pi + \gamma P_\pi V)$$

Theorem (Contraction Property)

$f(V)$ is a contraction mapping satisfying

$$\|f(V_1) - f(V_2)\| \le \gamma \|V_1 - V_2\|$$

∴ 不动点唯一存在 V^* 且可迭代获得.

7. 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

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

π^*, $v^* = v_{\pi^*}$

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?

只有最优公式是最优策略下的贝尔曼公式

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

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

For the proof, please see our book.

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

V^* 是最大的 state value.

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

8. Analyzing optimal policies

What factors determine the optimal policy?

It can be clearly seen from the BOE

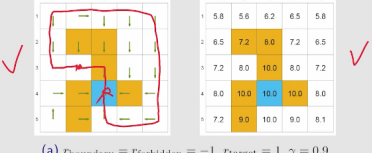
$$v(s) = \max_{\pi} \sum_{a} \pi(a|s) \left(\sum_{r} p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s') \right)$$

that there are three factors:

- Reward design: r
- System model: $p(s'|s,a)$, $p(r|s,a)$
- Discount rate: γ
- $v(s), v(s'), \pi(a|s)$ are unknowns to be calculated

Next, we use examples to show how changing r and γ can change the optimal policy.

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



(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!!

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

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

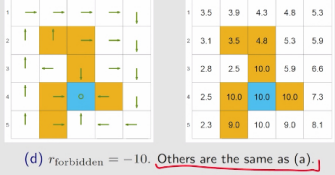
The optimal policy becomes shorted-sighted! Avoid all the forbidden areas!

If we change γ to 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!

If we increase the punishment when entering forbidden areas ($r_{\text{forbidden}} = -1$ to $r_{\text{forbidden}} = -10$)



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

The optimal policy would also avoid the forbidden areas.

What if we change $r \to ar + b$?
For example,

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

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!

策略不会变

相对量更关键

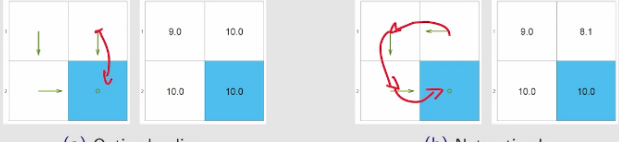
Theorem (Optimal Policy Invariance)

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 $a \neq 0$, 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 $\mathbf{1} = [1, \ldots, 1]^T$. Consequently, the optimal policies are invariant to the affine transformation of the reward signals.

Meaningless detour?



(a) Optimal policy (b) Not optimal

The policy in (a) is optimal, the policy in (b) is not.

Question: Why the optimal policy is not (b)? Why does the optimal policy not take meaningless detours? There is no punishment for taking detours!!

Due to the discount rate!

Policy (a): return $= 1 + \gamma 1 + \gamma^2 1 + \cdots = 1/(1-\gamma) = 10.$

Policy (b): return $= 0 + \gamma 0 + \gamma^2 1 + \gamma^3 1 + \cdots = \gamma^2/(1-\gamma) = 8.1$

否则 return = 0

也能搜索到最短路径

9. Summary:

① Bellman optimality equation:

elementwise form

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

Matrix-vector form

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

Questions about the Bellman optimality equation:

- Existence: does this equation have solutions?
  - Yes, by the contraction mapping Theorem
- Uniqueness: is the solution to this equation unique?
  - Yes, by the contraction mapping Theorem
- Algorithm: how to solve this equation?
  - Iterative algorithm suggested by the contraction mapping Theorem
- Optimality: why we study this equation
  - Because its solution corresponds to the optimal state value and optimal policy.

Finally, we understand why it is important to study the BOE!