**Subject:** Westlake University, Reinforce Learning, Lecture 3, Optimal Policy and Bellman Optimality Equation

**Date:** from December 21, 2024 to January 15, 2025

---

## Contents

# Lecture 3, Optimal Policy and Bellman Optimality Equation

Bilibili:Lecture 3, Bellman Optimality Equation

**Outline**

本节将介绍一个核心概念和一个重要工具：

**1 core concepts:** optimal state value (进而定义 optimal policy)

**1 fundamental tool:** Bellman optimality equation(BOE) (求解 optimal state value)

---

本节内容将回答如下问题：

1. 什么是 optimal state，optimal policy
2. optimal policy 是否一定存在？若存在长什么样？
3. 怎么求解 optimal state，optimal policy

## Optimal state values and optimal policies

**Optimal Policy**

State value 可以衡量一个 policy 的好坏，因此比较不同的 policy 需要使用 state value. 首先回忆 state value 的定义：

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

**Definition 1** (Better policy)**.** 对于两个 *policy* $\pi_1, \pi_2$，若他们的 *state value* 有如下关系：

$$v_{\pi_1}(s) \geqslant v_{\pi_2}(s), \quad \forall s \in \mathcal{S}$$

则称 *policy* $\pi_1$ 好于 $\pi_2$，其中 $\mathcal{S}$ 为 *state space.*

**Definition 2** (optimal policy & optimal state value)**.** 若 *policy* $\pi^*$ 的 *state value* 有：

$$v_{\pi^*}(s) \geqslant v_\pi(s), \quad \forall s \in \mathcal{S}, \forall \pi$$

则称 *policy* $\pi^*$ 为 *optimal policy.* 称 $\pi^*$ 的 *state values* 为 *optimal state values.*

这就产生了如下问题（都可以用贝尔曼最优公式回答）：

1. **Existence & Uniqueness**: optimal policy 是否存在？若存在，是否唯一？
2. **Stochasticity**: optimal policy 是确定性的(deterministic) 还是随机性的(stochastic)？
3. **Algorithm**: 如何得到 optimal policy 和 optimal state values？

## Bellman optimality equation

**Bellman optimality equation(BOE)**

分析 optimal policy 和 optimal state values 的重要工具是 Bellman optimality equation (BOE)，与 Bellman equation 相似，通过求解 BOE 就可以得到 optimal policy 和 optimal state values. 下面先回忆 Bellman equation :

$$\begin{aligned} v_\pi(s) &= \underbrace{\sum_{a\in\mathcal{A}}\pi(a|s)\sum_{r\in\mathcal{R}}p(r|s,a)r}_{\text{mean of immediate rewards}} + \underbrace{\gamma\sum_{a\in\mathcal{A}}\pi(a|s)\sum_{s'\in\mathcal{S}}p(s'|s,a)v_\pi(s')}_{\text{mean of future rewards}} \\ &= \sum_{a\in\mathcal{A}}\pi(a|s)\left[\sum_{r\in\mathcal{R}}p(r|s,a)r + \gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)v_\pi(s')\right], \quad \forall s\in\mathcal{S} \end{aligned}$$

Bellman optimality equation 则是将 policy $\pi$固定为最优的（取 max）：

$$\begin{aligned} v(s) &= \max_\pi \sum_{a\in\mathcal{A}}\pi(a|s)\left(\sum_{r\in\mathcal{R}}p(r|s,a)r + \gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)v(s')\right), \quad \forall s\in\mathcal{S} \\ &\triangleq \max_\pi \sum_{a\in\mathcal{A}}\pi(a|s)q(s,a) \end{aligned} \tag{1}$$

式(1)是 elementwise form 的 BOE. 写成 matrix-vector form 就是:

$$\boldsymbol{v} = \max_\pi(\boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi \boldsymbol{v}) \tag{2}$$

$$\begin{aligned} [r_\pi]_s &\triangleq \sum_{a\in\mathcal{A}}\pi(a|s)\sum_{r\in\mathcal{R}}p(r|s,a)r, \\ [P_\pi]_{s,s'} &\triangleq p(s'|s) \triangleq \sum_{a\in\mathcal{A}}\pi(a|s)\sum_{s'\in\mathcal{S}}p(s'|s,a) \end{aligned} \tag{3}$$

其中

$$\underbrace{\max_\pi \begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ \vdots \\ v_\pi(s_n) \end{bmatrix}}_{\max_\pi \boldsymbol{v}_\pi} = \begin{bmatrix} \max_\pi v_\pi(s_1) \\ \max_\pi v_\pi(s_2) \\ \vdots \\ \max_\pi v_\pi(s_n) \end{bmatrix}$$

**Note 1.** *1.* $p(r|s,a), p(s'|s,a)$ 作为模型均已知

*2.* $v(s)$, $v(s')$ 均为需求解的未知量

*3. Bellman equation* 中 *policy* 固定，但 *BOE(1)*中需要求解出最优 *policy*

同样地，也产生了如下问题:

1. **Existence & Uniqueness**: BOE 是否有解？若有解是否唯一？
2. **Algorithm**: 如何求解 BOE？
3. **Optimality**: BOE 的解与 optimal policy 有何关系？

### Solving an optimal policy from the BOE

**How to maximize the right-hand side of BOE**

由式(2)可以看出，式中有两个未知量需要求解，即 $\boldsymbol{v}$ 和 $\pi$，也就是说一个式子求两个未知量：

$$\boldsymbol{v} = \max_{\pi}(\boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi \boldsymbol{v})$$

解决方式是先固定住 $\boldsymbol{v}$，将其看作常量，然后对式子整体求 $\pi$ 能让整个式子达到最大的取值，记为 $\hat{\pi}$，固定之后得到方程 $\boldsymbol{v} = \boldsymbol{r}_{\hat{\pi}} + \gamma \boldsymbol{P}_{\hat{\pi}} \boldsymbol{v}$，即可求解得到 $\boldsymbol{v}$.

**Note 2.** 下面给出一个简单的例子说明求解过程，例如：

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

首先固定 $x$，显然当 $a = 0$ 时整个式子才可能有最大值，那么就得到

$$x = 2x - 1 \quad \Rightarrow \quad x = 1$$

在式(1)中，实际上会对 $v(s')$ 赋予初始值，这样实际上 $q(s, a)$ 就是已知的，那么只需要确定出 $\pi(a|s)$ 即可.

由于 $\sum_{a \in \mathcal{A}} \pi(a \mid s) = 1$，记 $\max\limits_{a \in \mathcal{A}(s)} q(s, a) = q(s, a^*)$由于

$$q(s, a^*) = \underbrace{\left(\sum_{a \in \mathcal{A}} \pi(a \mid s)\right)}_{=1} \cdot q(s, a^*) \geqslant \sum_{a \in \mathcal{A}} \pi(a|s) \cdot q(s, a)$$

那么实际上只需取最大的 $q(s, a)$ 即可：

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

$$\pi(a \mid s) = \begin{cases} 1, & a = a^* \\ 0, & a \neq a^* \end{cases}, \quad a^* = \arg\max_{a} q(s, a).$$

此时由于 $\pi$ 选择了最大的 $q$ 值，因此被称为 greedy policy.

**Solve the BOE**

根据我们最大化 BOE 右端项的思想，我们先固定$v$不动，再找到能够使整个式子最大的$\pi$，最后整个式子只剩下一个变量$v$，此时我们将其记为

$$f(v) \triangleq \max_{\pi} (r_\pi + \gamma P_\pi v) \tag{4}$$

这样 BOE 就变成

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

根据压缩映射原理(Contraction Mapping Theorem) / 不动点定理(Fixed Point Theorem)，对于形如 $x = f(x)$ 的方程，若 $f$ 是压缩映射（见附录A），那么有如下结论：

1. 存在唯一不动点满足 $x^* = f(x^*)$，即方程的解是存在唯一的
2. 算法 $x_{k+1} = f(x_k)$可以不断逼近此不动点，且以指数速率收敛

事实上，BOE 中的映射 $f$ 刚好是一个压缩映射，即定理1(证明见附录B).

**Theorem 1** (Contraction property of $f(v)$)**.** *BOE* 中的映射 $f(v)$ 是一个压缩映射，$\forall\, v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$，满足

$$\|f(v_1) - f(v_2)\|_\infty \le \gamma \|v_1 - v_2\|_\infty$$

其中 $\gamma \in (0,1)$ 为 *discount rate*，$\|\cdot\|_\infty$ 为 *maximum norm(*取绝对值最大者*).*

那么自然地就可以得到如下重要定理:

**Theorem 2** (Existence, Uniqueness, and Algorithm)**.** 对于 *BOE* $v = f(v) = \max\limits_{\pi}\,(r_\pi + \gamma P_\pi v)$，总存在唯一解，且该解可以被如下方式迭代逼近:

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

且产生的序列 $\{v_k\}$ 指数收敛至不动点 $v^*$，收敛速率由 $\gamma$ 控制.

**Policy optimality**

当我们求出了最优 policy $v^*$ 后，$v^*$ 显然满足

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

注意此时的 $\pi$ 都是固定的最优的，不妨将其记为 $\pi^*$，那么为达到 max，其满足

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

这样就将 BOE 转化为一个特殊的 BE:

$$v^* = r_{\pi^*} + \gamma P_{\pi^*} v^* \tag{5}$$

因此说 BOE 是特殊情形的 BE.

关于 BOE 解的最优性，有如下结论(证明见附录C):

**Theorem 3.** $v^*$ 是最优 *state value*，$\pi^*$是最优的 *policy*，即

$$v^* = v_{\pi^*} \geqslant v_\pi, \forall \pi, \forall v$$

**Theorem 4** (Greedy Optimal Policy)**.** $\forall s \in \mathcal{S}$，*BOE* 的最优 *policy (*也称为 *deterministic greedy policy)*为:

$$\pi^*(a \mid s)_\star = \begin{cases} 1 & \textit{if } a = a^*(s) \\ 0 & \textit{if } a \neq a^*(s) \end{cases} \tag{6}$$

其中

$$\begin{aligned} a^*(s) &= \arg\max_a q^*(a, s) \\ q^*(s,a) &:= \sum_{r \in \mathcal{R}} p(r \mid s, a) r + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) v^*(s') \end{aligned}$$

此外还有两点说明：最优策略的唯一性和最优策略的随机性

**Uniqueness of optimal policies**: 最优的 state value $v^*$ 是唯一确定的，但是 optimal policy 并不一定，有可能出现两个 policy 均为最优.

**Stochasticity of optimal policies**: 从 optimal policy 并不一定唯一就可以看出 optimal policy 既可以是随机的也可以是确定性的，但根据定理4 可以确定的是一定存在一个确定性的 optimal policy.

**Analyzing optimal policies**

对于 BOE

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

黑色部分是未知并需求解的量，红色部分为已知量，可能对最终结果造成影响，其中

1. $r$: 预先设计的 reward
2. $p(r|s,a), p(s'|s,a)$: 概率模型/系统模型(system model)
3. $\gamma$: discount rate

由于系统的模型一般难以改变，所以下面仅分析 $r$ 和 $\gamma$ 的改变对 BOE 结果的影响.

根据一些简答的例子（详见 textbook）可以发现:

1. 当 $\gamma$ 比较大时，agent 会比较“远视”，重视未来的 reward；较小时，会比较“近视”(short-sighted)，重视较近的 reward.
   (a) 当 reward；较小时，agent 会倾向于避免冒险，更多地选择眼前看起来较好的 action.
   (b) 当 reward = 0 时，agent 甚至不能成功抵达目标，因为此时只选择最大的 immediate reward 而不是最大的 total reward.
   (c) discount rate 的存在使得一些无意义的绕远路(meaningless detour)的策略被 pass，因为这样的 reward 会被延后且“打折”
2. 当对所有的 reward 作线性变换$r \to ar + b$时，optimal policy 并不会改变(定理5，证明见附录D)，因为重要的是不同 reward 相互间的相对差异(relative value)，而不是绝对差异

**Theorem 5** (Optimal Policy Invariance). 考虑一马尔可夫决策过程，其中$v^* \in \mathbb{R}^{|S|}$ 是满足

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

的 *optimal state value*. 如果每个奖励 $r \in \mathbb{R}$ 进行仿射变换 $\alpha r + \beta$，其中 $\alpha, \beta \in \mathbb{R}$ 且$\alpha > 0$，则相应的最优状态值$v'$ 也将是$v^*$ 的仿射变换:

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

其中$\gamma \in (0,1)$ 是 *discount rate*，且 $\mathbf{1} = [1, \ldots, 1]^T$。因此，从 $v'$ 导出的 *optimal state value* 对于 *reward* 的仿射变换是保持不变的。

## A Contraction Mapping Theorem

**Contraction Mapping Theorem**

**Definition 3** (fixed point)**.** 对于方程 $f(x), f:\mathbb{R}^d \to \mathbb{R}^d$，点 $x^*$ 被称为不动点*(fixed point)* 若

$$f(x^*) = x^*$$

**Definition 4** (contraction mapping / contractive function)**.** 函数 $f:\mathbb{R}^d \to \mathbb{R}^k$ 被称为压缩映射*(contraction mapping / contractive function)*，若 $\exists\, \gamma \in (0,1)$，使得

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

**Theorem 6** (Contraction mapping theorem)**.** 对于方程 $f(x) = x$，若 $f$ 为压缩映射，那么存在唯一不动点作为解满足 $f(x^*) = x^*$. 且可以设计如下算法迭代逼近不动点 $x^*$:

$$x_{k+1} = f(x_k)$$

满足 $x_k \xrightarrow{k\to\infty} x^*$，且以指数速率收敛.

*Proof.* 由于在完备赋范线性空间（即 Banach 空间）中，Cauchy 序列必收敛，因此需先证明序列 $\{x_k = f(x_{k-1})\}_{k=1}^{\infty}$ 是 Cauchy 列. 即 $\forall\, \epsilon > 0, \exists\, N > 0, s.t. \forall\, m, n > N, \|x_m - x_n\| < \epsilon$. 然后再证明该收敛点为唯一不动点.

① 先证明序列 $\{x_k = f(x_{k-1})\}_{k=1}^{\infty}$ 是 Cauchy 列. 根据压缩性，有

$$\begin{aligned}\|x_{n+1} - x_n\| &= \|f(x_n) - f(x_{n-1})\| \leqslant \gamma \|x_n - x_{n-1}\| = \gamma \|f(x_{n-1}) - f(x_{n-2})\| \\ &\leqslant \gamma^2 \|x_{n-1} - x_{n-2}\| = \gamma^2 \|f(x_{n-2}) - f(x_{n-3})\| \leqslant \cdots \leqslant \gamma^n \|x_1 - x_0\|\end{aligned}$$

$\forall\, \epsilon > 0$，取 $N \geqslant \log\left(\frac{\epsilon}{\|x_1 - x_0\|}(1-\gamma) - \gamma\right)$，那么 $\forall\, m, n > N$ 有

$$\begin{aligned}\|x_m - x_n\| &\leqslant \|x_m - x_{m-1}\| + \cdots + \|x_{n+1} - x_n\| \leqslant \left(\gamma^{m-1} + \cdots + \gamma^n\right) \|x_1 - x_0\| \\ &= \frac{\gamma^n (1 - \gamma^{m-n})}{1-\gamma} \|x_1 - x_0\| \leqslant \frac{\gamma^n}{1-\gamma} \|x_1 - x_0\| \leqslant \epsilon\end{aligned}$$

因此 $\{x_k = f(x_{k-1})\}_{k=1}^{\infty}$ 是 Cauchy 列.

② 设 Cauchy 列 $\{x_k = f(x_{k-1})\}_{k=1}^{\infty}$ 收敛至 $x^*$. 由于 $\|f(x_k) - x_k\| = \|x_{k+1} - x_k\| \leqslant \gamma^k \|x_1 - x_0\| \xrightarrow{k\to\infty} 0$，因此取极限 $k \to \infty$ 后有 $f(x^*) = x^*$. 因此 $x^*$ 是不动点.

③ 证明不动点的唯一性. 使用反证法，若不动点不唯一，设 $x_1^*, x_2^*$ 均为不动点，那么就有

$$\|x_2^* - x_1^*\| = \|f(x_2^*) - f(x_1^*)\| \leqslant \gamma \|x_2^* - x_1^*\| < \|x_2^* - x_1^*\|$$

显然矛盾.

综上所述，方程 $f(x) = x$ 存在唯一解. □

## B Contraction property of $f(v)$

**Contraction property of $f(v)$**

**Theorem 7.** *BOE 中的映射 $f(v)$ 是一个压缩映射，$\forall\, v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$，满足*

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

*其中 $\gamma \in (0,1)$ 为 discount rate，$\|\cdot\|_\infty$ 为 maximum norm.*

*Proof.* 根据定义，$f(v) \triangleq \max_\pi (r_\pi + \gamma P_\pi v)$. $\forall\, v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$，假设 $\pi_1^* \triangleq \arg\max_\pi (r_\pi + \gamma P_\pi v_1)$, $\pi_2^* \triangleq \arg\max_\pi (r_\pi + \gamma P_\pi v_2)$，那么就有

$$\begin{aligned}
f(v_1) &= \max_\pi (r_\pi + \gamma P_\pi v_1) = r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1 \geqslant r_{\pi_2^*} + \gamma P_{\pi_2^*} v_1, \\
f(v_2) &= \max_\pi (r_\pi + \gamma P_\pi v_2) = r_{\pi_2^*} + \gamma P_{\pi_2^*} v_2 \geqslant r_{\pi_1^*} + \gamma P_{\pi_1^*} v_2,
\end{aligned}$$

其中 $\geqslant$ 为逐元素比较. 这样就可以得到

$$\begin{aligned}
f(v_1) - f(v_2) &= r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1 - (r_{\pi_2^*} + \gamma P_{\pi_2^*} v_2) \\
&\leq r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1 - (r_{\pi_1^*} + \gamma P_{\pi_1^*} v_2) \\
&= \gamma P_{\pi_1^*} (v_1 - v_2).
\end{aligned}$$

同理可得

$$\gamma P_{\pi_2^*}(v_1 - v_2) \leqslant f(v_1) - f(v_2) \leqslant \gamma P_{\pi_1^*}(v_1 - v_2)$$

定义

$$z \triangleq \max\left\{|\gamma P_{\pi_2^*}(v_1 - v_2)|, |\gamma P_{\pi_1^*}(v_1 - v_2)|\right\} \in \mathbb{R}^{|\mathcal{S}|}, z_i = \max\{\gamma |p_i^T (v_1 - v_2)|, \gamma |q_i^T (v_1 - v_2)|\}$$

其中 $p_i, q_i$ 分别为 $P_{\pi_1^*}, P_{\pi_2^*}$ 元素，满足行和为 1，$p_i \leqslant 1, q_i \leqslant 1$，

$$\begin{aligned}
&\|f(v_1) - f(v_2)\|_\infty \leqslant \|z\|_\infty \\
&|p_i^T (v_1 - v_2)| \leqslant p_i^T |v_1 - v_2| \leqslant \|v_1 - v_2\|_\infty \Rightarrow z_i \leqslant \gamma \|v_1 - v_2\|_\infty
\end{aligned}$$

从而

$$\|f(v_1) - f(v_2)\|_\infty \leqslant \|z\|_\infty = \max_i |z_i| \leqslant \gamma \|v_1 - v_2\|_\infty$$

□

## C Optimality

**Optimality**

**Theorem 8.** $v^*$ 是最优 *state value*，$\pi^*$是最优的 *policy*，即

$$v^* = v_{\pi^*} \geqslant v_\pi, \forall \pi, \forall v$$

*Proof.* 已知 $v_\pi = r_\pi + \gamma P_\pi v_\pi$，根据定义

$$\begin{aligned} v^* &= \max_\pi (r_\pi + \gamma P_\pi v^*) = r_{\pi^*} + \gamma P_{\pi^*} v^* \geqslant r_\pi + \gamma P_\pi v^* \\ &\Rightarrow v^* - v_\pi \geqslant (r_\pi + \gamma P_\pi v^*) - (r_\pi + \gamma P_\pi v_\pi) = \gamma P_\pi (v^* - v_\pi) \\ &\Rightarrow v^* - v_\pi \geqslant \lim_{n\to\infty} \gamma^n P_\pi^n (v^* - v_\pi) = 0 \end{aligned}$$

最后的不等式是由于 $\gamma < 1$，$P_\pi$ 每个元素 $0 \leqslant p_i \leqslant 1$. □

## D Optimal Policy Invariance

**Optimal Policy Invariance**

$$r' = \alpha r + \beta \Rightarrow v' = \alpha v^* + \frac{\beta}{1-\gamma}\mathbf{1},$$

*Proof.* 根据 matrix form BOE，令 $r_\pi = [\ldots, r_\pi(s_k), \ldots]^T$，其中

$$r_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{r\in\mathcal{R}} p(r|s,a) r, \quad s \in \mathcal{S}$$

由于 $r' = \alpha r + \beta$，那么 $r'_\pi(s) = \alpha r_\pi(s) + \beta \Rightarrow r'_\pi = \alpha r_\pi + \beta\mathbf{1}$，其中 $\mathbf{1} = [1, \ldots, 1]^T$. 这样得到

$$v' = \max_{\pi\in\Pi} (\alpha r_\pi + \beta \mathbf{1} + \gamma P_\pi v') \tag{7}$$

不妨设 $v' = \alpha v^* + c\mathbf{1}$，带入式(7)中就得到

$$\alpha v^* + c\mathbf{1} = \max_{\pi\in\Pi} (\alpha r_\pi + \beta\mathbf{1} + \gamma P_\pi(\alpha v^* + c\mathbf{1})) \xlongequal{P_\pi \mathbf{1} = \mathbf{1}} \max_{\pi\in\Pi} (\alpha r_\pi + \beta\mathbf{1} + \alpha\gamma P_\pi v^* + c\gamma\mathbf{1})$$

进而得到

$$\begin{aligned} \alpha v^* &= \max_{\pi\in\Pi} (\alpha r_\pi + \alpha\gamma P_\pi v^*) + \beta\mathbf{1} + c\gamma\mathbf{1} - c\mathbf{1} \\ &\Rightarrow \beta\mathbf{1} + c\gamma\mathbf{1} - c\mathbf{1} = 0 \Rightarrow c = \frac{\beta}{(1-\gamma)} \end{aligned}$$

因此得到

$$v' = \alpha v^* + \frac{\beta}{1-\gamma}\mathbf{1}$$

□

First updated: January 16, 2025
Last updated: March 21, 2025

## References