Lecture Notes 3

Minfei Chen

2025 年 8 月 14 日

1 最优策略（optimal policy）

1.1 定义

一个策略 π^* 是最优策略，如果 $v_{\pi^*}(s) \geq v_\pi(s)$ 对于所有状态 s 和其他所有策略 π.

2 贝尔曼最优公式（BOE）

2.1 elementwise form

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

* 已知：系统信息（状态转移概率、reword、γ）；求解：最优策略 π

2.2 最优化问题

对于某个状态，最优的策略是选择 action value 最大的那一个 action，即把它的概率设为 1，其他的 action 的概率设为 0。

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

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

2.3 压缩映射定理（contract mapping theorem）

BOE 可以写成：

$$v = f(v)$$

不动点（fixed point）：

$$x \in X \text{ is a fixed point of } f : X \to X \text{ if}$$

$$f(x) = x$$

压缩映射（contract mapping）：

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

$$\text{其中} \gamma \in (0, 1)$$

压缩映射定理：

对于任何形如 $x = f(x)$ 的方程，如果函数 f 是一个**压缩映射 (contraction mapping)**，那么：

- **存在性 (Existence):** 存在一个不动点 x^* 满足 $f(x^*) = x^*$。
- **唯一性 (Uniqueness):** 该不动点 x^* 是唯一的。
- **算法/收敛性 (Algorithm):** 对于一个序列 $\{x_k\}$，如果 $x_{k+1} = f(x_k)$，那么当 $k \to \infty$ 时，序列 x_k 将收敛到 x^* ($x_k \to x^*$)。并且，该收敛是指数级快速的 (exponentially fast)。

2.4 利用压缩映射定理求解 BOE

第一步：证明 $f(v)$ 是一个 contract mapping，公式中的 γ 恰好为 discount rate

第二步：利用压缩映射定理来求出不动点 v^*，即最高的状态值，并且可以知道解是唯一存在的，可以通过迭代的方法逐渐收敛到这个解

2.5 BOE 的最优性

BOE 实际上是一个特殊的贝尔曼公式 $v^* = r_{\pi^*} + \gamma P_{\pi^*} v^*$, 它所对应的策略是最优策略。

从状态值的角度看，假设 v^* 是贝尔曼最优方程 $v = \max_\pi(r_\pi + \gamma P_\pi v)$ 的唯一解，并且对于任何给定的策略 π，v_π 是满足贝尔曼期望方程 $v_\pi = r_\pi + \gamma P_\pi v_\pi$ 的状态价值函数，那么：

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

贝尔曼最优公式所代表的其实是一个**贪心**最优策略，在这个策略下，每个状态都选择 action value 最大的 action。

3 最优策略分析

影响策略的主要有三个因素：

- 奖励设计 (Reward design): r
- 系统模型 (System model): $p(s'|s,a), p(r|s,a)$
- 折扣率 (Discount rate): γ

在系统模型固定的情况下，通过**调整 reward 和 γ**, 最优策略会发生很大的改变。

对于不同的 γ,γ 越大，agent 会更注重长远的 reward。γ 越小，agent 会更注重 immediate reward。

对于不同的 reward，调整 reward 的大小不一定会使最优策略发生改变。影响最优策略的是 **reward 之间的关系（relative reward）**。对奖励函数的线性修改不会改变 action value 的相对大小关系

考虑一个马尔可夫决策过程，其最优状态价值函数为 $v^* \in \mathbb{R}^{|S|}$，满足贝尔曼最优方程 $v^* = \max_\pi(r_\pi + \gamma P_\pi v^*)$。

如果将每一个奖励 r 都进行一次仿射变换 (affine transformation) 得到新的奖励 $ar + b$，其中 $a, b \in \mathbb{R}$ 且 $a > 0$，那么对应的新最优状态价值函数 v' 也是 v^* 的一个仿射变换：

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

其中 $\gamma \in (0,1)$ 是折扣率，$\mathbf{1} = [1, \ldots, 1]^T$ 是全 1 向量。

因此可以推断，**最优策略对于奖励信号的仿射变换是保持不变的**。

*Discount rate 实际上是对绕远路（meaningless detour）的一种惩罚。