# 第三课 bellman optimality equation(BOE)

bellman optimality equation只是在bellman equation前面加了max pai

**Bellman optimality equation (elementwise form):**

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

**Bellman optimality equation (matrix-vector form):**

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

等式右侧，对每个给定的v，都能找到使等式右侧取最大值的pai。因此可以将该等式看作v=f(v)形式。

可以用以下迭代算法求解BOE，得到唯一最优解state value，和与之对应的policy

注：该迭代算法被称为值迭代算法value iteration

**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*

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

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

得到的唯一最优解state value，是所有state value中最大的

**Theorem (Policy Optimality)**

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

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

最优策略是deterministic greedy policy

What does an optimal policy $\pi^*$ look like?

**Theorem (Greedy Optimal Policy)**

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

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

*is an optimal policy solving the BOE. Here,*

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

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

附录：线性改变reward，并不会影响最优策略

$$v' = \max_\pi (r'_\pi + \gamma P_\pi v')$$

**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.*