Policy Optimization

Suppose we aim to find an optimal policy π^*, i.e.,

$$v^{\pi^*}(s) \geq v^{\pi}(s) \qquad \forall s \in S,\ \forall \pi \in \Pi$$

or

$$v^{\pi^*} \geq v^{\pi} \qquad \forall \pi \in \Pi$$

or

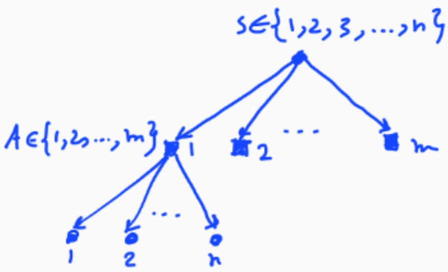
$$v^{\pi^*} = v^* = \max_{\pi \in \Pi} v^{\pi}$$

Bellman Optimality Equations:

The optimal value functions, i.e., v^* and Q^*, satisfy the following equations:

$$v^*(s) = \max_{a \in A} \left[R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[v^*(s') \right] \right] \qquad \forall s \in S$$

$$Q^*(s,a) = R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in A} Q^*(s',a') \right] \qquad \forall s \in S,\ \forall a \in A$$



$v^{\pi}(s) = \mathbb{E}_{a \sim \pi}\left[R(s,a) + \gamma\, \mathbb{E}_{s'}\left[v^{\pi}(s') \right] \right]$ Bellman consistency equation

$v^* = \max_{a \in A} \left[R(s,a) + \gamma\, \mathbb{E}_{s'}\left[v^*(s') \right] \right]$ Bellman optimality equation

Theorem [Bellman Optimality]: π^* is an optimal policy if and only if v^{π^*} satisfies the Bellman optimality equation.

proof (sketch):

① π^* an optimal policy $\Longrightarrow$ v^{π^*} satisfies the Bellman optimality equation

Consider the following deterministic, stationary policy

$$\tilde{\pi}(s) = \arg\max_{a \in \mathcal{A}} \left[R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[v^{\pi^*}(s') \right] \right] \quad \forall s \in \mathcal{S}$$

One can show that $v^{\pi^*} = v^{\tilde{\pi}}$ and $v^{\tilde{\pi}}$ satisfies the Bellman optimality equation.

② v^{π} satisfies the Bellman optimality equation $\Longrightarrow$ π an optimal policy

Consider an optimal policy π^*.

By ①, we know that v^{π^*} satisfies the Bellman optimality equation.

One can show that $|v^{\pi}(s) - v^{\pi^*}(s)| \leq 0$ for all $s \in \mathcal{S}$. ■

Corollary: The deterministic stationary policy

$$\pi(s) = \underset{a \in \mathcal{A}}{\arg\max} \Big[\underbrace{R(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \big[V^*(s') \big]}_{Q^*(s,a)} \Big]$$

or $$\pi(s) = \underset{a \in \mathcal{A}}{\arg\max} \; Q^*(s,a)$$

is an optimal policy.

The optimal value function satisfies the Bellman optimality equation:

$$V^*(s) = \max_{a \in \mathcal{A}} \Big[R(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} \big[V^*(s') \big] \Big], \quad \forall s \in \mathcal{S}$$

Value Iteration

VI aims to find (approximate) the optimal value function v^* through the fixed-point iteration algorithm.

Value Iteration Algorithm

- Initialize $\vec{v}_0$
- For $t = 0, 1, 2, \ldots, T-1$:

$$v_{t+1}(s) = \max_{a \in A} \left[R(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} [v_t(s')] \right] \quad \forall s \in S$$

- Return $\vec{v}_T \approx \vec{v}^*$

or

$$\pi_T(s) = \arg\max_{a \in A} \left[R(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} [v_T(s')] \right] \quad \forall s \in S$$

$\approx$

$v^{\pi_T} \approx v^*$

Bellman optimality operator: $\mathcal{T} : \mathbb{R}^{|S|} \longrightarrow \mathbb{R}^{|S|}$

$$(\mathcal{T} v)(s) = \max_{a \in A} \left[R(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} [v(s')] \right] \quad \forall s \in S$$