

AIL 8027 / AIL 841

Advanced RL

Sem 2, 2025-26

Lec 2: Bellman Equation

Wed, 7 Jan, 2025

Raunak P. Bhattacharyya, ScAI



ScAI

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Bellman Equation

Value Function

Given an MDP $\mathcal{M} = (S, A, P, r, \gamma)$, how to find $\pi^*: S \rightarrow A$

$$\theta^* = \arg \max_{\theta} E_{\tau \sim p_{\theta}(\tau)} \left[\sum_t r(s_t, a_t) \right]$$

Value function $V^{\pi}(s) = E \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t), s_{t+1} \sim P(\cdot \mid s_t, a_t) \right]$

Q function $Q^{\pi}(s, a) = E \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid (s_0, a_0) = (s, a), a_t = \pi(s_t), s_{t+1} \sim P(\cdot \mid s_t, a_t) \right]$

Bellman Equation for Value Function

Value function $V^\pi(s) = E \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t), s_{t+1} \sim P(\cdot \mid s_t, a_t) \right]$

$$V^\pi(s) = r(s, \pi(s)) + \gamma E_{s' \sim P(\cdot \mid s, \pi(s))} [V^\pi(s')]$$

Why is this true?

Q function $Q^\pi(s, a) = E \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid (s_0, a_0) = (s, a), a_t = \pi(s_t), s_{t+1} \sim P(\cdot \mid s_t, a_t) \right]$

What is the Bellman equation for the Q function?

2 Questions

How do we bring Q in the RHS of the Bellman equation for Q?

$$V^\pi(s) = Q^\pi(s, \pi(s))$$

What more is needed if we have stochastic policies? And why do we not worry about stoch policies?

For infinite horizon discounted MDPs, there always exists a deterministic optimal policy

$$\pi^*: S \rightarrow A, s.t., V^{\pi^*}(s) \geq V^\pi(s), \forall s, \pi$$

Preview: Policy Iteration

Policy Iteration

Given an MDP $\mathcal{M} = (S, A, P, r, \gamma)$, how to find $\pi^*: S \rightarrow A$

$$\theta^* = \arg \max_{\theta} E_{\tau \sim p_{\theta}(\tau)} \left[\sum_t r(s_t, a_t) \right]$$

Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

Policy Improvement: $\pi^{t+1}(s) := \arg \max_a Q^{\pi^t}(s, a), \forall s$

Known as Policy Iteration

Policy Evaluation: Fixed Point

Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

Consider $x^* = f(x^*)$, how do we find this?

$$\forall s, V^{\pi}(s) = r(s, \pi(s)) + \gamma E_{s' \sim P(s, \pi(s))} V^{\pi}(s')$$

Approach to find fixed point:

$$V^{\pi} = R + \gamma P V^{\pi}$$

Initialise $x^0 \in [a, b]$, repeat: $x^{t+1} = f(x^t)$

If f is a contraction mapping,

i.e., $\forall x, x', |f(x) - f(x')| \leq \gamma |x - x'|$, for some $\gamma \in [0, 1)$, then:

$$x^t \rightarrow x^*, \text{ as } t \rightarrow \infty$$