Policy Iteration

Recap: Bellman Optimality

$$\mathcal{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto A$

Bellman Optimality—the Q version

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

Recap: Value Iteration

1. VI
(a fix point iteration alg):
$Q^{t+1} \Leftarrow \mathcal{T} Q^t$

Contraction →

2. VI convergence: exponentially fast,
i.e., $\|Q^t - Q^\star\|_\infty \leq \gamma^t \|Q^0 - Q^\star\|_\infty$

1. How to extract a policy from VI?

2. We could set $\pi^t(s) = \arg\max_a Q^t(s, a)$, does $\pi^t \rightarrow \pi^\star$ when t increases?

3. Can we still hope π^t being a good policy?

Recap: Value Iteration

Policy Performance: $V^{\pi^t}(s) \geq V^\star(s) - \frac{2}{1-\gamma}\|Q^t - Q^\star\|_\infty \forall s \in S$

Error in Q is amplified by $1/(1-\gamma)$

(Because π^t could disagree w/ $\pi^\star$ at every step)

Question for Today:

Given an MDP $\mathscr{M} = (S, A, P, r, \gamma)$, How to directly search for $\pi^\star : S \mapsto A$

Outline:

1: An Iterative Algorithm: Policy Iteration

2: Convergence? How fast?

3: A new model: Finite horizon MDP

Algorithm: Policy Iteration

1. Initialization: $\pi^0 : S \mapsto \Delta(A)$

2. For $t = 0 \ldots,$

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

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

Outline:

1: An Iterative Algorithm: Policy Iteration ✓

2: Convergence? How fast?

3: A new model: Finite horizon MDP

Key properties of Policy Iterations:

1. Monotonic improvement:

$$Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s,a$$

2. When $\pi^{t+1} = \pi^t$, then π^t is equal to $\pi^\star$

(You will explore this question in hw1)

Monotonic Improvement

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

Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s, a$

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$= \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) + Q^{\pi^t}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) \right] \geq \ldots, \geq -\gamma^{\infty}/(1-\gamma) = 0$$

$V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s), \forall s, ??$

Summary of Policy Iteration

Iterate between Policy Evaluation and Policy Improvement:

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

Monotonic improvement

$$Q^{\pi^{t+1}}(s, a) \geq Q^{\pi^t}(s, a), \forall s, a$$

Policy Iteration convergence

How many iterations (computation complexity) need to find the EXACT optimal policy?

We will explore this problem in HW1

Outline:

1: An Iterative Algorithm: Policy Iteration

2: Convergence? How fast?

3: A new model: Finite horizon MDP

Finite horizon Markov Decision Process

$$\mathcal{M} = \left\{S, A, r, P, H, \mu_0\right\},$$
$$r : S \times A \mapsto [0,1], H \in \mathbb{N}^+, P : S \times A \mapsto \Delta(S),\ s_0 \sim \mu_0$$

i.e., the task always starts from $s_0 \sim \mu_0$, and lasts for H total steps

Very common in control,
e.g., keep tracking a pre-specified trajectory with fixed length and fixed initial state

Finite horizon Markov Decision Process

$$\mathcal{M} = \left\{S, A, r, P, H, \mu_0\right\},$$
$$r : S \times A \mapsto [0,1], H \in \mathbb{N}^+, P : S \times A \mapsto \Delta(S), s_0 \sim \mu_0$$

Note that in finite horizon setting, we will consider time-dependent policies, i.e.,

$$\pi := \left\{\pi_0, \pi_1, \ldots, \pi_{H-1}\right\}, \pi_h : S \mapsto A, \forall h$$

Policy interacts with the MDP as follows:

$$\tau = \{s_0, a_0, s_1, a_1, \ldots, s_H, a_H\}, s_0 \sim \mu_0, a_0 = \pi_0(s_0), s_1 \sim P(\cdot \mid s_0, a_0), a_1 = \pi_1(s_1), \ldots$$

V/Q functions in Finite horizon MDP

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \,|\, s_h = s, \pi\right]$$

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \,|\, (s_h, a_h) = (s, a), a_\tau \sim \pi \text{ for } \tau > h\right]$$

Compute Optimal Policy via DP

$$\pi^\star = \{\pi_0^\star, \pi_1^\star, \ldots, \pi_{H-1}^\star\}$$

We use Dynamic Programming, and do DP backward in time; start at $H-1$

$$Q_{H-1}^\star(s,a) = r(s,a) \quad \pi_{H-1}^\star(s) = \arg\max_a Q_{H-1}^\star(s,a) \quad V_{H-1}^\star(s) = \max_a Q_{H-1}^\star(s,a)$$

Now assume that we have already computed $V_{h+1}^\star$, $h \leq H-2$

(i.e., we know how to perform optimally starting at $h+1$)

$$Q_h^\star(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} V_{h+1}^\star(s') \quad \pi_h^\star(s) = \arg\max_a Q_h^\star(s,a)$$

Summary so far

1. Basics of MDPs (e.g., Bellman equation, Bellman optimality)

2. How to perform policy evaluation and how to compute optimal policies