Policy Iteration

Recap: Bellman Optimality

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

$$Q^{\star}(s,a) = r(s,a) + q$$

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

Policy $\pi: S \mapsto A$

Bellman Optimality—the Q version $\gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[\max_{\substack{a' \in A}} Q^{\star}(s', a') \right]$

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

1. How to extract a policy from VI?

- $\boldsymbol{\mathcal{A}}$

Recap: Value Iteration

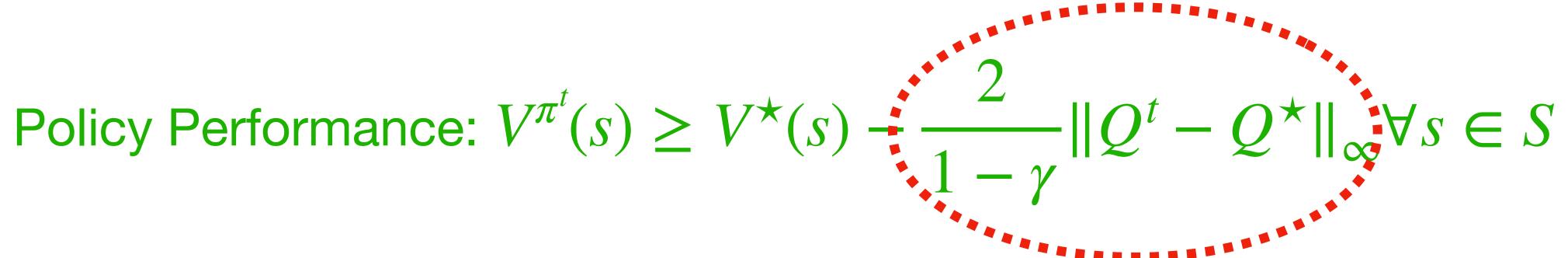
2. VI convergence: exponentially fast, i.e., $\|Q^t - Q^{\star}\|_{\infty} \le \gamma^t \|Q^0 - Q^{\star}\|_{\infty}$

2. We could set $\pi^t(s) = \arg \max Q^t(s, a)$, does $\pi^t \to \pi^*$ when t increases?

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



Recap: Value Iteration



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

(Because π^t could disagree w/ π^* at every step)



Question for Today:

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

2: Convergence? How fast?

3: A new model: Finite horizon MDP

Outline:

1: An Iterative Algorithm: Policy Iteration

1. Initialization: $\pi^0 : S \mapsto \Delta(A)$ 2. For t = 0...

Algorithm: Policy Iteration

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



2: Convergence? How fast?

3: A new model: Finite horizon MDP

Outline:

1: An Iterative Algorithm: Policy Iteration

Key properties of Policy Iterations:

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

1. Monotonic improvement:

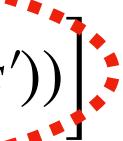
$$\geq Q^{\pi^t}(s,a), \forall s,a$$

- 2. When $\pi^{t+1} = \pi^t$, then π^t is equal to π^{\star}
 - (You will explore this question in hw1)

Monotonic Improvement

$$\begin{aligned} \text{Recall: Policy Improvement } \pi^{t+1}(s) &= \arg\max_{a} Q^{\pi^{t}}(s,a), \forall s \\ \text{Monotonic improvement } Q^{\pi^{t+1}}(s,a) &\geq Q^{\pi^{t}}(s,a), \forall s, a \\ Q^{\pi^{t+1}}(s,a) - Q^{\pi^{t}}(s,a) &= \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^{t}}(s', \pi^{t}(s')) \right] \\ &= \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^{t}}(s', \pi^{t+1}(s')) + Q^{\pi^{t}}(s', \pi^{t+1}(s')) - Q^{\pi^{t}}(s', \pi^{t+1}(s')) \right] \\ &\geq \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^{t}}(s', \pi^{t+1}(s')) \right] \geq \dots, \geq -\gamma^{\infty}/(1-\gamma) = 0 \end{aligned}$$

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



Summary of Policy Iteration

- - $\pi^{t+1}(s) := \arg(s)$

Monotonic improvement

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

Iterate between Policy Evaluation and Policy Improvement:

$$g \max_{a} Q^{\pi^{t}}(s, a), \forall s$$

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

Policy Iteration convergence

We will explore this problem in HW1

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



Outline:

1: An Iterative Algorithm: Policy Iteration

2: Convergence? How fast?

3: A new model: Finite horizon MDP

Finite horizon Markov Decision Process

 $\mathcal{M} = \{S$ $r: S \times A \mapsto [0,1], H \in$

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

$$S, A, r, P, H, \mu_0 \},$$

$$\mathbb{N}^+, P : S \times A \mapsto \Delta(S), s_0 \sim \mu_0$$

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

Finite horizon Markov Decision Process

$$\mathcal{M} = \{S, A, r, P, H, \mu_0\},\$$

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

Note that in finite horizon setting, we will consider time-dependent policies, i.e., $\pi := \{\pi_0, \pi_1, \dots, \pi_{H-1}\}, \pi_h : S \mapsto A, \forall h$

Policy interacts with the MDP as follows:

$$\tau = \{s_0, a_0, s_1, a_1, \dots, s_H, a_H\}, s_0 \sim \mu_0,$$

 $a_0 = \pi_0(s_0), s_1 \sim P(\cdot | s_0, a_0), a_1 = \pi_1(s_1), \dots$

V/Q functions in Finite horizon MDP

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

Compute Optimal Policy via DP

 $\pi^{\star} = \{\pi_0^{\star}\}$

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

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

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

$$\overset{\star}{}, \pi_1^{\star}, \ldots, \pi_{H-1}^{\star} \}$$

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)

Summary so far

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

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