Policy Iteration

Recap: Bellman Optimality

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

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

Policy $\pi : S \mapsto A$

Recap: Bellman Optimality

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

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

Policy $\pi : S \mapsto A$

Bellman Optimality—the Q version

$V^\star(s')$

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

Recap: Value Iteration

1. VI
(a fix point iteration alg):

$$Q^{t+1} \Leftarrow \mathscr{T} Q^t$$

$Q \cdot \xrightarrow{T} \bullet$

$Q' \cdot \xrightarrow{T} \bullet$

Recap: Value Iteration

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

$\xrightarrow{\text{Contraction}}$

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

Recap: Value Iteration

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

$\xrightarrow{\text{Contraction}}$

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

1. How to extract a policy from VI?

$\pi^\star(s) = \arg\max_a Q^\star(s,a)$

$Q^t \approx Q^\star$

$\pi^t(s) = \arg\max_a Q^t(s,a)$

Recap: Value Iteration

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

$\xrightarrow{\text{Contraction}}$

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?

$Q^\star(s, a_1) = 100$

$Q^\star(s, a_2) = 100 + 1e^{-5}$

$t = 1000$

$Q^t(s, a_1) = 100 + 1e^{-4}$

$Q^t(s, a_2) = 100 + 1e^{-5}$

$\|Q^t - Q^\star\|_\infty = 1e^{-4}$

Recap: Value Iteration

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

$\xrightarrow{\text{Contraction}}$

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 $\pi^t \to \pi^\star$ when t increases?

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

$V^{\pi^t}(s) \approx V^{\pi^*}(s)$ if $Q^t \approx Q^*$

Recap: Value Iteration

Recap: Value Iteration

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

Recap: Value Iteration

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

$\to 0$

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

$1 + \gamma + \gamma^2 + \gamma^3 + \dots \gamma^\infty + \dots$

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)

$\sum_{h=0}^{\infty} \gamma^h \|Q^t - Q^\star\|$

$\underline{\underline{\pi^t}} \xrightarrow{\times} \pi^*$

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 $t = 0 \ldots,$

Algorithm: Policy Iteration

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

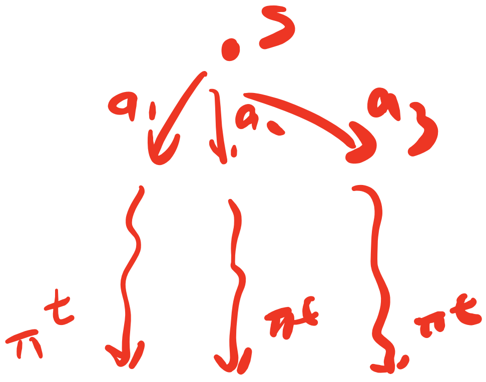
2. For $t = 0 \ldots,$

Policy Evaluation

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

$\arg\max_a Q^{\pi^t}(s, a)$

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

Monotonic Improvement

① − ② = ① − ③ + ③ − ②

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

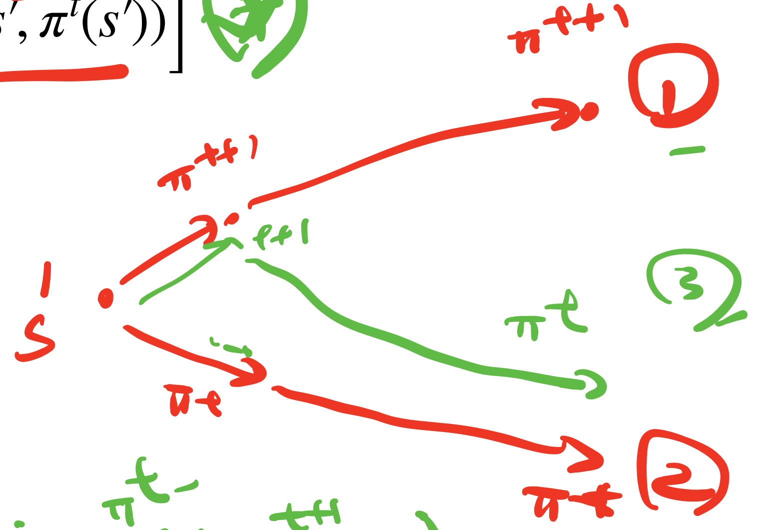
Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s, a$

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \gamma \mathbb{E}_{s' \sim P(s,a)} \Big[\underbrace{Q^{\pi^{t+1}}(s', \pi^{t+1}(s'))}_{①} - \underbrace{Q^{\pi^t}(s', \pi^t(s'))}_{②} \Big] \quad (*)$$

$$Q^{\pi^{t+1}}(s,a) = \cancel{r(s,a)} + \gamma \mathbb{E}_{s'} Q^{\pi^{t+1}}(s', \pi^{t+1}(s'))$$

BE

$$Q^{\pi^{t}}(s,a) = \cancel{r(s,a)} + \gamma \mathbb{E}_{s'} Q^{\pi^{t}}(s', \pi^{t}(s'))$$



③ : $Q^{\pi^t}(s, \pi^{t+1}(s))$

Monotonic Improvement

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

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

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

$$= \gamma \mathbb{E}_{s' \sim P(s,a)} \left[\underbrace{Q^{\pi^{t+1}}(s', \pi^{t+1}(s'))}_{1} - \underbrace{Q^{\pi^t}(s', \pi^{t+1}(s'))}_{3} + \underbrace{Q^{\pi^t}(s', \pi^{t+1}(s'))}_{3} - \underbrace{Q^{\pi^t}(s', \pi^t(s'))}_{2} \right]$$

Monotonic Improvement

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

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

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

?? >0

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

3 − 2 ≥ 0

Monotonic Improvement

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

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

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

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

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

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

π^{t+1} (1)

(1) − (3)

γ^{∞}

Monotonic Improvement

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

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

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

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

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

$\in \left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma} \right]$

∞ Δ

Monotonic Improvement

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

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

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

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

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

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

$\geq Q^{\pi^t}(s, \pi^t(s)) = V^{\pi^t}(s)$

Summary of Policy Iteration

Iterate between Policy Evaluation and Policy Improvement:

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

PE

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

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

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

Policy Iteration convergence π^*

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

$\gamma \in [0, 1)$

$V^{\pi} = \left[\sum_{h=0}^{\infty} \gamma^h \cdot r_h \right]$

Finite horizon Markov Decision Process

H Integer

horizon

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

Initial state distribution

Finite horizon Markov Decision Process

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

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

Finite horizon Markov Decision Process

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

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

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

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

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:

- - - H

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

V/Q functions in Finite horizon MDP

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

V/Q functions in Finite horizon MDP

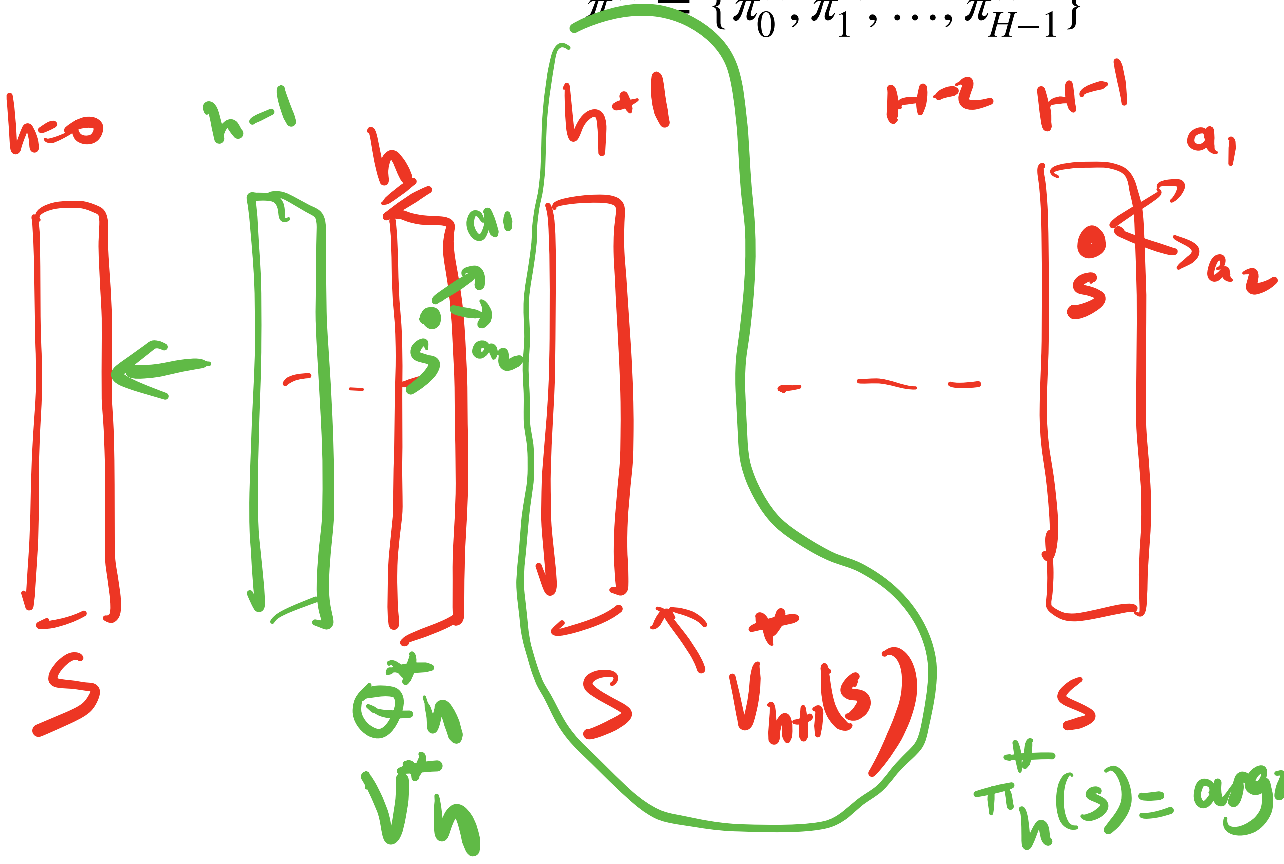
$$V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \,|\, s_h = s, \pi\right]$$

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

$$Q_h^\pi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V_{h+1}^\pi(s')$$

Compute Optimal Policy via DP

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



$$\pi^\star_{H-1}(s) = \arg\max_a r(s,a)$$

$$Q^\star_{H-1}(s,a) = r(s,a)$$

$$V^\star_{H-1}(s) = \max_a r(s,a) = \max_a Q^\star_{H-1}(s,a)$$

$$\pi^\star_h(s) = \arg\max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V^\star_{h+1}(s') \right]$$

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 $H - 1$

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 $H - 1$

$$Q_{H-1}^\star(s, a) = r(s, a)$$

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 $H-1$

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

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 $H - 1$

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

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 $h+1$)

$\pi^\star_h.\quad Q^\star_h.\quad V^\star_h$

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 $h + 1$)

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

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 $h+1$)

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

$$V_h^\star(s) = \max_a Q_h^\star(s,a)$$

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