

Policy Iteration

Recap: Bellman Optimality

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

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

$$\text{Policy } \pi : S \mapsto A$$

Recap: Bellman Optimality

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

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

$$\text{Policy } \pi : S \mapsto A$$

Bellman Optimality—the Q version

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

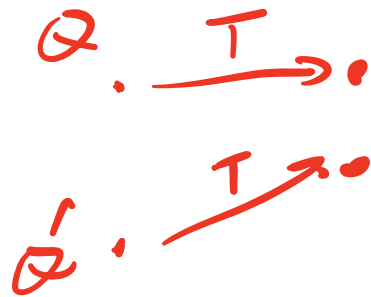
Handwritten annotations: A red triangle is under the Q^* term. A red circle is around the maximization term. A red arrow points from the circle to the handwritten $V^*(s')$ above it.

Recap: Value Iteration

1. VI

(a fix point iteration alg):

$$Q^{t+1} \leftarrow \mathcal{T} Q^t$$



Recap: Value Iteration

1. VI

(a fix point iteration alg):

$$Q^{t+1} \leftarrow \mathcal{T} Q^t$$

Contraction



2. VI convergence: exponentially fast,

$$\text{i.e., } \underline{\|Q^t - Q^*\|_\infty} \leq \underbrace{\gamma^t}_{\triangle} \|Q^0 - Q^*\|_\infty$$

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., $\|Q^t - Q^*\|_\infty \leq \gamma^t \|Q^0 - Q^*\|_\infty$

1. How to extract a policy from VI?

$$\pi^*(s) = \operatorname{argmax}_a Q^*(s, a)$$

$$Q^t \approx Q^*$$

$$\pi^t(s) = \operatorname{argmax}_a Q^t(s, a)$$

Recap: Value Iteration

1. VI

(a fix point iteration alg):

$$Q^{t+1} \leftarrow \mathcal{T} Q^t$$

Contraction



2. VI convergence: exponentially fast,

$$\text{i.e., } \|Q^t - Q^*\|_\infty \leq \gamma^t \|Q^0 - Q^*\|_\infty$$

1. How to extract a policy from VI?

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

$$Q^*(s, a_1) = 100$$

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

$t = 1000$

$$Q^t(s, a_1) = 100 + 1e^{-4}$$
$$Q^t(s, a_2) = 100 + 1e^{-5}$$
$$\|Q^t - Q^*\|_\infty = 1e^{-4}$$

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., $\|Q^t - Q^*\|_\infty \leq \gamma^t \|Q^0 - Q^*\|_\infty$

1. How to extract a policy from VI?

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

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

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

Recap: Value Iteration

Recap: Value Iteration

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

Recap: Value Iteration

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

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

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

Recap: Value Iteration

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

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

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

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

π^t $\xrightarrow{\lambda}$ π^*

Question for Today:

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

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 $t = 0 \dots$,

|

Algorithm: Policy Iteration

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

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

$$\underset{a}{\operatorname{argmax}} Q^{\pi^t}(s, a)$$



Policy Evaluation

Algorithm: Policy Iteration

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

2. For $t = 0 \dots$,

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 π^\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

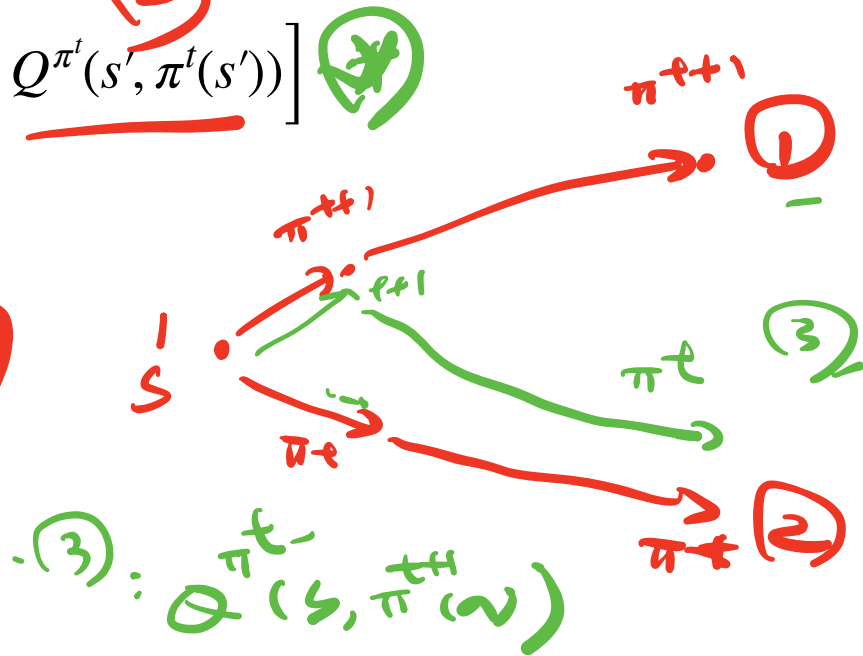
Recall: Policy Improvement $\pi^{t+1}(s) = \arg \max_a Q^{\pi^t}(s, a), \forall s$

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

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

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

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



Monotonic Improvement

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

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

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

Monotonic Improvement

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

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

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

Handwritten annotations in red:
- A circle around $\pi^{t+1}(s)$ in the first line.
- A circle around $??$ with > 0 next to it, pointing to the first term in the second line.
- A dashed red oval around the second term in the second line, with a red line striking through it.
- The expression $(3) - (2) \geq 0$ written below the oval.

Monotonic Improvement

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

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

$$\begin{aligned}
 Q^{\pi^{t+1}}(s, a) - Q^{\pi^t}(s, a) &= \gamma \mathbb{E}_{s' \sim P(s, a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right] \\
 &= \gamma \mathbb{E}_{s' \sim P(s, a)} \left[\underbrace{Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s'))}_{\text{green box}} + \underbrace{Q^{\pi^t}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s'))}_{\text{red dashed box}} \right] \\
 &\geq \gamma \mathbb{E}_{s' \sim P(s, a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) \right] \\
 &\geq \gamma \mathbb{E}_{s'' \sim P(s, \pi^{t+1}(s))} \left[Q^{\pi^{t+1}}(s'', \pi^{t+1}(s'')) - Q^{\pi^t}(s'', \pi^{t+1}(s'')) \right]
 \end{aligned}$$

Handwritten annotations include:

- Green triangles under the first two terms of the first line.
- A red circle around the inequality sign in the third line, with a green triangle below it and the symbol δ written below.
- Red annotations π^{t+1} and π^t with arrows pointing to the corresponding policies in the last line.
- A green arrow labeled π^{t+1} pointing from the first term in the last line to the second term in the last line.
- Handwritten text $\delta \mathbb{E}_{s'' \sim P(s, \pi^{t+1}(s))}$ in red and green.
- A green box around the last line of the equation.
- A red arrow labeled π^{t+1} pointing from the first term in the last line to the second term in the last line.
- A red arrow labeled π^t pointing from the second term in the last line to the second term in the last line.
- Handwritten (1) and $(2)-(3)$ in red.

Monotonic Improvement

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

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

$$\begin{aligned} Q^{\pi^{t+1}}(s, a) - Q^{\pi^t}(s, a) &= \gamma \mathbb{E}_{s' \sim P(s, a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right] \\ &= \gamma \mathbb{E}_{s' \sim P(s, a)} \left[\underbrace{Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s'))}_{\in [-\frac{\gamma}{1-\gamma}, \frac{\gamma}{1-\gamma}]} + \underbrace{Q^{\pi^t}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s'))}_{\geq 0} \right] \\ &\geq \gamma \mathbb{E}_{s' \sim P(s, a)} \left[\underbrace{Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s'))}_{\geq -\frac{\gamma}{1-\gamma}} \right] \geq \dots \geq \underbrace{-\gamma^\infty / (1 - \gamma)}_{= 0} \end{aligned}$$

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

Monotonic Improvement

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

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

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

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

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

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

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

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

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?



$$\gamma \in [0, 1)$$
$$V^\pi = \left[\sum_{h=0}^{\infty} \gamma^h \cdot r_h \right]$$

3: A new model: Finite horizon MDP

Finite horizon Markov Decision Process

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

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

H Integer horizon

Initial state distribution

Finite horizon Markov Decision Process

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

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

Finite horizon Markov Decision Process

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

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

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

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.,

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

▶ ▶ ▶

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:

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

--- H

V/Q functions in Finite horizon MDP

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

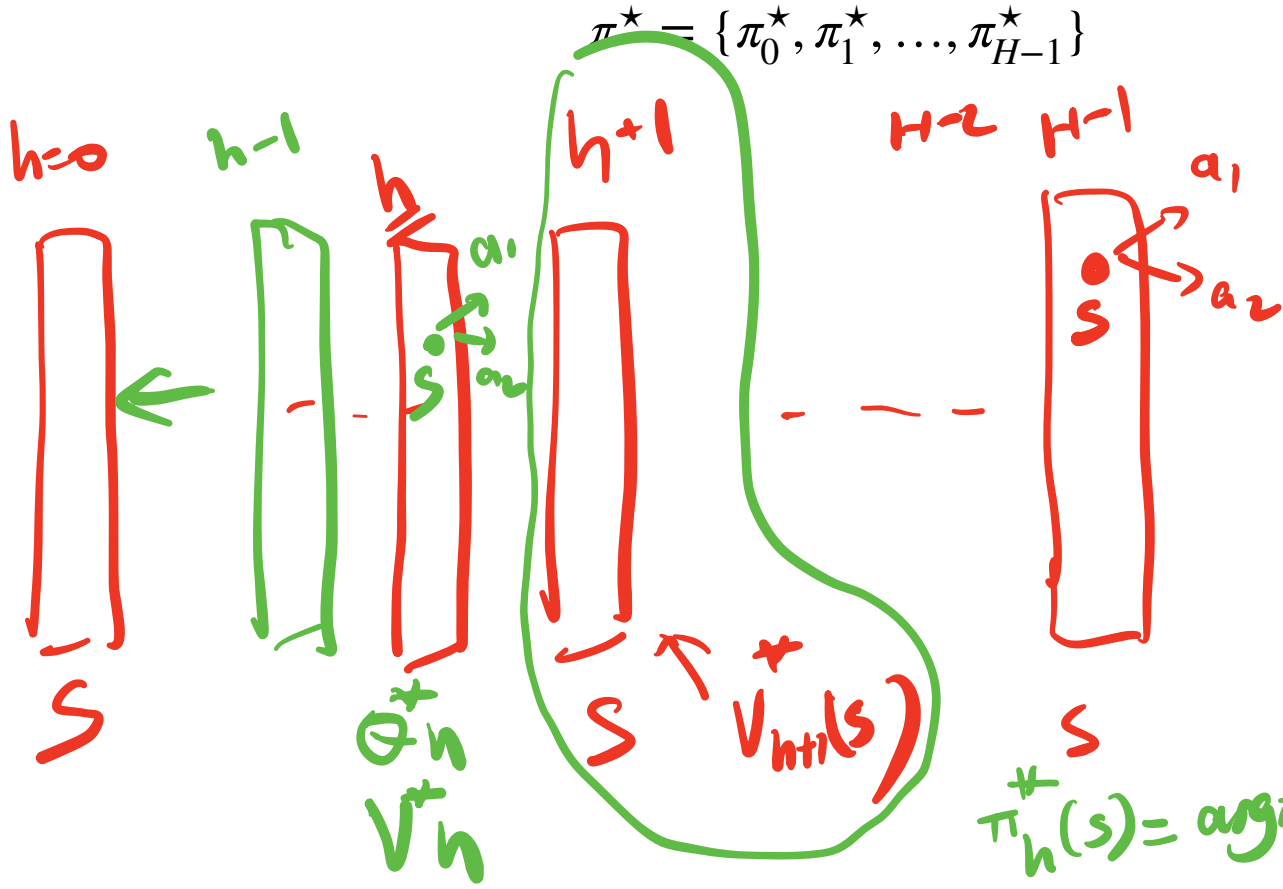
V/Q functions in Finite horizon MDP

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

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

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

Compute Optimal Policy via DP



$$\pi_{H-1}^*(s) = \operatorname{argmax}_a \Gamma(s, a)$$

$$Q_{H-1}^*(s, a) = \Gamma(s, a)$$

$$V_{H-1}^*(s)$$

$$= \max_a \Gamma(s, a)$$

$$= \max_a Q_{H-1}^*(s, a)$$

$$\pi_h^*(s) = \operatorname{argmax}_a \left[\Gamma(s, a) + \mathbb{E}_{s'} [V_{h+1}^*(s')] \right]$$

Compute Optimal Policy via DP

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

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

Compute Optimal Policy via DP

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

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

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

Compute Optimal Policy via DP

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

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

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

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

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

Compute Optimal Policy via DP

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

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

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

Now assume that we have already computed V_{h+1}^* , $h \leq H - 2$
(i.e., we know how to perform optimally starting at $h + 1$)

$$\pi_h^* \quad Q_h^* \quad V_h^*$$

Compute Optimal Policy via DP

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

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

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

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

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

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

Compute Optimal Policy via DP

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

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

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

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

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

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

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

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