Policy Iteration

Recap: Policy Evaluation

Given a policy $\pi : S \mapsto A$, compute V^{π}:

Recap: Policy Evaluation

Given a policy $\pi : S \mapsto A$, compute V^{π}:

1. Solve a linear system:

$$\forall s : V(s) = r(s, \pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V(s')$$

Recap: Policy Evaluation

Given a policy $\pi : S \mapsto A$, compute V^{π}:

1. Solve a linear system:

$$\forall s : V(s) = r(s, \pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V(s')$$

2. Fix-point iteration

$$\forall s : V^{t+1}(s) \Leftarrow r(s, \pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V^{t}(s')$$

Recap: Policy Evaluation

Given a policy $\pi : S \mapsto A$, compute V^π:

1. Solve a linear system:

$$\forall s : V(s) = r(s, \pi(s)) + \gamma \sum_{s'} P(s' | s, \pi(s)) V(s')$$

2. Fix-point iteration

$$\forall s : V^{t+1}(s) \Leftarrow r(s, \pi(s)) + \gamma \sum_{s'} P(s' | s, \pi(s)) V^t(s')$$

Q: once we get V^π, how to get Q^π?

$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} V^\pi(s')$$

Recap: Value Iteration

1. VI
(a fix point iteration alg):

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

$$Q^* = \mathcal{T} Q^*$$

$\forall s, a$

$$Q^{t+1}(s,a) \Leftarrow r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in A} \left(Q^t(s', a') \right)$$

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

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

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

Recap: Value Iteration

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

Contraction →

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

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

3. Policy Performance: $V^{\pi^t}(s) \ge V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|Q^0 - Q^\star\|_\infty \forall s \in S$

$Q^0 \in [0, \frac{1}{1-\gamma}]$

$Q^\star \in [0, \frac{1}{1-\gamma}]$

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

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

3. Policy Performance: $V^{\pi^t}(s) \geq V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|Q^0 - Q^\star\|_\infty \forall s \in S$

$Q^\star$

Note that $Q^t \in \mathbb{R}^{|S||A|}$ is our estimator from VI,
it does not correspond a Q^{π^t}!

$Q^{\pi^t} \neq Q^t$

Question for Today:

~~(approximately)~~

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

Outline:

1: An Iterative Algorithm: Policy Iteration

2: Convergence? How fast?

3: A new model: Finite horizon MDP

Algorithm: Policy Iteration

$\{\pi^0, \pi^1, \dots \pi^T\}$

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

2. For $t = 0 \ldots,$

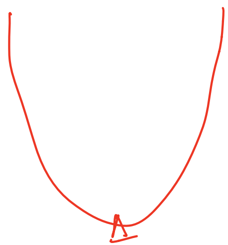
Algorithm: Policy Iteration

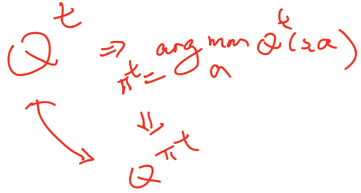
1. Initialization: $\pi^0 : S \mapsto \Delta(A)$

2. For $t = 0\ldots,$

linear program to compute $V^{\pi^t} \Rightarrow Q^{\pi^t}$

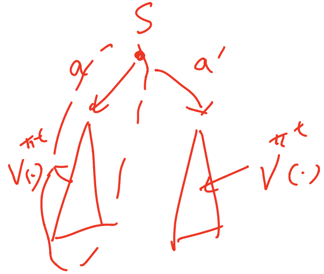
3. **Policy Evaluation**: $Q^{\pi^t}(s, a), \forall s, a$





Algorithm: Policy Iteration

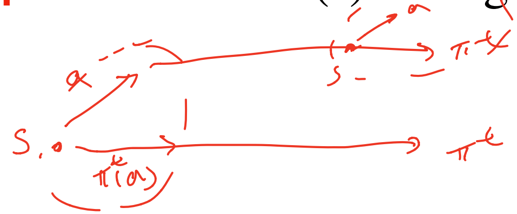
1. Initialization: $\pi^0 : S \mapsto \Delta(A)$



2. For $t = 0 \ldots,$

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

    4. **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? γ^t

3: A new model: Finite horizon MDP

Key properties of Policy Iterations:

1. Monotonic improvement:

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

Key properties of Policy Iterations:

1. Monotonic improvement:

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

2. Convergence:

$$\| V^\star - V^{\pi^t} \|_\infty \leq \gamma^t \| V^\star - V^{\pi^0} \|_\infty$$

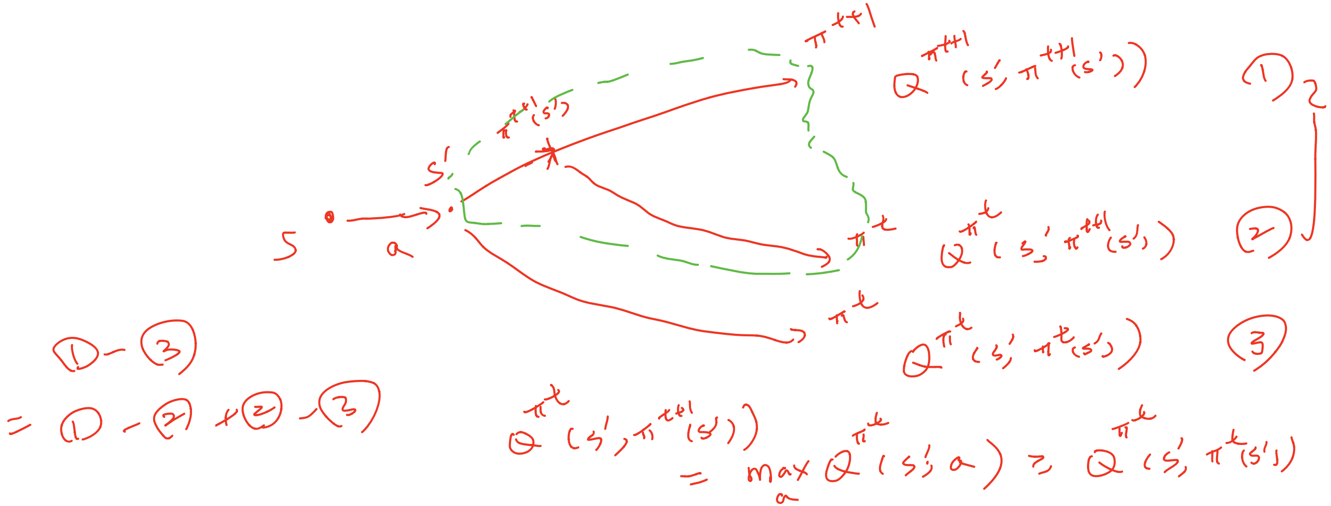
Monotonic Improvement

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

Monotonic Improvement

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

Let us compare $Q^{\pi^{t+1}}(s, a)$ & $Q^{\pi^t}(s, a)$:



Monotonic Improvement

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

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

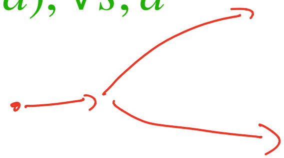
Monotonic Improvement

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

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

cancell r(s,a)

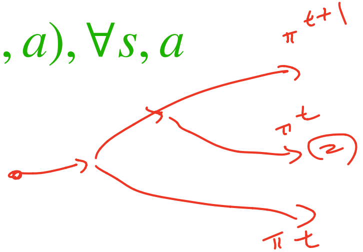
$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$



Monotonic Improvement

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

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



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

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

(2) apply Recursion

(2) ≥ 0

Monotonic Improvement

$\gamma \in [0, 1]$

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

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

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

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

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

Recursion

$-\frac{1}{1-\gamma} \leq \ldots \leq \frac{1}{1-\gamma}$

$Q^{\pi^t}(s', \pi^{t+1}(s')) = \max_a Q^{\pi^t}(s', a) \geq Q^{\pi^t}(s', \pi^t(s'))$

Monotonic Improvement

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

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

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

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

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

Monotonic Improvement

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

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

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

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

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

$\geq -\frac{1}{1-\gamma}$

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

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

Lemma

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

Convergence analysis via Monotonic Improvement

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

Theorem: Convergence $\|V^{\pi^{t+1}} - V^{\star}\|_\infty \le \gamma \|V^{\pi^t} - V^{\star}\|_\infty$ $\le \cdots \gamma^{t+1} \|V^{\pi^0} - V^{\star}\|_\infty$

Convergence analysis via Monotonic Improvement

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

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

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

Convergence analysis via Monotonic Improvement

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

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \le \gamma \|V^{\pi^t} - V^\star\|_\infty$

monotonic improvement on $V^{\pi^{t+1}}$

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

$$\le \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^{t+1}(s))} V^{\pi^t}(s') \right]$$

Convergence analysis via Monotonic Improvement

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

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \le \gamma \|V^{\pi^t} - V^\star\|_\infty$

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

$$\le \max_a \left[r(s,a) + \gamma \mathbb{E}_{s'\sim P(s,a)} V^\star(s') \right] - \left[r(s,\pi^{t+1}(s)) + \gamma \mathbb{E}_{s'\sim P(s,\pi^{t+1}(s))} V^{\pi^t}(s') \right]$$

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

$$= \max_a \left(r(s,a) + \mathbb{E}_{s'\sim P(s,a)} \gamma V^\star(s') \right) - \max_a \left(r(s,a) + \gamma \mathbb{E}_{s'\sim P(s,a)} V^{\pi^t}(s') \right)$$

Convergence analysis via Monotonic Improvement

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

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \le \gamma \|V^{\pi^t} - V^\star\|_\infty$

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

$$\le \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^{t+1}(s))} V^{\pi^t}(s') \right]$$

$$= \max_a \left(r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \gamma V^\star(s') \right) - \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s') \right)$$

$$\le \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s') \right) \right)$$

Convergence analysis via Monotonic Improvement

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

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \le \gamma \|V^{\pi^t} - V^\star\|_\infty$

monotonic improvement

$\forall s$

$s' \sim P(\cdot | s, a)$

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

$$\le \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s') \right]$$

$$= \max_a \left(r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \gamma V^\star(s') \right) - \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s') \right)$$

$$\le \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s') \right) \right)$$

$$\le \gamma \|V^\star - V^{\pi^t}\|_\infty$$

$\|V^\star - V^{\pi^{t+1}}\|_\infty \le \gamma \|V^\star - V^{\pi^t}\|_\infty$

Summary of Policy Iteration

Iterate between Policy Evaluation and Policy Improvement:

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

Exact Q^{π^t}

Summary of Policy Iteration

Iterate between Policy Evaluation and Policy Improvement:

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

← policy evaluation is expensive

Monotonic improvement + convergence:

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

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

$$\| V^\star - V^{\pi^t} \|_\infty \leq \gamma^t \| V^\star - V^{\pi^0} \|_\infty$$

Value Iteration vs Policy Iteration

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

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

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

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

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

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

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

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

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

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

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

no γ any more

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

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

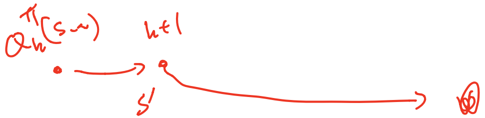
V/Q functions in Finite horizon MDP

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

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

Bellman Equation:

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



$$Q^\pi(s,a) = r(s,a) + E\left[V_{h+1}^\pi(s')\right]$$

Compute Optimal Policy via DP

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

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 $H-1$

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

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 $H-1$

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

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

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 $H-1$

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

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

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 $h+1$)

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 $H-1$

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

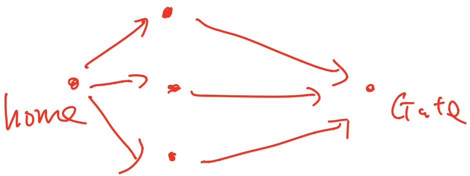
$$V_{H-1}^\star = \max_a Q_{H-1}^\star(s,a) = Q_{H-1}^\star(s, \pi_{H-1}^\star(s))$$

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 $H-1$

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

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

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

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

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

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

Summary on Finite horizon MDP

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

Comparing to the infinite horizon MDP:

1. Policy will be time dependent
2. Value Iteration takes H steps (DP from $H - 1 \to 0$)
3. Compute exact $\pi^\star$—-no need to use γ^t argument
4. No more discount factor