Planning in MDPs

Sham Kakade and Kianté Brantley

CS 2824: Foundations of Reinforcement Learning

Announcements

HW0 is **due** Mon Feb. 2nd

First reading assignment **due** Wed. Feb 4th

Waitlist

Recap: Infinite Horizon MDPs

state space

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

action space

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

Recap: Infinite Horizon MDPs

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

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

Stationary Policy $\pi : S \mapsto \cancel{\Delta}(A)$

$\pi^* : S \to A$

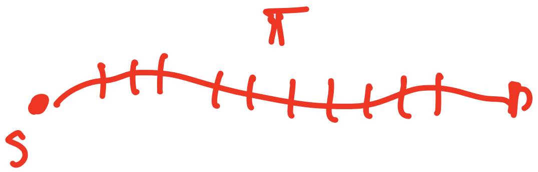
Recap: Infinite Horizon MDPs

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

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

Stationary Policy $\pi : S \mapsto \Delta(A)$

Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$



Recap: Infinite Horizon MDPs

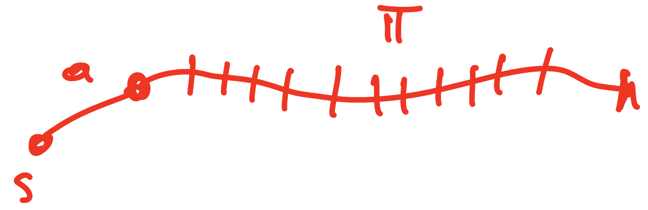
$$\mathcal{M} = \{S, A, P, r, \gamma\}$$

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

Stationary Policy $\pi : S \mapsto \Delta(A)$

Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

Q function $Q^\pi(s, a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s, a), a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$



Recap: Bellman Optimality

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

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

Theorem 1: Bellman Optimality (Q-version)

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

Main Question for Today:

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

Outline

1. Bellman optimality — property of $V^\star$

2. Optimal planning: Value Iteration

Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

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

Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$|V(s) - V^\star(s)| = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

Conditie

Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$|V(s) - V^\star(s)| = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|$$

Bellman Optimality

Theorem 2:

For any $V: S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$\underbrace{|V(s) - V^\star(s)|}_{\text{State } s} = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right| \quad ①$$

$$\le \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right| \quad ②$$

$$\le \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| \overset{\text{State } s'}{\underline{V(s') - V^\star(s')}} \right| \quad ③$$

$$\left| \mathbb{E}_{x \sim p} f(x) - \mathbb{E}_{x \sim p} g(x) \right| \le \mathbb{E}_{x \sim p} \left| f(x) - g(x) \right|$$

Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$
\begin{aligned}
\underbrace{|V(s) - V^\star(s)|}_{} &= \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right| \\
&\le \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right| \\
&\le \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| \underline{V(s') - V^\star(s')} \right| \\
&\le \underbrace{\max_a \gamma \mathbb{E}_{s' \sim P(s,a)}}_{\text{state } s} \left(\underbrace{\max_{a'} \gamma \mathbb{E}_{s'' \sim P(s',a')}}_{\text{state } s'} \left| \underline{V(s'') - V^\star(s'')} \right| \right)
\end{aligned}
$$

Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$|V(s) - V^\star(s)| = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$V(s) = V^\star(s)$

$\forall s$

$$\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| V(s') - V^\star(s') \right|$$

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left(\max_{a'} \gamma \mathbb{E}_{s'' \sim P(s',a')} \left| V(s'') - V^\star(s'') \right| \right)$$

$$\leq \underbrace{\max_{a_1, a_2, \ldots a_{k-1}} \gamma^k \mathbb{E}_{s_k} |V(s_k) - V^\star(s_k)|}_{\geq 0}$$

$k \to \infty$

$\gamma^k \to 0$

Bellman Optimality for $Q^\star$

What about $Q^\star$?

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

Bellman Optimality for $Q^\star$

What about $Q^\star$?

We should have:

For any $Q : S \times A \to \mathbb{R}$, if $\boxed{Q(s,a)} = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} \boxed{Q(s',a')}$ for all s, then $Q(s,a) = Q^\star(s,a), \forall s, a$

Outline

1. Bellman optimality — property of $V^\star$

2. Optimal planning: Value Iteration

Define Bellman Operator $\mathcal{T}$:

$f : S \times A \to \mathbb{R}$

Given a function $f : S \times A \mapsto \mathbb{R}$,

$\mathcal{T}f : S \times A \to \mathbb{R}$

$$\mathcal{T}f : S \times A \mapsto \mathbb{R},$$

$$\boxed{(\mathcal{T}f)}(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in A} f(s', a'), \forall s, a \in S \times A$$

Define Bellman Operator $\mathcal{T}$:

Given a function $f : S \times A \mapsto \mathbb{R}$,

$$\mathcal{T}f : S \times A \mapsto \mathbb{R},$$

$$(\mathcal{T}f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in A} f(s',a'), \forall s,a \in S \times A$$

$$(\mathcal{T}Q^*)(s,a) = r(s,a) + \gamma \mathbb{E}\left[\max_{a' \in A} Q^*(s',a')\right] = Q^*(s,a)$$

Q: what is $\mathcal{T}Q^\star$?

Value Iteration Algorithm:

$\gamma \in [0,1)$

$Q: S \times A \to \mathbb{R}$

$\sum \gamma^t = \frac{1}{1-\gamma}$

1. Initialization: $Q^0 : \|Q^0\|_\infty \in (0, \frac{1}{1-\gamma})$

2. Iterate until convergence: $Q^{t+1} = \mathcal{T} Q^t$

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

Intuition:

Via Bellman optimality theorem:

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

Intuition:

Via Bellman optimality theorem:

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

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Intuition:

Via Bellman optimality theorem:

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

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

$\ell : \mathbb{R} \to \mathbb{R}$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

Intuition:

Via Bellman optimality theorem:

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

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, \boxed{x_{t+1}} = \ell(x_t), t = 0, \ldots,$$

$t = 0, 1, \ldots \infty$

$t \to \infty$

$x_t = x^*$

Intuition:

Via Bellman optimality theorem:

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

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

$$|x_t - x^\star| =$$

Intuition:

Via Bellman optimality theorem:

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

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$= \ell(x_{t-1})$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

$$|x_t - x^\star| = |\ell(x_{t-1}) - \ell(x^\star)|$$

$= \ell(x^\star)$

Intuition:

Via Bellman optimality theorem:

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

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

$$|x_t - x^\star| = |\ell(x_{t-1}) - \ell(x^\star)| \leq L |x_{t-1} - x^\star|$$

Intuition:

Via Bellman optimality theorem:

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

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

$$|x_t - x^\star| = |\ell(x_{t-1}) - \ell(x^\star)| \leq L|x_{t-1} - x^\star| \leq L^2 |x_{t-2} - x^\star|$$

......

If $L < 1$ (i.e., contraction), then it converges exponentially fast

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

$\|Q(s,a)\|_\infty$

$= \max_{s,a} \|Q(s,a)\|_\infty$

Proof:

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$$|\underline{\mathcal{T}Q(s,a)} - \underline{\mathcal{T}Q'(s,a)}| = \left| \underline{r(\cancel{s,a}) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a')} - \left(r(\cancel{s,a}) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right|$$

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$$|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \boxed{\gamma \sum_{s'} P(s' | s,a)} \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$$|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s'\sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s'\sim P(s,a)} \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \gamma \sum_{s'} P(s'|s,a) \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \gamma \boxed{\sum_{s'} P(s'|s,a)} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

$\mathbb{E}_{s'\sim P} f(s) \leq \max_s f(s)$

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$$|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \gamma \sum_{s'} P(s'|s,a) \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \gamma \sum_{s'} P(s'|s,a) \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

$$\leq \gamma \max_{s'} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \le \gamma \|Q - Q'\|_\infty$$

Proof:

$$|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right|$$

$$\le \gamma \sum_{s'} P(s' | s,a) \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

$$\le \gamma \sum_{s'} P(s' | s,a) \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

$$\le \gamma \max_{s'} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right| \quad = \gamma \|Q - Q'\|_\infty$$

$\gamma \in [0,1)$

Convergence of Value Iteration:

Lemma [Convergence]: Given Q^0, we have:

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

Proof:

Convergence of Value Iteration:

Lemma [Convergence]: Given Q^0, we have:

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

Proof:

$= \tau Q^t$ $= \tau Q^*$

$$\|Q^{t+1} - Q^\star\|_\infty = \|\mathcal{T}Q^t - \mathcal{T}Q^\star\|_\infty \leq \gamma \|Q^t - Q^\star\|_\infty$$

Convergence of Value Iteration:

Lemma [Convergence]: Given Q^0, we have:

$$\|Q^t - Q^\star\|_\infty \le \gamma^t \|Q^0 - Q^\star\|_\infty$$

Proof:

$$\|Q^{t+1} - Q^\star\|_\infty = \|\mathcal{T} Q^t - \mathcal{T} Q^\star\|_\infty \le \underline{\gamma \|Q^t - Q^\star\|_\infty}$$

$$\dots \le \underline{\gamma^{t+1} \|Q^0 - Q^\star\|_\infty}$$

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

Final Quality of the Policy:

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

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

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

Theorem: $V^{\pi^t}(s) \geq V^\star(s) - \dfrac{2\gamma^t}{1-\gamma} \|Q^0 - Q^\star\|_\infty \forall s \in S$

Proof:

Final Quality of the Policy:

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

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

Proof:

Bell Eq.

$$\underline{V^{\pi^t}(s)} - \underline{V^\star(s)} = Q^{\pi^t}(s, \underline{\pi^t(s)}) - Q^\star(s, \pi^\star(s))$$

Final Quality of the Policy:

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

Theorem: $V^{\pi^t}(s) \geq V^\star(s) - \dfrac{2\gamma^t}{1-\gamma}\|Q^0 - Q^\star\|_\infty \forall s \in S$

Proof:

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

+/-

$$= Q^{\pi^t}(s,\pi^t(s)) - \boxed{Q^\star(s,\pi^t(s)) + Q^\star(s,\pi^t(s))} - Q^\star(s,\pi^\star(s))$$

Final Quality of the Policy:

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

Theorem: $V^{\pi^t}(s) \geq V^\star(s) - \dfrac{2\gamma^t}{1-\gamma}\|Q^0 - Q^\star\|_\infty \forall s \in S$

Proof:

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

state s

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^\star(s, \pi^t(s)) + Q^\star(s, \pi^t(s)) - Q^\star(s, \pi^\star(s))$$

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

state s'

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

Final Quality of the Policy:

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

$\geq V^* - \dfrac{\|Q^t - Q^*\|_\infty}{1-\gamma}$

Theorem: $V^{\pi^t}(s) \geq V^\star(s) - \dfrac{2\gamma^t}{1-\gamma}\|Q^0 - Q^\star\|_\infty \forall s \in S$

Proof:

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

$Q^t(s, \pi^t(s)) \geq Q^t(s, \pi^\star(s))$

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

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^\star(s, \pi^t(s)) + Q^\star(s, \pi^t(s)) - Q^\star(s, \pi^\star(s))$$

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

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

Final Quality of the Policy:

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

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

Proof:

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

$$= Q^{\pi^t}(s, \pi^t(s)) - Q^\star(s, \pi^t(s)) + Q^\star(s, \pi^t(s)) - Q^\star(s, \pi^\star(s))$$

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

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

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^t(s))} \left(V^{\pi^t}(s') - V^\star(s') \right) - 2\gamma^t \|Q^0 - Q^\star\|_\infty$$

Final Quality of the Policy:

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

Theorem: $V^{\pi^t}(s) \geq V^\star(s) - \dfrac{2\gamma^t}{1-\gamma}\|Q^0 - Q^\star\|_\infty \forall s \in S$

Proof:

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

$$= Q^{\pi^t}(s,\pi^t(s)) - Q^\star(s,\pi^t(s)) + Q^\star(s,\pi^t(s)) - Q^\star(s,\pi^\star(s))$$

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

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

$$\geq \gamma\mathbb{E}_{s'\sim P(s,\pi^t(s))}\left(V^{\pi^t}(s') - V^\star(s')\right) - 2\gamma^t\|Q^0 - Q^\star\|_\infty$$ ...Recursion

Outline

1. Bellman optimality — property of $V^\star$

2. Optimal planning: Value Iteration

3. State-action distribution

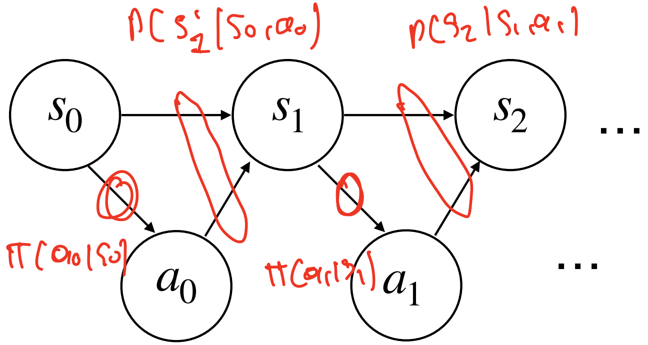
Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{\underline{s_0}, \underline{a_0}, s_1, a_1, \ldots, \underline{s_h}, \underline{a_h}\}$?

$$\pi = \begin{cases} 1 & a = \pi(s) \\ 0 & \end{cases}$$

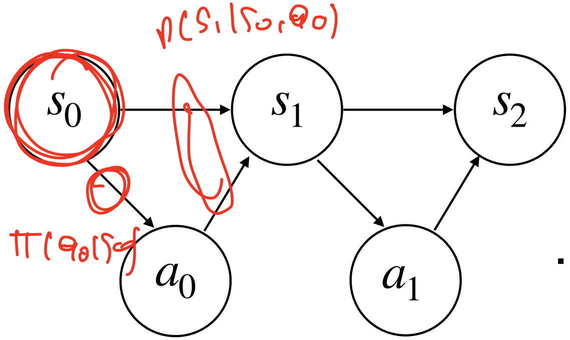
Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?



Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?

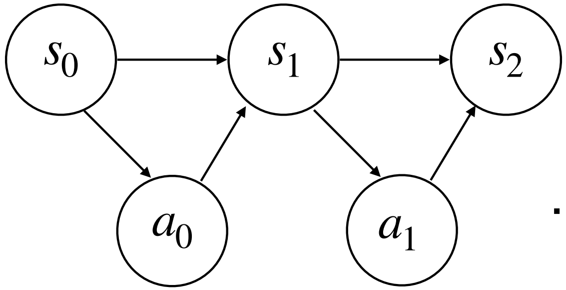


$$\ldots \quad \mathbb{P}^{\pi}(s_0, a_0, \ldots, s_h, a_h)$$

$$\ldots \quad = \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)P(s_2 \mid s_1, a_1)\ldots P(s_h \mid s_{h-1}, a_{h-1})\pi(a_h \mid s_h)$$

Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?



… $\mathbb{P}^{\pi}(s_0, a_0, \ldots, s_h, a_h)$

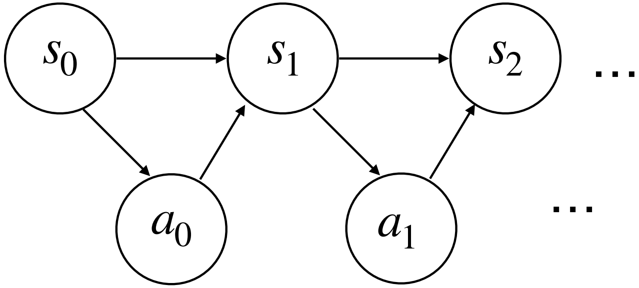
… $= \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)P(s_2 \mid s_1, a_1)\ldots P(s_h \mid s_{h-1}, a_{h-1})\pi(a_h \mid s_h)$

$P(A, B)$ $\qquad$ $P(A=a) = \sum_b P(A=a, B=b)$

Q: what's the probability of π visiting state (s,a) at time step h?

Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?



$$\mathbb{P}^{\pi}(s_0, a_0, \ldots, s_h, a_h)$$

$$= \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)P(s_2 \mid s_1, a_1)\ldots P(s_h \mid s_{h-1}, a_{h-1})\pi(a_h \mid s_h)$$

Q: what's the probability of π visiting state (s,a) at time step h?

$$\mathbb{P}_h^{\pi}(s, a) = \sum_{s_0, a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^{\pi}(s_0, a_0, \ldots, s_{h-1}, a_{h-1}, s_h = s, a_h = a)$$

Averaged state action occupancy measure

$\mathbb{P}_h^\pi(s, a)$: probability of π visiting (s, a) at time step $h \in \mathbb{N}$

$$d^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a)$$

$\gamma^0 P_0^\pi(s,a) + \gamma^1 P_1^\pi(s,a) + \ldots\ldots +$

Averaged state action occupancy measure

$d^\pi(s) = (1-\gamma)\sum_h \gamma^h P_h^\pi(s)$ $\quad d^\pi(s) = (1-\gamma)\cdot 1 + \gamma \sum_{s'} P(s'|s,a) d^\pi(s')$

$\mathbb{P}_h^\pi(s,a)$: probability of π visiting (s,a) at time step $h \in \mathbb{N}$

$$d^\pi(s,a) = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathbb{P}_h^\pi(s,a)$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s,a \sim d^\pi}[r(s,a)]$$

$$\mathbb{E}_{s_0\sim\mu}V^\pi(s_0) = \frac{1}{1-\gamma}\sum_{s,a} d^\pi(s,a) r(s,a)$$

Summary for today

Planning algorithm (no learning so far):

VI: fixed point iteration $Q^{t+1} = \mathcal{T} Q^t$

1. Bellman operator is a contraction map

2. $\|Q^t - Q^\star\|_\infty$ being small implies V^{π^t} & $V^\star$ are close