Planning in Markov Decision Processes

Announcements

HW0: due this Thursday 11:59pm
Gradescope (please self-enroll)

Announcements

HW0: due this Thursday 11:59pm
Gradescope (please self-enroll)

Waiting List

Recap: Infinite Horizon MDPs

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

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

Recap: Infinite Horizon MDPs

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

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

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

Recap: Infinite Horizon MDPs

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

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

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

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

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

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

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

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

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

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

Recap: 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

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

Theorem 2:

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

Main Question for Today:

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

Main Question for Today:

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

Two Approaches:

1. Value Iteration
2. Policy Iteration

Define Bellman Operator $\mathscr{T}$:

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

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

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

Value Iteration Algorithm:

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

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

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 $f = \mathcal{T} f$

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 $x^\star = \ell(x^\star)$

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

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 $x^\star = \ell(x^\star)$

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

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

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 $x^\star = \ell(x^\star)$

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

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

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

Convergence of Value Iteration:

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

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

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:

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

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:

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

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

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:

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

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

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

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:

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

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:

$$\begin{aligned}
|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| &= \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right| \\
&\leq \gamma \sum_{s'} P(s'|s,a) \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right| \\
&\leq \gamma \sum_{s'} P(s'|s,a) \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right| \\
&\leq \gamma \max_{s'} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right| \quad = \gamma \|Q - Q'\|_\infty
\end{aligned}$$

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:

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

Convergence of Value Iteration:

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

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

Proof:

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

$$\ldots \leq \gamma^{t+1} \|\widehat{Q}^0 - Q^\star\|_\infty$$

Final Quality of the Policy:

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

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

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:

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

Final Quality of the Policy:

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

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

Proof:

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

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

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:

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

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

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

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:

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

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

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

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

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

Proof:

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

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

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

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

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

Policy Iteration Algorithm:

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

Policy Iteration Algorithm:

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

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

Policy Iteration Algorithm:

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

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

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

Policy Iteration Algorithm:

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

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

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

Policy Iteration Algorithm:

Closed-form for PE
(see 1.1.3 in Monograph)

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

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

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

Analysis of Policy Iteration

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

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

Analysis of Policy Iteration

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

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

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

Analysis of Policy Iteration

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

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

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

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

Analysis of Policy Iteration

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

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

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

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

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

Analysis of Policy Iteration

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

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

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

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

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

Analysis of Policy Iteration

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

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

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

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

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

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

Analysis of Policy Iteration

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

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

Analysis of Policy Iteration

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

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

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

Analysis of Policy Iteration

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

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

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

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

Analysis of Policy Iteration

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

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

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

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

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

Analysis of Policy Iteration

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

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

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

Analysis of Policy Iteration

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

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

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

Value Iteration vs Policy Iteration

Which one is faster?

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

A key Lemma: Performance Difference Lemma

Q: What's the performance difference of two policies π, π', i.e.,$V^{\pi}(s) - V^{\pi'}(s)$

A key Lemma: Performance Difference Lemma

PDL Statement:

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[Q^{\pi'}(s,a) - V^{\pi'}(s)\right]$$

$$d^{\pi}_{s_0}(s,a) = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathbb{P}_h(s,a;s_0,\pi)$$

$\mathbb{P}_h(s,a;s_0,\pi)$: probability of π visiting (s,a) at time step $h \in \mathbb{N}$, starting at s_0

Proof of PDL

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_{s_0}^{\pi}}\left[A^{\pi'}(s,a)\right]$$

Proof of PDL $V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

Proof of PDL

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

Proof of PDL $V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0)$$

Proof of PDL

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$\begin{aligned} & V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0) \\ &= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(r(s_n,\pi(s_n)) + \gamma\mathbb{E}_{s_{n+1}\sim P_{s_n,\pi(s_n)}} V^{\pi'}(s_{n+1}) - V^{\pi'}(s_n)\right) \end{aligned}$$

Proof of PDL $V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_{s_0}^{\pi}}\left[A^{\pi'}(s,a)\right]$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$\begin{aligned}
&V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(r(s_n,\pi(s_n)) + \gamma\mathbb{E}_{s_{n+1}\sim P_{s_n,\pi(s_n)}} V^{\pi'}(s_{n+1}) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(Q^{\pi'}(s_n,\pi(s_n)) - V^{\pi'}(s_n)\right)
\end{aligned}$$

Proof of PDL $V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_{s_0}^{\pi}}\left[A^{\pi'}(s,a)\right]$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$\begin{aligned}
&V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(r(s_n,\pi(s_n)) + \gamma\mathbb{E}_{s_{n+1}\sim P_{s_n,\pi(s_n)}} V^{\pi'}(s_{n+1}) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(Q^{\pi'}(s_n,\pi(s_n)) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(A^{\pi'}(s_n,\pi(s_n))\right)
\end{aligned}$$

Proof of PDL $V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$\begin{aligned}
&V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(r(s_n,\pi(s_n)) + \gamma\mathbb{E}_{s_{n+1}\sim P_{s_n,\pi(s_n)}} V^{\pi'}(s_{n+1}) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(Q^{\pi'}(s_n,\pi(s_n)) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(A^{\pi'}(s_n,\pi(s_n))\right)
\end{aligned}$$

Telescoping: $V^{\pi'} - V^{\pi} = \sum_{n=0}^{\infty} V^{\pi_n} - V^{\pi_{n+1}}$