Planning in MDPs

Wen Sun

CS 6789: Foundations of Reinforcement Learning

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

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

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

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

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

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

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

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

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

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)

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

$s \cdot \xrightarrow{a} \xrightarrow{\pi^*}$

Main Question for Today:

Given an MDP $\mathcal{M} = (S, A, P, r, \gamma)$, 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 $V(s) = V^\star(s), \forall s$

$$\min_V \sum_{s \in S} \left(V(s) - \left(\max_a r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right) \right)^2$$

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

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

$f(x).\ g(x)$

$$\left| \max_x f(x) - \max_x g(x) \right|$$

$$\leq \max_x | f(x) - g(x) |$$

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

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

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

$$\left| \mathbb{E}_x f(x) \right| \leq \mathbb{E}_x \left| f(x) \right|$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$K \to +\infty$

Bellman Optimality for $Q^\star$

What about $Q^\star$?

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

Bellman Optimality for $Q^\star$

What about $Q^\star$?

We should have:

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

for all s, then $Q(s,a) = Q^\star(s,a), \forall s, a$

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

Outline

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

2. Optimal planning: Value Iteration

Define Bellman Operator $\mathcal{T}$:

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

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

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

Define Bellman Operator $\mathcal{T}$:

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

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

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

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

$= Q^\star$

Value Iteration Algorithm:

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

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

For All $s, a \in S \times A$

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

Intuition:

Via Bellman optimality theorem:

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

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 $f = \mathcal{T} f$

$x \in \mathbb{R}$

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

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

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

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

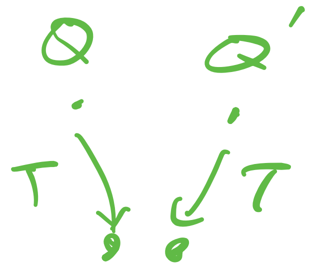
$$|x_t - x^\star| = |\ell(x_{t-1}) - \ell(x^\star)| \leq L|x_{t-1} - x^\star| \leq L^2 |x_{t-2} - x^\star|$$

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

Convergence of Value Iteration:

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

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



Proof:

Convergence of Value Iteration:

$x \in \mathbb{R}^d$

$\|x\|_\infty$

$= \max_i |x_i|$

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

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

Proof:

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

Convergence of Value Iteration:

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

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

Proof:

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

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

Convergence of Value Iteration:

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

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

Proof:

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

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

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

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:

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

Proof:

$\forall s, a$

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

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

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

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

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:

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:

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

$$\le \gamma^2 \|Q^{t-1} - Q^\star\|_\infty$$

$$\vdots$$

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

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:

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

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

Final Quality of the Policy:

$Q^t \leftarrow VI$

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

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

Proof:

$$\frac{2\gamma^t}{1-\gamma}\underbrace{\|Q^0 - Q^\star\|_\infty}_{\leq 2\cdot\frac{1}{1-\gamma}} \approx \varepsilon$$

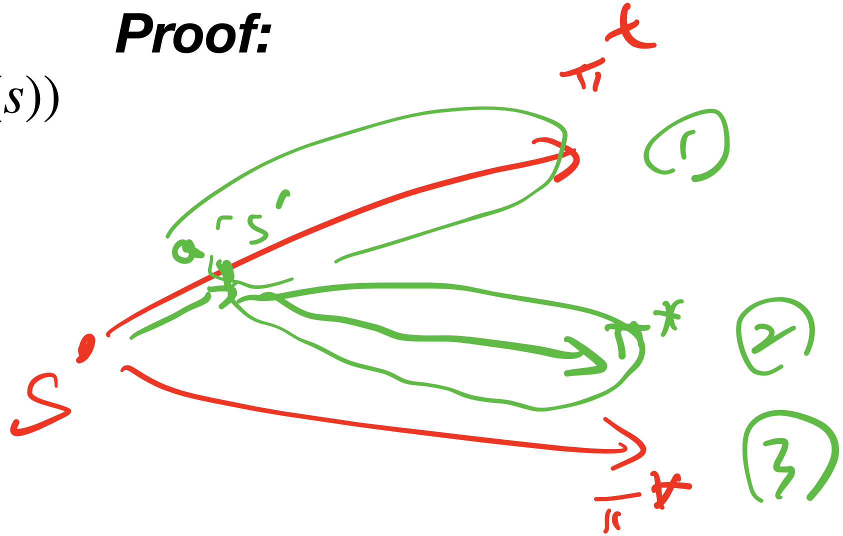
Final Quality of the Policy:

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

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:

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



$$(1) - (3) = (1) - (2) + (2) - (3)$$

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:

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

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

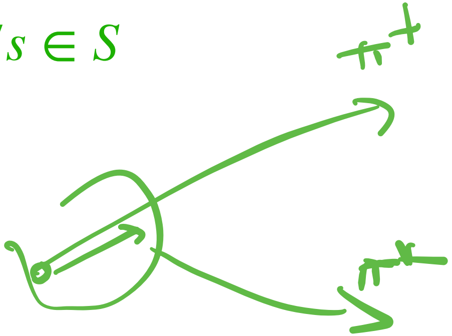
π^t

$s' \longrightarrow \pi^\star$

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:

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

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

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

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

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

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:

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

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

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

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

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:

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

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

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

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

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

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

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

Summary for today

Planning algorithm (no learning so far):

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

1. Bellman operator is a contraction map

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