Value Iteration

The Learning Strategies Center

Find Study Partners!

Studying with peers is a great way to connect with other Cornell students and is a powerful tool for learning.

Cornell's Learning Strategies Center (LSC) helps match you with study partners.

To learn more, visit the LSC's Studying Together webpage or scan the code →

Scan the QR code to find out more about Study Partners

or visit
http://lsc.cornell.edu/studying-together/

lsc.cornell.edu

Recap: Bellman Optimality

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

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

Policy $\pi : S \mapsto A$

Recap: Bellman Optimality

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

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

Policy $\pi : S \mapsto A$

Bellman Optimality—-the Q version (HW0 problem)

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

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

$$V^*(s') = \max_{a'} \Big[\overbrace{r(s',a') + \gamma \mathbb{E}_{s'' \sim P_{s',a'}} V^*(s'')}^{Q^*(s',a')} \Big]$$

Recap: Bellman Optimality

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

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

Policy $\pi : S \mapsto A$

Bellman Optimality—-the Q version (HW0 problem)

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

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

Recap: Fix~~ed~~-point solution

Find the fixed point solution of $x^\star = f(x^\star),\ x \in \mathbb{R}$

Recap: Fixed-point solution

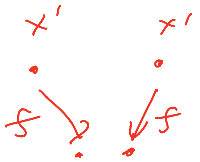
Find the fixed point solution of $x^\star = f(x^\star), x \in \mathbb{R}$

Start with some $x_0,$ set $x_{t+1} \Leftarrow f(x_t)$

Recap: Fixed-point solution

Find the fixed point solution of $x^\star = f(x^\star), x \in \mathbb{R}$

Start with some x_0, set $x_{t+1} \Leftarrow f(x_t)$



Suppose f is contraction, i.e., $\forall x, x', |f(x') - f(x)| \le \gamma |x' - x|, \gamma \in [0,1)$,
then $x_t \to x^\star$

Recap: Fixed-point solution

Find the fixed point solution of $x^\star = f(x^\star)$, $x \in \mathbb{R}$

Start with some x_0, set $x_{t+1} \Leftarrow f(x_t)$

Suppose f is contraction, i.e., $\forall x, x', |f(x') - f(x)| \leq \gamma |x' - x|, \gamma \in [0,1)$, then $x_t \to x^\star$

For Policy Evaluation (i.e., given $\mathcal{M}$ and π, compute V^π)

Recap: Fixed-point solution

Find the fixed point solution of $x^\star = f(x^\star), x \in \mathbb{R}$

Start with some x_0, set $x_{t+1} \Leftarrow f(x_t)$

Suppose f is contraction, i.e., $\forall x, x', |f(x') - f(x)| \leq \gamma |x' - x|, \gamma \in [0,1)$,
then $x_t \to x^\star$

For Policy Evaluation (i.e., given $\mathscr{M}$ and π, compute V^π)

$$V^\pi = \underbrace{R + \gamma P V^\pi}_{:=\mathcal{T}^\pi V^\pi}$$

$V^\pi = T^\pi V^\pi$

Recap: Fixed-point solution

Find the fixed point solution of $x^\star = f(x^\star)$, $x \in \mathbb{R}$

Start with some x_0, set $x_{t+1} \Leftarrow f(x_t)$

Suppose f is contraction, i.e., $\forall x, x', |f(x') - f(x)| \leq \gamma |x' - x|, \gamma \in [0,1)$,
then $x_t \to x^\star$

For Policy Evaluation (i.e., given $\mathscr{M}$ and π, compute V^π)

$$V^\pi = \underbrace{R + \gamma P V^\pi}_{:= \mathcal{T}^\pi V^\pi} \qquad V^{t+1} \Leftarrow \mathcal{T}^\pi V^t$$

$\| V^t - V^\pi \|_\infty \leq \gamma^t \| V_0 - V^\pi \|_\infty$

Question for Today:

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

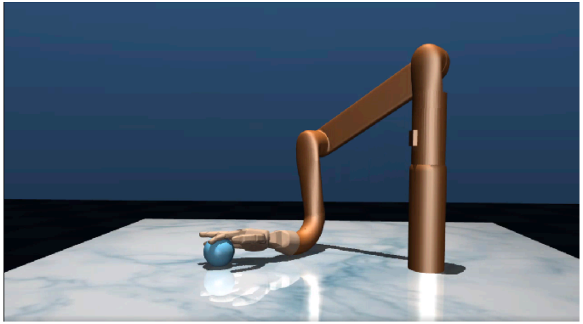
Motivation for Finding the Optimal Policy

Motivation for Finding the Optimal Policy

Find the strategy w/ the highest prob of winning
(i.e., a policy that maps the board position to the next move)

Motivation for Finding the Optimal Policy

Find the strategy w/ the highest prob of winning
(i.e., a policy that maps the board position to the next move)



Find the strategy (i.e., a mapping from robot & ball configuration to torques) that picks the ball and moves it to a goal position ASAP

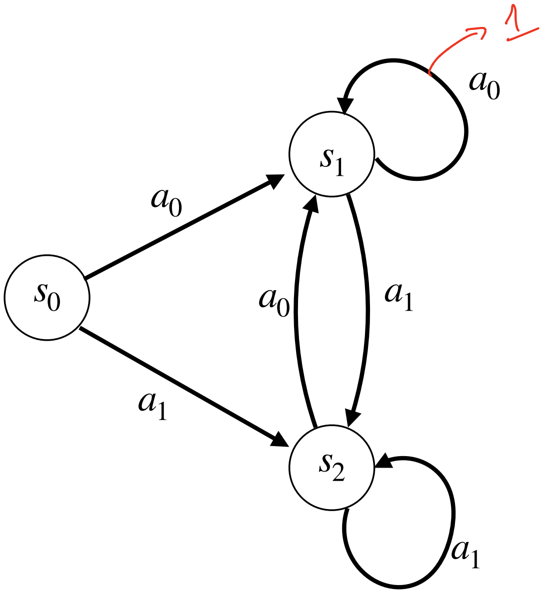
Outline:

1: An Iterative Algorithm: Value Iteration
(a fix-point iteration algorithm again!)

2: Convergence? How fast?
(Via the contraction argument again!)

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions

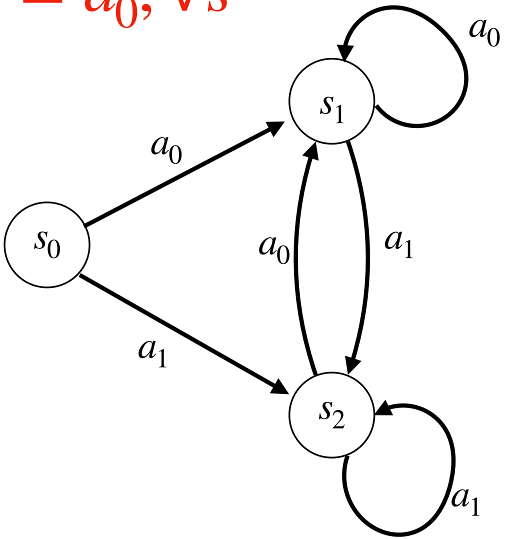


Reward: $r(s_1, a_0) = 1$, 0 everywhere else

Example of Optimal Policy $\pi^\star$

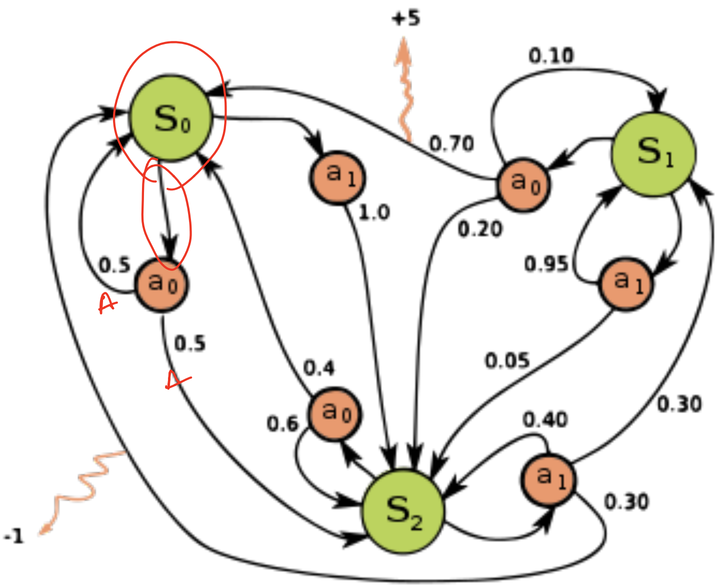
Consider the following **deterministic** MDP w/ 3 states & 2 actions

$\pi^\star(s) = a_0, \forall s$

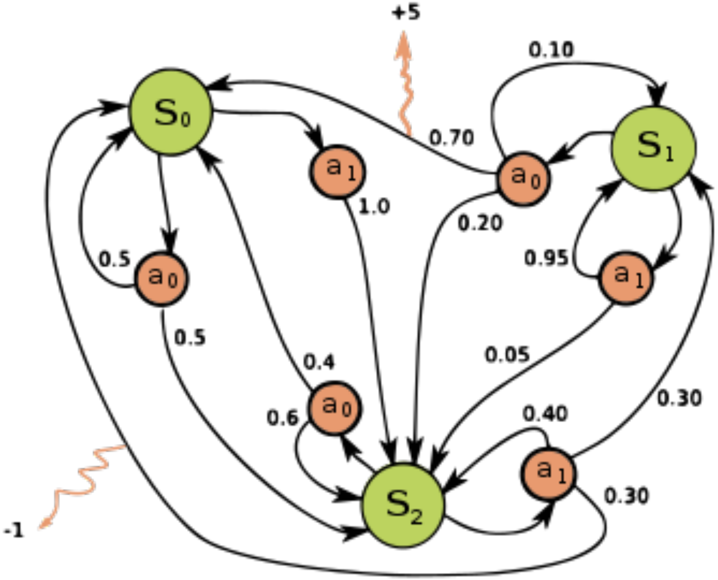


Reward: $r(s_1, a_0) = 1$, 0 everywhere else

What about this one…



What about this one…



Let's design an algorithm that computes $V^\star / Q^\star$ for any given $r \in \mathbb{R}^{|S|\times|A|}$ & $P \in \mathbb{R}^{|S|\times(|S||A|)}$

A Naive Approach (not computationally efficient)

Well, we know how to do policy evaluation for any given $\pi : S \mapsto A$, so…

A Naive Approach (not computationally efficient)

Well, we know how to do policy evaluation for any given $\pi : S \mapsto A$, so…

A^S

Enumeration:

$O(S^3)$

$\forall \pi \in S \mapsto A$, do PE, i.e., $V^\pi =$ Exact-PE(π),

then pick the policy π', such that:

$$V^{\pi'}(s) \geq V^\pi(s), \forall s, \pi$$

A Naive Approach (not computationally efficient)

Well, we know how to do policy evaluation for any given $\pi : S \mapsto A$, so...

Enumeration:

$\forall \pi \in S \mapsto A$, do PE, i.e., $V^\pi = \text{Exact-PE}(\pi)$,

then pick the policy π', such that:

$$V^{\pi'}(s) \geq V^\pi(s), \forall s, \pi$$

Computation time: $O(A^S \cdot S^3)$

A Naive Approach (not computationally efficient)

Well, we know how to do policy evaluation for any given $\pi : S \mapsto A$, so...

Enumeration:

$\forall \pi \in S \mapsto A$, do PE, i.e., $V^\pi =$ Exact-PE(π),

then pick the policy π', such that:

$$V^{\pi'}(s) \geq V^\pi(s), \forall s, \pi$$

Computation time: $O(A^S \cdot S^3)$

Can we do better? We definitely want to avoid A^S...

Define Bellman Operator $\mathcal{T}$:

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

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

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

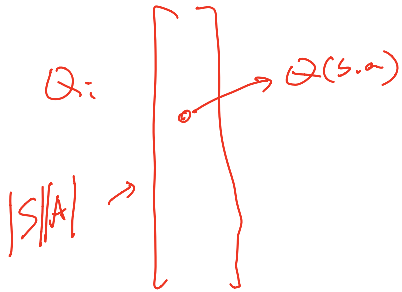
$$s, a \to \boxed{TQ} \to \text{scalar: } r(s,a) + \gamma \mathbb{E}_{s' \sim P_{sa}} \max_{a' \in A} Q(s',a')$$

Define Bellman Operator $\mathcal{T}$:

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

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

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



We can express $Q \in \mathbb{R}^{|S||A|}$, so $\underline{\mathcal{T}Q} \in \mathbb{R}^{|S||A|}$

$R + \gamma P \cdot V \Rightarrow R + \gamma P(\alpha V + \beta \cdot V') = \alpha(R + \gamma P V) + \beta(R + \gamma P V')$

$(\alpha + \beta = 1)$

Define Bellman Operator $\mathcal{T}$:

$\mathcal{T}(\alpha Q + \beta Q')$

$\neq \alpha \cdot \mathcal{T} Q + \beta \cdot \mathcal{T} Q'$

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

$\max_x (f(x) + g(x))$

$\neq \max_x f(x) + \max_x g(x)$

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

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

We can express $Q \in \mathbb{R}^{|S||A|}$, so $\mathcal{T} Q \in \mathbb{R}^{|S||A|}$

i.e., think about $\mathcal{T}$ as a (non-linear) mapping that maps from $\mathbb{R}^{|S||A|}$ to $\mathbb{R}^{|S||A|}$

High Level idea for Algorithm Design

$\boxed{Q^*}$ ⇐ Goal;

Fix-point iteration again!

High Level idea for Algorithm Design

Fix-point iteration again!

$s \to \boxed{TQ}$, $a \to \boxed{TQ}$ $\to r + \gamma \mathbb{E}\, \cancel{Q}^{*}(s')$

Recall Bellman Optimality for $Q^\star$:

$$Q^\star(s,a) = \underbrace{r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} \max_{a'} Q^\star(s',a')}_{(TQ)(s,a)}$$

High Level idea for Algorithm Design

Fix-point iteration again!

Recall Bellman Optimality for $Q^\star$:

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

We have $Q^\star = \mathcal{T} Q^\star$,

i.e., $Q^\star$ is a fix-point solution of $Q = \mathcal{T} Q$

$Q =$

$Q(s,a)$

$|S||A|$

Value Iteration Algorithm:

1. Initialization: $Q^0 : \|Q^0\|_\infty \in \left[0, \frac{1}{1-\gamma}\right]$

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

$\forall s, a$

$Q^0(s,a) \in \left[0, \frac{1}{1-\gamma}\right]$

$1+\gamma+\gamma^2+\cdots = \frac{1}{1-\gamma}$

Value Iteration Algorithm:

1. Initialization: $Q^0 : \|Q^0\|_\infty \in \left[0, \dfrac{1}{1-\gamma}\right]$

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

Guarantee of VI:

The fix-point iteration converges, i.e., $Q^t \to Q^\star$, as $t \to \infty$

Summary so far:

Zooming in $Q^{t+1} \Leftarrow \underset{A}{\mathcal{T}} Q^t$**:** $\quad \mathcal{T}\mathcal{T}Q^{t-1}$

$= \mathcal{T}\mathcal{T}\mathcal{T}Q^{t-2} \cdots$

Summary so far:

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

For:

At Iteration t:

Given Q^t, we set:

For:

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

Outline:

1: An Iterative Algorithm: Value Iteration
(a fix-point iteration algorithm again!)

2: Convergence? How fast?
(Via the **contraction** argument again!)

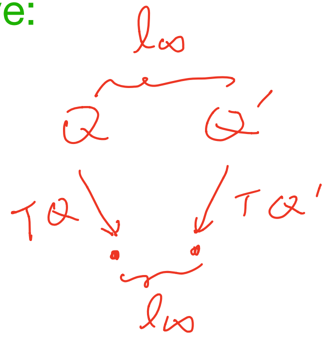
Convergence of Value Iteration:

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

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

$\|x\|_\infty = \max_i |x[i]|$

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:

$\forall s, a$

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

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:

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

$$\leq \underline{\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}} \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

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:

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

$$\leq \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} \left| \left(\underline{\max_{a'}} Q(s',a') - \underline{\max_{a'}} Q'(s',a') \right) \right|$$

$$\leq \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

$$\left| \max_{x} f(x) - \max_{x} g(x) \right|$$

$$\leq \max_{x} \left| f(x) - g(x) \right|$$

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:

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

$$\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

Avg ≤ max

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

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:

$\forall s, a$

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

$$\checkmark \leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

$$\checkmark \leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

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

Δ

$\|TQ - TQ'\|_\infty$

$\leq \gamma \|Q - Q''\|_\infty$

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:

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

$t \to \infty$

$Q^t \to Q^\star,$

$Q^t(s,a) \to Q^\star(s,a),$

$\forall s,a$

Proof ??

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

Bell-opt

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

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 ??

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

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

Summary so far:

VI (a fix point iteration alg):

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

VI convergence (via contraction)

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

Summary so far:

VI (a fix point iteration alg):

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

VI convergence (via contraction)

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

4

Next: what about the policy? Ultimately, we do want $\pi^\star$…

From Q functions to policies…

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

Recall that VI ensures that $Q^t(s, a) \approx Q^\star(s, a), \forall s, a,$ …

From Q functions to policies…

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

Recall that VI ensures that $Q^t(s, a) \approx Q^\star(s, a), \forall s, a,$ …

then maybe $\pi(s) := \arg\max_a Q^t(s, a)$ is a good choice?

The Quality of Policy:

$Q^t \to Q^\star$

$\pi^t \to \pi^\star$

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

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

$t \to \infty$

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

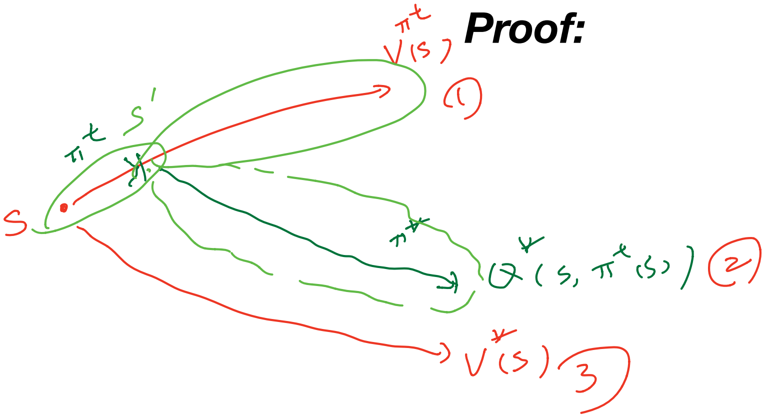
$\Rightarrow \pi^t$ is an optimal policy.

The Quality of 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:



① − ③

= ① − ② + ② − ③

The Quality of 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:

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

The Quality of 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:

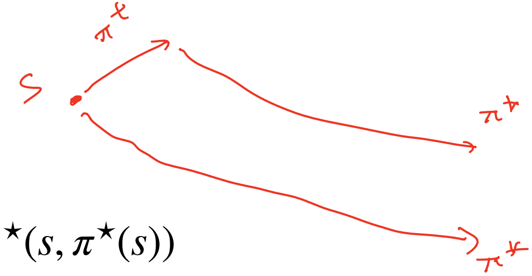
$$V^{\pi^t}(s) - V^\star(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^\star(s, \pi^\star(s))$$

$$= \underset{1}{Q^{\pi^t}(s, \pi^t(s))} \underbrace{- Q^\star(s, \pi^t(s))}_{2} + \underbrace{Q^\star(s, \pi^t(s))}_{2} - \underbrace{Q^\star(s, \pi^\star(s))}_{3}$$

The Quality of 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:

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

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

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

Cancell reward $r(s, \pi^t(s))$

The Quality of Policy:

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

by def of π^t:

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

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

$-Q^t(s, \pi^t(s)) + Q^t(s, \pi^\star(s)) \leq 0$

Proof:

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

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

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

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

The Quality of 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:

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

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

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

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

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

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

repeat!

The Quality of Policy:

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

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

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

solve for t;

Proof:

$$|Q^\star(s,\pi^t(s)) - Q^t(s,\pi^t(s))| \leq \gamma^t \|Q^0 - Q^\star\|_\infty$$

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

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

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

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

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

Summary for VI:

1. VI
(a fix point iteration alg):

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

Summary for VI:

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

$\xrightarrow{\text{Contraction}}$

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

Summary for VI:

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

$$\text{VI} \xrightarrow{\text{Contraction}} \text{VI convergence}$$

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

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

Summary for VI:

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

$\xrightarrow{\text{Contraction}}$

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

$\to 0, \; t \to \infty$

Summary for this week

Bellman Equation:

$$V^{\pi} = R + \gamma P V^{\pi}$$

Bellman Optimality (Q-version):

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

Summary for this week

Bellman Equation:

$$V^{\pi} = R + \gamma P V^{\pi}$$

Fix-point Iteration framework

Bellman Optimality (Q-version):

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

Summary for this week

Bellman Equation:

$$V^\pi = R + \gamma P V^\pi$$

Iterative PE:

$$V^{t+1} \Leftarrow R + PV^t$$

Fix-point Iteration framework

Bellman Optimality (Q-version):

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

Summary for this week

Bellman Equation:

$$V^\pi = R + \gamma P V^\pi$$

Fix-point Iteration framework

Bellman Optimality (Q-version):

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

Iterative PE:

$$V^{t+1} \Leftarrow R + PV^t$$

VI:

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

Summary for this week

Contraction

Bellman Equation:

$$V^{\pi} = R + \gamma P V^{\pi}$$

Iterative PE:

$$V^{t+1} \Leftarrow R + P V^{t}$$

Fix-point Iteration framework

VI:

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

Bellman Optimality (Q-version):

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

Summary for this week

Contraction

$$\|V^t - V^\pi\|_\infty \le \gamma^t \|V^0 - V^\pi\|_\infty$$

Bellman Equation:

$$V^\pi = R + \gamma P V^\pi$$

Fix-point Iteration framework

Bellman Optimality (Q-version):

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

Iterative PE:

$$V^{t+1} \Leftarrow R + PV^t$$

VI:

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

Summary for this week

Contraction

Bellman Equation:
$V^\pi = R + \gamma P V^\pi$

Iterative PE:
$V^{t+1} \Leftarrow R + P V^t$

$\|V^t - V^\pi\|_\infty \le \gamma^t \|V^0 - V^\pi\|_\infty$

Fix-point Iteration framework

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

Bellman Optimality (Q-version):
$Q^\star = \mathcal{T} Q^\star$

Contraction

Summary for this week

Contraction

Bellman Equation:

$V^\pi = R + \gamma P V^\pi$

Iterative PE:

$V^{t+1} \Leftarrow R + PV^t$

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

Fix-point Iteration framework

VI:

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

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

Bellman Optimality (Q-version):

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

Contraction

Summary for this week

Bellman Equation:

$$V^\pi = R + \gamma P V^\pi$$

Contraction

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

Iterative PE:

$$V^{t+1} \Leftarrow R + PV^t$$

Fix-point Iteration framework

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

$$\downarrow$$

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

Bellman Optimality (Q-version):

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

Contraction

VI:

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

V^{π^t} versus $V^\star$

Next week:

1. One more algorithm (Policy Iteration) for computing $\pi^\star$

2. A continuous control model: Linear Quadratic Regulator