Value Iteration

Recap: Optimal Policy $\pi^\star$

$\gamma < 1$

For infinite horizon discounted MDP, there always exists a deterministic policy

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^{\pi}(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph—no need to understand the proof]

Recap: Optimal Policy $\pi^\star$

For infinite horizon discounted MDP, there always exists a deterministic policy

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^{\pi}(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph—no need to understand the proof]

i.e., $\pi^\star$ dominates any other policy π, everywhere!

Recap: Optimal Policy $\pi^\star$

For infinite horizon discounted MDP, there always exists a deterministic policy

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^{\pi}(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph—no need to understand the proof]

i.e., $\pi^\star$ dominates any other policy π, everywhere!

We often denote $V^\star$, $Q^\star$ in short for $V^{\pi^\star}$, $Q^{\pi^\star}$

Recap: Optimal Policy $\pi^\star$

For infinite horizon discounted MDP, there always exists a deterministic policy

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^{\pi}(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph—no need to understand the proof]

i.e., $\pi^\star$ dominates any other policy π, everywhere!

We often denote $V^\star, Q^\star$ in short for $V^{\pi^\star}, Q^{\pi^\star}$

$V^\star(s)$: the maximum value we could possibly achieve

Question for Today and Wed:

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

$\underline{\pi}$ $\quad \underline{V^\pi}$ or $\underline{Q^\pi}$

Motivation for Finding the Optimal Policy

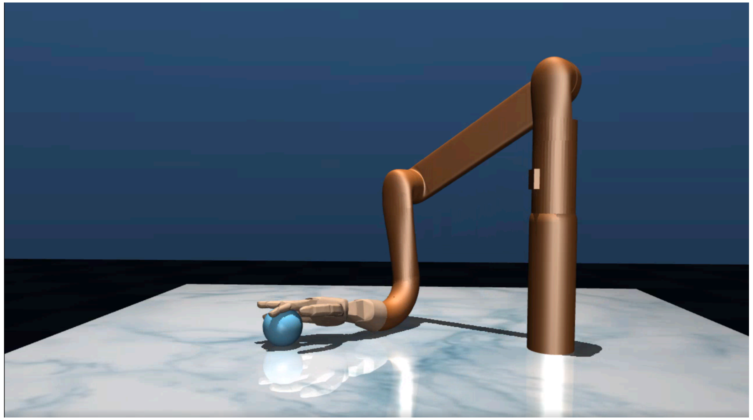
Motivation for Finding the Optimal Policy

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:

Bellman Equation

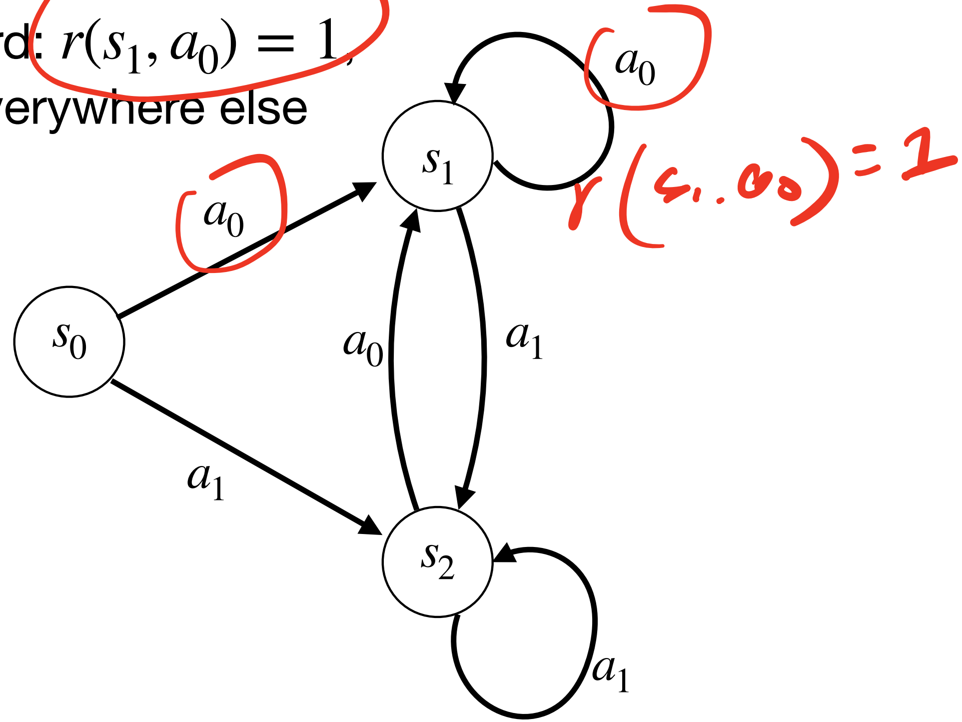
1: Bellman optimality

2: An Iterative Algorithm: Value Iteration

Example of Optimal Policy $\pi^\star$

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

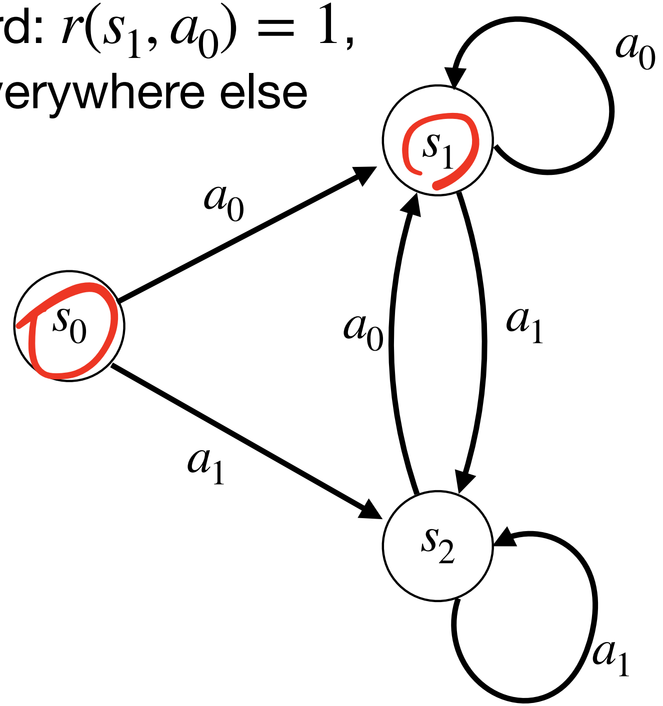
Reward: $r(s_1, a_0) = 1$,
0 everywhere else



Example of Optimal Policy $\pi^\star$

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

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



Let's say $\gamma \in (0,1)$
What's the optimal policy?

$\pi^*(s_0) = a_0$

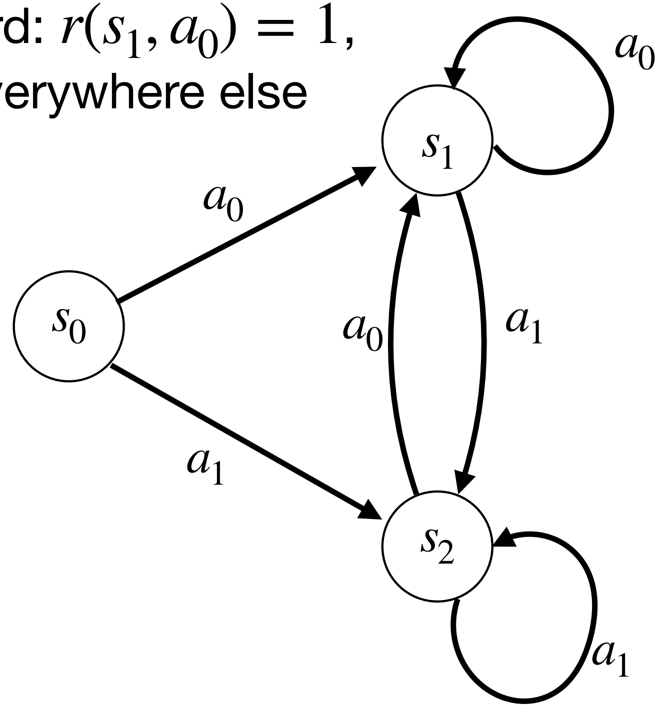
$\pi^*(s_1) = a_0$

$\pi^*(s_2) = a_0$

Example of Optimal Policy $\pi^\star$

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

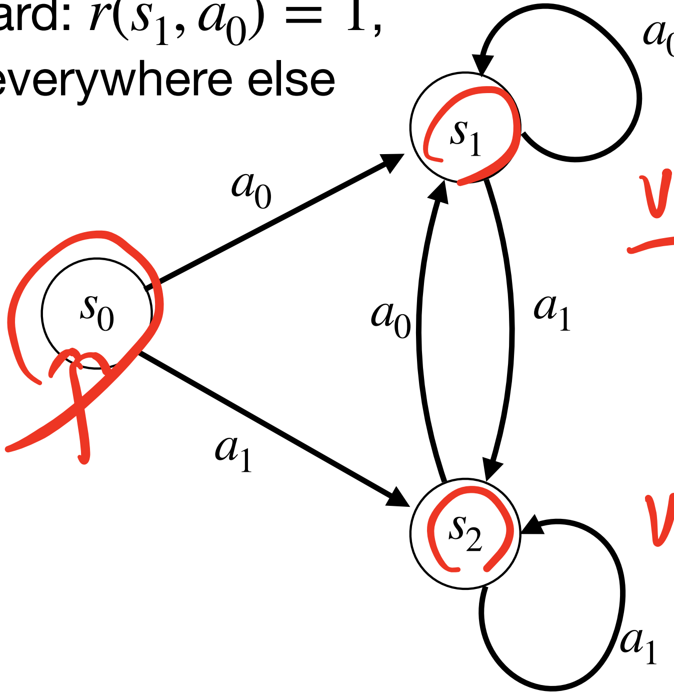
Reward: $r(s_1, a_0) = 1$,
0 everywhere else



Example of Optimal Policy $\pi^\star$

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

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



$V^\star(s_1)$

$V^\star(s_2)$

If we were told $V^\star(s_1)$ & $V^\star(s_2)$,
how to compute $V^\star(s_0)$

$$\max\left\{ \underbrace{r(s_0, a_0)}_{=0} + \gamma \cdot \underline{V^\star(s_1)},\ \underbrace{r(s_0, a_1)}_{=0} + \gamma \cdot \underline{V^\star(s_2)} \right\}$$

Bellman Optimality

Bellman Optimality

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

Bellman Optimality

$V^\star(s)$

Bellman Optimality

Value function of $\pi^\star$ $\stackrel{\pi^\star}{=}$ $$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right], \forall s$$

$$V^\star(s) = \left[r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi^\star(s))} V^\star(s') \right] \qquad \text{(By BE of } \pi^\star\text{)}$$

Bellman Optimality

Bellman Optimality

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

$$V^\star(s) = \left[r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi^\star(s))} V^\star(s') \right]$$ (By BE of $\pi^\star$)

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

$\leq$

Bellman Optimality

Bellman Optimality

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

$$V^\star(s) = \left[r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi^\star(s))} V^\star(s') \right] \qquad \text{(By BE of } \pi^\star\text{)}$$

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

$> V^\star(s)$

If we took this $\arg\max_a$ at s, then follow $\pi^\star$, we **would have higher value**

Bellman Optimality

Bellman Optimality

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

$$V^\star(s) = \left[r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi^\star(s))} V^\star(s') \right] \quad \text{(By BE of } \pi^\star\text{)}$$

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

If we took this $\arg\max_a$ at s, then follow $\pi^\star$, we **would have higher value**

Contradicts to the fact that $V^\star(s)$ is the maximum value at s one could possibly achieve

=

Bellman Optimality

Bellman Optimality

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

$Q^*(s,a)$

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

Q^π

Q^{π^*} (Q^*)

$Q^*(s,a)$ $s \xrightarrow{a} \xrightarrow{\pi^*}$

Bellman Optimality

Bellman Optimality

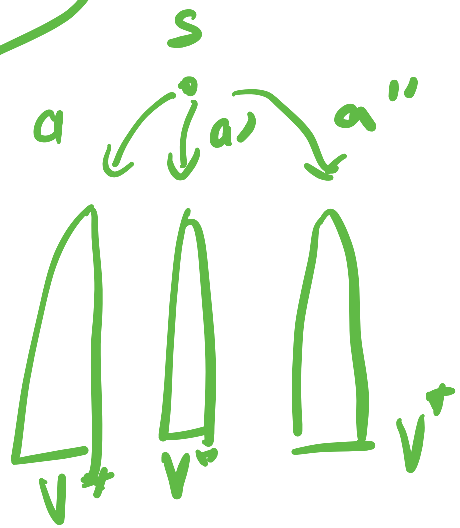
$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

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

This implies that $\arg\max_a Q^\star(s,a)$ is an optimal policy

$Q \approx Q^*$

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



Bellman Optimality

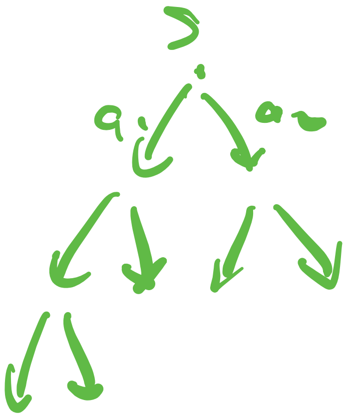
Bellman Optimality

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

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

This implies that $\arg\max_a Q^\star(s,a)$ is an optimal policy

An optimal policy should pick this action at s



Outline:

1: Bellman optimality

2: An Iterative Algorithm: Value Iteration

$\Rightarrow \hat{Q} \approx Q^*$

Define the Bellman optimality for $Q^\star$

We now know:

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

What's the Q version?

Q^* ?? $=$

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

Define the Bellman optimality for $Q^\star$

We now know:

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

What's the Q version?

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

$= V^\star(s')$

Define the Bellman optimality for $Q^\star$

We now know:

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

What's the Q version?

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

To estimate $Q^\star$, we will use the fix-point iterative approach again

Define the Bellman operator

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

$\underline{\underline{\mathcal{T}}}$

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

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

Q

$\underline{\underline{\mathcal{T}}}$ $\quad Q(s,a) \in \mathbb{R}^{|S||A|} \quad \Rightarrow \quad r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} \max_{a'} Q(s',a') \in \mathbb{R}^{|S||A|}$

Define the Bellman operator

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 $\mathcal{T}Q \in \mathbb{R}^{|S||A|}$

Define the Bellman operator

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 $\mathcal{T}Q \in \mathbb{R}^{|S||A|}$

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

$$\mathcal{T}(Q + Q') \neq \mathcal{T}Q + \mathcal{T}Q'$$

High Level idea for Algorithm Design

Fix-point iteration again!

$\top$

$\sqsupseteq$

High Level idea for Algorithm Design

Fix-point iteration again!

Recall Bellman Optimality for $Q^\star$:

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

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

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 $Q = \mathcal{T} Q$

Value Iteration Algorithm:

$\mathcal{T}(Q_1 + Q_2)$
$\neq \mathcal{T}Q_1 + \mathcal{T}Q_2$

1. Initialization: Q^0

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

For t in t=0 to T:

For All $(s,a) \in S \times A$:

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

Value Iteration Algorithm:

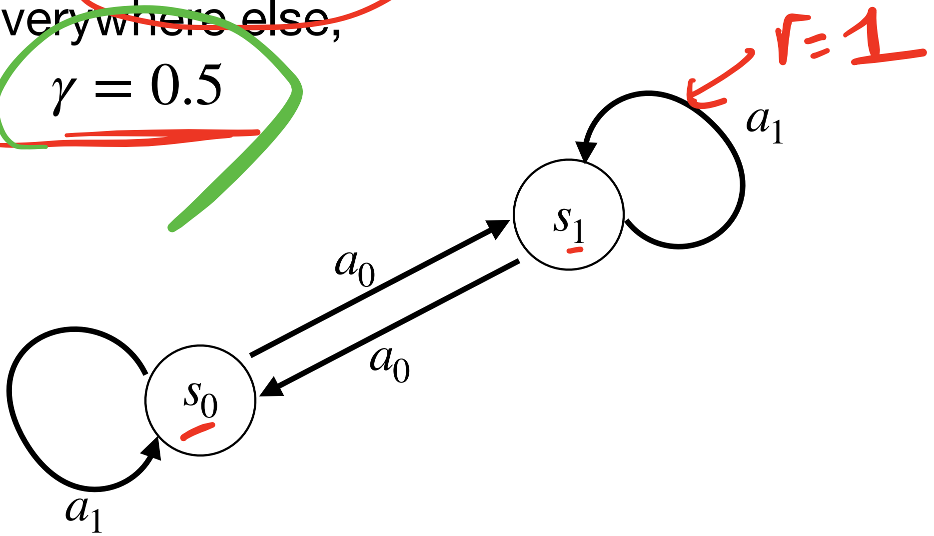
1. Initialization: Q^0

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

We hope $Q^t \rightarrow Q^\star$, as $t \rightarrow \infty$

Exercise

Reward: $r(s_1, a_1) = 1$,
0 everywhere else,
$\gamma = 0.5$



$$Q^0(s, a) = 0, \forall s, a$$

1. Compute
$Q^1(s, a), Q^2(s, a), \forall s, a$

2. Compute
$\|Q^i - Q^\star\|_\infty$ for $i \in \{0,1,2\}$

3. How does $\|Q^i - Q^\star\|_\infty$ behave as i increases

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

$\|Q^0 - Q^*\|_\infty = 2$

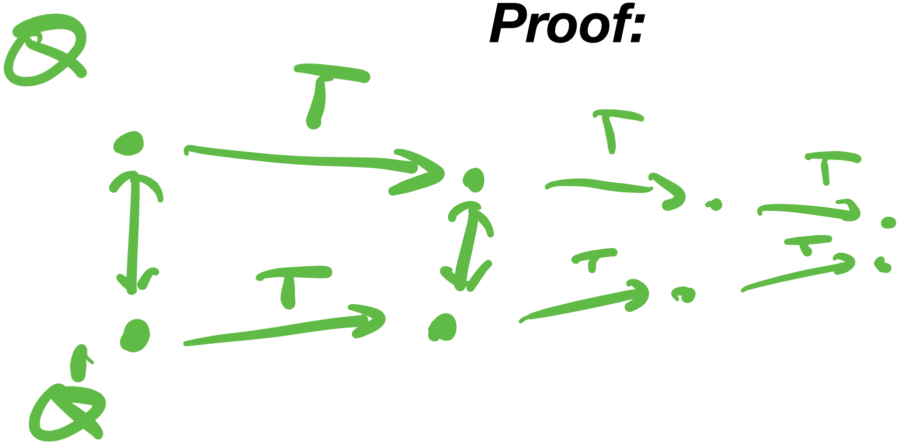
$\|Q^1 - Q^*\|_\infty = 1$ $\|Q^2 - Q^*\|_\infty = \frac{1}{2}$

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:

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

Proof:

$$|(\mathcal{T}Q)(s,a) - (\mathcal{T}Q')(s,a)| = \Big| \underbrace{\cancel{r(s,a)} + \gamma \mathbb{E}_{s'\sim P(s,a)} \max_{a'} Q(s',a')}_{\text{Bell-op on } Q} - \Big(\underbrace{\cancel{r(s,a)} + \gamma \mathbb{E}_{s'\sim P(s,a)} \max_{a'} Q'(s',a')}_{\text{Bell-op on } Q'} \Big)$$

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:

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

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

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:

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

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

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:

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

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

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

$\mathbb{E} \le \max_{s'}$

Convergence of Value Iteration:

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

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

$Q \xrightarrow{T} \cdot$

$Q' \xrightarrow{T} \cdot$

Proof:

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

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

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

Convergence of Value Iteration:

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

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

$$Q^\star \xrightarrow{T} Q^\star \xrightarrow{T} Q^\star \xrightarrow{T} Q^\star \to \to$$

$$Q^0 \xrightarrow{T} Q^1 \xrightarrow{T} Q^2 \xrightarrow{T} Q^3 \to \to$$

Convergence of Value Iteration:

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

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

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

Convergence of Value Iteration:

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

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

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

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

Summary so far:

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

$Q^\star$

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

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

$t = 1000 \quad Q^{1000} \approx Q^\star$

What about the policy? Ultimately, we do want $\pi^\star$…

$\hat{\pi} \leftarrow \arg\max_a Q^{1000}(s, a)$