Policy Evaluation & Policy Iteration

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning
Fall 2022

Today

- HW2 posted
- Recap
- Today:
  - Value Iteration works directly with a vector V which converging to V*.
    Is there an iterative algorithm that more directly works with policies?
    - Part 1: policy evaluation.
    - Part 2: policy iteration.

Recap

Define Bellman Operator $\mathcal{T}$:

Bellman Equations: $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$

- Any function $V : S \mapsto \mathbb{R}$ can also be viewed as a vector in $V \in \mathbb{R}^{|S|}$.
- Define $\mathcal{T} : \mathbb{R}^{|S|} \mapsto \mathbb{R}^{|S|}$, where

$$\left(\mathcal{T} V\right)(s) := \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$$

- Bellman equations in terms of $\mathcal{T}$:

$$\mathcal{T} V = V$$

Value Iteration Algorithm:

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

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

What is the Per-Iteration Computational Complexity?

- Making the update $V^{t+1} \leftarrow \mathcal{T} V^t$ explicit:
  - Define Q^{t+1}:

$$\forall s, a \;\; Q^{t+1}(s,a) = r(s,a) + \gamma \sum_{s' \in S} P(s' \mid s, a) V^t(s')$$

  - Set V^{t+1}:

$$\forall s \;\; V^{t+1}(s) = \max_a Q^{t+1}(s,a)$$

implicity corresponds to a policy

- What is the order of the number of basic arithmetic operations?

$O(|S|^2 |A|)$

Convergence of Value Iteration:

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

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

$$\Longrightarrow$$

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

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

Computational Complexity of VI

(for approximating $V^\star$)

Runtime: VI will return a V^t s.t. $\|V^t - V^\star\|_\infty \leq \epsilon$ in no more than

$$\frac{\ln(\| V^0 - V^\star \|_\infty / \epsilon)}{(1-\gamma)}$$ iterations.

if $\|V^0 - V^*\|_\infty \leq \frac{1}{1-\gamma}$

and then

So the computational complexity for an ϵ-accurate solution is

$$O\left(\frac{|S|^2|A|}{1-\gamma} \ln\left(\frac{1}{\epsilon(1-\gamma)}\right)\right)$$

But what about the policy we find with VI?

Theorem: For any V, let $\pi(s) = \arg\max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$, then

$$V^*(s) \geq V^\pi(s) \geq V^\star(s) - \frac{2\gamma}{1-\gamma} \| V - V^\star \|_\infty$$

But what about the policy we find with VI?

Theorem: For any V, let $\pi(s) = \arg\max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$, then

$$V^{\pi}(s) \geq V^{\star}(s) - \frac{2\gamma}{1-\gamma} \| V - V^{\star} \|_{\infty}$$

Runtime: After $\frac{\ln\left(2/\left((1-\gamma)^2 \epsilon\right)\right)}{1-\gamma}$ iterations of PI, we have: $V^{\pi^t}(s) \geq V^{\star}(s) - \epsilon$,

and, the total runtime of VI is:

$$O\left(\frac{|S|^2 |A|}{1-\gamma} \ln\left(1/\left((1-\gamma)^2 \epsilon\right)\right) \right)$$

But what about the policy we find with VI?

$Q^*(s, \pi^*(s))$

Theorem: For any V, let $\pi(s) = \arg\max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$, then

$$V^\pi(s) \geq V^\star(s) - \frac{2\gamma}{1-\gamma} \| V - V^\star \|_\infty$$

$V^*(s) = Q^*(s, \pi^*(s))$

$V^\pi(s) = Q^\pi(s, \pi^t(s))$

$- Q^*(s, \pi^t(s))$

Runtime: After $\dfrac{\ln\left(2/\left((1-\gamma)^2\epsilon\right)\right)}{1-\gamma}$ iterations of ~~PI~~, we have: $V^{\pi^t}(s) \geq V^\star(s) - \epsilon$,

VI

and, the total runtime of VI is:

$$O\left(\frac{|S|^2 |A|}{1-\gamma} \ln\left(1/\left((1-\gamma)^2\epsilon\right)\right) \right)$$

$\frac{(1-\gamma)\varepsilon}{2} \leq \frac{(1-\gamma)\varepsilon}{2\gamma}$

We replace $\epsilon \leftarrow (1-\gamma)\epsilon/2$, then VI will return V^t s.t. $\|V^t - V^\star\|_\infty \leq (1-\gamma)\epsilon/2$.

Thus, $V^{\pi^t}(s) \geq V^\star(s) - \dfrac{2\gamma}{1-\gamma} \|V^t - V^\star\|_\infty \geq V^\star(s) - \epsilon$

Today:

Let's start with Policy Evaluation

Given MDP $\mathscr{M} = (S, A, r, P, \gamma)$ & a policy $\pi : S \mapsto A$, how do we compute $V^{\pi}(s)$?

Exact Policy Evaluation

Exact Policy Evaluation

- V^π satisfies the Bellman consistency conditions:

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(s, \pi(s))} V^\pi(s')$$

Exact Policy Evaluation

- V^π satisfies the Bellman consistency conditions:

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(s, \pi(s))} V^\pi(s')$$

- or, equivalently,

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V^\pi(s')$$

Exact Policy Evaluation

- V^π satisfies the Bellman consistency conditions:

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(s, \pi(s))} V^\pi(s')$$

- or, equivalently,

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V^\pi(s')$$

- This gives us $|S|$ linear constraints.

Exact Policy Evaluation

- V^π satisfies the Bellman consistency conditions:

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(s, \pi(s))} V^\pi(s')$$

- or, equivalently,

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V^\pi(s')$$

- This gives us $|S|$ linear constraints.

dependent on π

- **Exact algorithm:** Find V that solves the following linear system:

$$\forall s, V(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V(s')$$

Exact Policy Evaluation

- V^π satisfies the Bellman consistency conditions:

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(s, \pi(s))} V^\pi(s')$$

- or, equivalently,

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V^\pi(s')$$

- This gives us $|S|$ linear constraints.

- **Exact algorithm:** Find V that solves the following linear system:

$$\forall s, V(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V(s')$$

- **Theorem:** This system of linear equations has a unique solution, which is V^π.

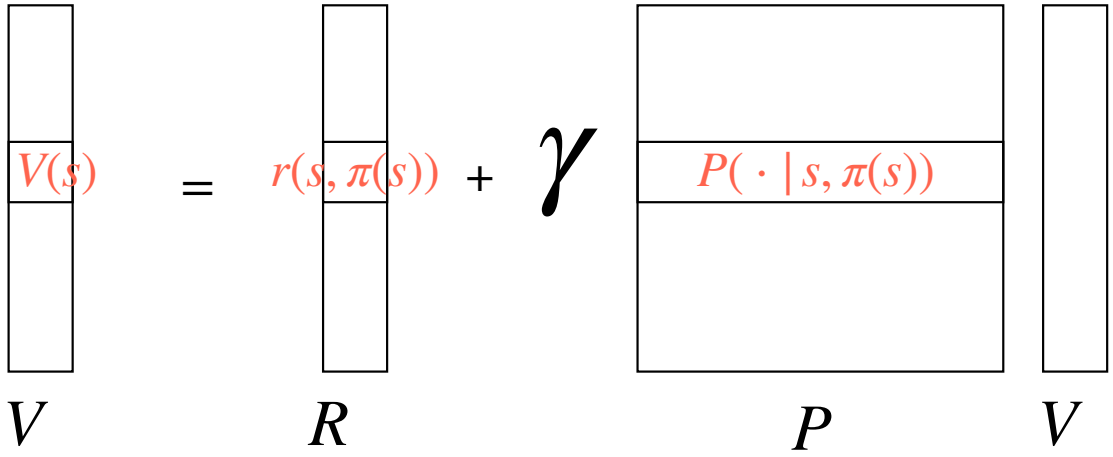
Exact Policy Evaluation: Matrix Version

Exact Policy Evaluation: Matrix Version

- Define: $R^{\pi} \in \mathbb{R}^{|S|}$, where $R^{\pi}_{s} = r(s, \pi(s))$, and $P^{\pi} \in \mathbb{R}^{|S| \times |S|}$, where $P^{\pi}_{s',s} = P(s' \mid s, \pi(s))$

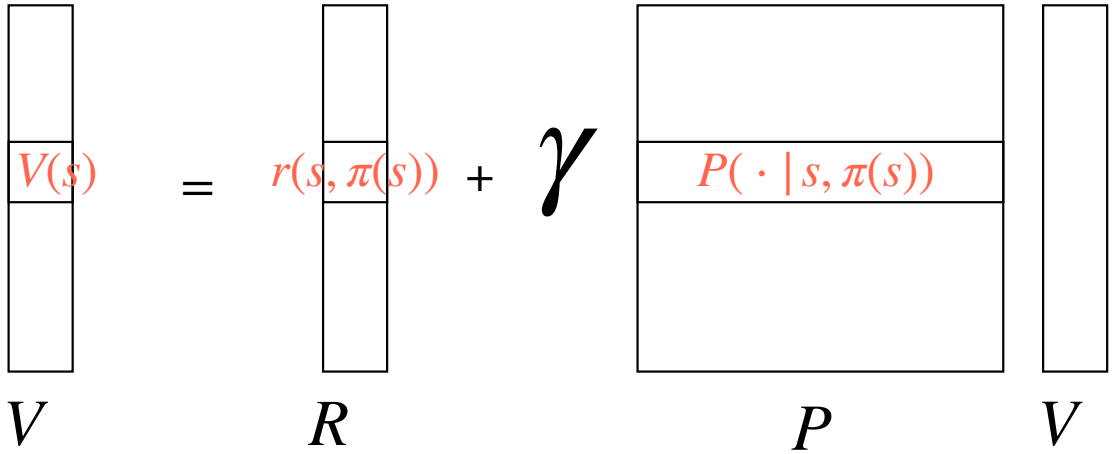
Exact Policy Evaluation: Matrix Version

- Define: $R \in \mathbb{R}^{|S|}$, where $R_s^\pi = r(s, \pi(s))$, and $P^\pi \in \mathbb{R}^{|S| \times |S|}$, where $P^\pi_{s',s} = P(s' \mid s, \pi(s))$
- So we want to find $V \in \mathbb{R}^{|S|}$, s.t. $V = R^\pi + \gamma P^\pi V$



Exact Policy Evaluation: Matrix Version

- Define: $R \in \mathbb{R}^{|S|}$, where $R_s^\pi = r(s, \pi(s))$, and $P^\pi \in \mathbb{R}^{|S| \times |S|}$, where $P^\pi_{s',s} = P(s' \mid s, \pi(s))$
- So we want to find $V \in \mathbb{R}^{|S|}$, s.t. $V = R^\pi + \gamma P^\pi V$ $\quad (I - \gamma P^\pi) V = R^\pi$



- **Algo:** compute $V = (I - \gamma P^\pi)^{-1} R^\pi$

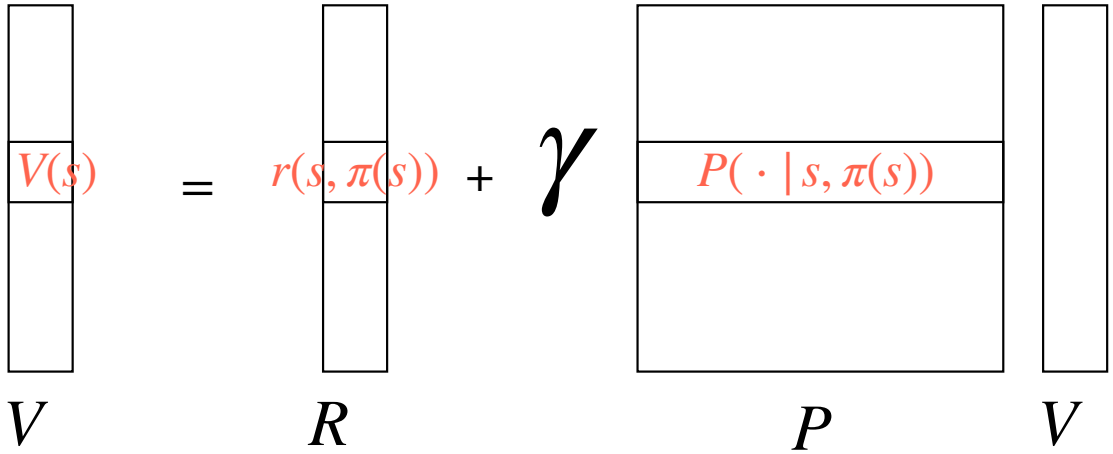
  One can show that $I - \gamma P^\pi$ is full rank (thus invertible).

$\forall \vec{x} \neq 0$

$(I - \gamma P^\pi) \vec{x} \neq 0$

Exact Policy Evaluation: Matrix Version

- Define: $R \in \mathbb{R}^{|S|}$, where $R_s^\pi = r(s, \pi(s))$, and $P^\pi \in \mathbb{R}^{|S| \times |S|}$, where $P_{s',s}^\pi = P(s' \mid s, \pi(s))$
- So we want to find $V \in \mathbb{R}^{|S|}$, s.t. $V = R^\pi + \gamma P^\pi V$



- **Algo:** compute $V = (I - \gamma P^\pi)^{-1} R^\pi$

  One can show that $I - \gamma P^\pi$ is full rank (thus invertible).
- **Runtime:** This approach runs in time $O(|S|^3)$.

Is there an iterative version?

(that is faster, but approximate?)

Algorithm (Iterative PE):

1. Initialization: $V^0 : \|V^0\|_\infty \in \left[0, \frac{1}{1-\gamma}\right]$
2. Iterate until convergence: $V^{t+1} \leftarrow R^\pi + \gamma P^\pi V^t$

Is there an iterative version?

(that is faster, but approximate?)

Algorithm (Iterative PE):

1. Initialization: $V^0 : \|V^0\|_\infty \in \left[0, \frac{1}{1-\gamma}\right]$
2. Iterate until convergence: $V^{t+1} \leftarrow R + \gamma P V^t$

What's the computational complexity per iteration?

$O(S^2)$

Contraction of Iterative PE

Theorem: After t iterations, we have:

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

Contraction of Iterative PE

Theorem: After t iterations, we have:

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

Proof: (really the same as before)

Contraction of Iterative PE

Theorem: After t iterations, we have:

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

Proof: (really the same as before)

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

Runtime Comparison:

Runtime Comparison:

- **Runtime of Iterative PE:** After $\ln\left(\left\| V^0 - V^{\pi} \right\|_\infty / \epsilon \right) / (1-\gamma)$ iterations of iterative PE, we have $\|V^t - V^{\pi}\|_\infty \leq \epsilon$.

  Thus, the total runtime is: $O\left(\frac{|S|^2}{1-\gamma} \ln\left(1/\left((1-\gamma)\epsilon \right) \right) \right)$.

Runtime Comparison:

- **Runtime of Iterative PE:** After $\ln\left(\left\| V^0 - V^\star \right\|_\infty / \epsilon \right) / (1-\gamma)$ iterations of iterative PE, we have $\|V^t - V^\star\|_\infty \leq \epsilon$.

  Thus, the total runtime is: $O\left(\frac{|S|^2}{1-\gamma} \ln\left(1/\left((1-\gamma)\epsilon \right) \right) \right)$.

- Contrast this to the exact algo which is $O(S^3)$.

Outline:

Part 1: Policy Evaluation

Part 2: Policy Iteration

Policy Iteration (PI)

- Initialization: choose a policy $\pi^0 : S \mapsto A$
- For $t = 0,1,\ldots$
  1. **Policy Evaluation**: compute $V^{\pi^t}(s)$ and $Q^{\pi^t}(s,a)$, where
     $$Q^{\pi^t}(s,a) = r(s,a) + \gamma \sum_{s'} \underbrace{P(s' \mid s,a) V^{\pi^t}(s')}_{\mathbb{E}_{s' \sim P(\cdot \mid s,a)}\left[V^{\pi^t}(s')\right]}$$
  2. **Policy Improvement**: set
     $$\pi^{t+1}(s) := \arg\max_a Q^{\pi^t}(s,a)$$

Policy Iteration (PI)

- Initialization: choose a policy $\pi^0 : S \mapsto A$
- For $t = 0,1,\ldots$
  1. **Policy Evaluation**: compute $V^{\pi^t}(s)$ and $Q^{\pi^t}(s,a)$, where
     $$Q^{\pi^t}(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^{\pi^t}(s')$$
  2. **Policy Improvement**: set
     $$\pi^{t+1}(s) := \arg\max_a Q^{\pi^t}(s,a)$$

V^π comp. Q^π comp.

What's the computational complexity per iteration?

$O(|S|^3 + |S|^2 |A|)$

Policy Iteration (PI)

- Initialization: choose a policy $\pi^0 : S \mapsto A$
- For $t = 0,1,\ldots$
  1. **Policy Evaluation**: compute $V^{\pi^t}(s)$ and $Q^{\pi^t}(s,a)$, where
     $$Q^{\pi^t}(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^{\pi^t}(s')$$
  2. **Policy Improvement**: set
     $$\pi^{t+1}(s) := \arg\max_a Q^{\pi^t}(s,a)$$

What's the computational complexity per iteration?

$O(|S|^3 + |S|^2|A|)$

Policy Iteration (PI)

- Initialization: choose a policy $\pi^0 : S \mapsto A$
- For $t = 0,1,\ldots$
  1. **Policy Evaluation**: compute $V^{\pi^t}(s)$ and $Q^{\pi^t}(s,a)$, where
     $$Q^{\pi^t}(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^{\pi^t}(s')$$
  2. **Policy Improvement**: set
     $$\pi^{t+1}(s) := \arg\max_a Q^{\pi^t}(s,a)$$

$\max_a \{ r(s,a) + \gamma E_{s'} [V^{\pi^t}(s')] \}$

What's the computational complexity per iteration?

$O(|S|^3 + |S|^2|A|)$

What about convergence?

Two Properties of Policy Iteration:

1. Monotonic improvement:

$\forall s$

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

Two Properties of Policy Iteration:

1. Monotonic improvement:

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

2. Convergence to $V^\star$:

$$\left\| V^\star - V^{\pi^{t+1}} \right\|_\infty \leq \gamma \left\| V^\star - V^{\pi^t} \right\|_\infty$$

Monotonic Improvement of PI

Lemma: We have $V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s)$.

Monotonic Improvement of PI

Lemma: We have $V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s)$.

$\mathcal{T} V^{\pi^t}(s) \geq V^{\pi^t}(s)$

Proof:

- First, let us show that $\mathcal{T} V^{\pi^t} \geq V^{\pi^t}$.

Monotonic Improvement of PI

Lemma: We have $V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s)$.

Proof:

- First, let us show that $\mathcal{T} V^{\pi^t} \geq V^{\pi^t}$.

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

Monotonic Improvement Proof

Monotonic Improvement Proof

- By construction of π^{t+1}:
  $\mathcal{T} V^{\pi^t}(s) = r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s')$

Monotonic Improvement Proof

- By construction of π^{t+1}:
  $\mathcal{T} V^{\pi^t}(s) = r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s')$

- Using last two claims:
  $V^{\pi^{t+1}}(s) - V^{\pi^t}(s) \geq V^{\pi^{t+1}}(s) - \mathcal{T} V^{\pi^t}(s)$
  $= \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right]$

$\mathcal{T} V^{\pi^t} > V^{\pi^t}$

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

$= r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}[V^{\pi^{t+1}}]$

Monotonic Improvement Proof

- By construction of π^{t+1}:
  $\mathcal{T} V^{\pi^t}(s) = r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s')$

this is helpful in next proof.

- Using last two claims:
  $$V^{\pi^{t+1}}(s) - V^{\pi^t}(s) \geq V^{\pi^{t+1}}(s) - \mathcal{T} V^{\pi^t}(s)$$
  $$= \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right]$$

- Recursing,
  $$V^{\pi^{t+1}}(s) - V^{\pi^t}(s) \geq \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right]$$
  $$\geq \gamma^2 \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[\mathbb{E}_{s'' \sim P(s', \pi^{t+1}(s'))} \left[V^{\pi^{t+1}}(s'') - V^{\pi^t}(s'') \right] \right]$$
  $$\vdots \quad \gamma^k (\quad)$$
  $$\to 0$$

same argument

Convergence to $V^\star$

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

Convergence to $V^\star$

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

Proof:

Convergence to $V^\star$

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

just showed.

Proof:

- First, let us show that $V^{\pi^{t+1}}(s) \geq \mathcal{T} V^{\pi^t}(s) \geq V^{\pi^t}(s)$

Convergence to $V^\star$

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

Proof:

- First, let us show that $V^{\pi^{t+1}}(s) \geq \mathcal{T} V^{\pi^t}(s)$
  - As we observed in our previous proof:

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

Convergence to $V^\star$

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

Proof:

- First, let us show that $V^{\pi^{t+1}}(s) \geq \mathcal{T} V^{\pi^t}(s)$
  - As we observed in our previous proof:
    $$V^{\pi^{t+1}}(s) - \mathcal{T} V^{\pi^t}(s) = \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right]$$
  - The claim is completed since $V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \geq 0$ by monotonicity.

Convergence to $V^\star$

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

unlike VI, this conv. is on the value of the policies.

Proof:

- First, let us show that $V^{\pi^{t+1}}(s) \geq \mathcal{T} V^{\pi^t}(s)$
  - As we observed in our previous proof:
    $$V^{\pi^{t+1}}(s) - \mathcal{T} V^{\pi^t}(s) = \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right]$$
  - The claim is completed since $V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \geq 0$ by monotonicity.

- Now the proof follows using the contraction of the $\mathcal{T}$ operator:

$$0 \leq V^\star(s) - V^{\pi^{t+1}}(s) \leq V^\star(s) - \mathcal{T} V^{\pi^t}(s) = \mathcal{T} V^\star(s) - \mathcal{T} V^{\pi^t}(s)$$

$$\leq \gamma \|V^\star - V^{\pi^t}\|_\infty$$

contraction of $\mathcal{T}$.

Runtime of PI:

Runtime of PI:

Runtime of PI:

After $\dfrac{\ln\left(\left\| V^{\pi^0} - V^\star \right\|_\infty / \epsilon \right)}{1-\gamma}$ iterations of PI, we have:

$$V^{\pi^t}(s) \geq V^\star(s) - \epsilon.$$

Thus, the total runtime of PI is:

$$O\left(\frac{|S|^3 + |S|^2 |A|}{1-\gamma} \ln\left(1/\left((1-\gamma)\epsilon\right)\right) \right).$$

Runtime of PI:

Runtime of PI:

After $\dfrac{\ln\left(\left\| V^{\pi^0} - V^\star \right\|_\infty / \epsilon \right)}{1-\gamma}$ iterations of PI, we have:

$$V^{\pi^t}(s) \geq V^\star(s) - \epsilon.$$

Thus, the total runtime of PI is:

$$O\left(\frac{|S|^3 + |S|^2 |A|}{1-\gamma} \ln\left(1/\left((1-\gamma)\epsilon\right)\right) \right).$$

Comparison of VI and PI:

Runtime of PI:

Runtime of PI:

After $\dfrac{\ln\left(\left\| V^{\pi^0} - V^\star \right\|_\infty / \epsilon \right)}{1-\gamma}$ iterations of PI, we have:

$$V^{\pi^t}(s) \geq V^\star(s) - \epsilon.$$

Thus, the total runtime of PI is:

$$O\left(\frac{|S|^3 + |S|^2 |A|}{1-\gamma} \ln\left(1/\left((1-\gamma)\epsilon\right)\right) \right).$$

Per iteration comp. of VI is

$O(|S|^2|A|)$

Comparison of VI and PI:

- Per iteration complexity of VI is less than that of PI.

Runtime of PI:

Runtime of PI:

After $\dfrac{\ln\left(\left\| V^{\pi^0} - V^\star \right\|_\infty / \epsilon \right)}{1-\gamma}$ iterations of PI, we have:

$$V^{\pi^t}(s) \geq V^\star(s) - \epsilon.$$

Thus, the total runtime of PI is:

$$O\left(\frac{|S|^3 + |S|^2 |A|}{1-\gamma} \ln\left(1/\left((1-\gamma)\epsilon\right)\right) \right).$$

Comparison of VI and PI:

- Per iteration complexity of VI is less than that of PI.
- PI and VI have the same upper bound on the # of iterations.

Runtime of PI:

Runtime of PI:

After $\dfrac{\ln\left(\left\| V^{\pi^0} - V^\star \right\|_\infty / \epsilon\right)}{1-\gamma}$ iterations of PI, we have:

$$V^{\pi^t}(s) \geq V^\star(s) - \epsilon.$$

Thus, the total runtime of PI is:

$$O\left(\frac{|S|^3 + |S|^2|A|}{1-\gamma} \ln\left(1/\left((1-\gamma)\epsilon\right)\right)\right).$$

This is not an easy result to prove and certainly does not follow from what we proved in class.

Thm: PI exactly hits a π^* in $O\left(\frac{|S|^2}{1-\gamma} \log \frac{|S|^2}{1-\gamma}\right)$ iterations.

Comparison of VI and PI:

- Per iteration complexity of VI is less than that of PI.
- PI and VI have the same upper bound on the # of iterations.
- In practice, PI reaches a better policy more quickly than VI.

(see HW "Comments on Computational Complexity" for theoretical justification)

want to find V s.t.

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

some extras (not necessary)

Linear Program Formulation:

$$\min_{\vec{V} \in \mathbb{R}^{|S|}} \sum_s V(s)$$

$$\text{s.t.} \left\{ \forall s,a \quad V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \right\}$$

(This is the "feasible set" $\mathcal{F}$, which is a convex polytope with $|S||A|$ linear constraints.

claim 1: $V^* \in \mathcal{F}$ (easy to see)

claim 2: if $V \in \mathcal{F} \Rightarrow \vec{V} \geq \vec{V}^*$

(not too difficult to prove)

so $\mathcal{F}$ contains vectors greater than V^*

claim 3: The solution to the LP is V^*

(immediately follows from 1 & 2).

[Interesting Question: What is the dual LP ??]