

# **Policy Evaluation**

# Announcements

HW0 is out today (due March 2nd, 11:59 pm ET)

(Gradescope entry code: BP3B3N)

Office hours start this week:

**Wen:** Tuesday and Thursday, 10:55am - 11:30am

**Wen-Ding:** Friday 3pm-4pm

**Hadi:** Wednesday 2:30-3:30pm

## Recap: Definitions

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

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

# Recap: Definitions

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

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

Policy  $\pi : S \mapsto A^S$

# Recap: Definitions

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

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

Policy  $\pi : S \mapsto A$

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

# Recap: Definitions

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

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

Policy  $\pi : S \mapsto A$

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

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

## Recap: Optimal Policy

For Discounted infinite horizon MDP,  $\exists$  a deterministic policy  $\pi^* : S \mapsto A$ :

$$V^*(s) \geq V^\pi(s), \forall s, \forall \pi$$

# Recap: Optimal Policy

For Discounted infinite horizon MDP,  $\exists$  a deterministic policy  $\pi^* : S \mapsto A$ :

$$V^*(s) \geq V^\pi(s), \forall s, \forall \pi$$

## Bellman Optimality (DP):

1. For  $V^*$ , we have  $V^*(s) = \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s') \right], \forall s$

# Recap: Optimal Policy

For Discounted infinite horizon MDP,  $\exists$  a deterministic policy  $\pi^* : S \mapsto A$ :

$$V^*(s) \geq V^\pi(s), \forall s, \forall \pi$$

## Bellman Optimality (DP):

1. For  $V^*$ , we have  $V^*(s) = \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s') \right], \forall s$
2. For  $V$  that satisfies  $V(s) = \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s') \right], \forall s$ ,  
we have  $V(s) = V^*(s), \forall s$

# Recap: Optimal Policy

For Discounted infinite horizon MDP,  $\exists$  a deterministic policy  $\pi^* : S \mapsto A$ :

$$V^*(s) \geq V^\pi(s), \forall s, \forall \pi$$

**Bellman Optimality (DP):**

$$\begin{aligned} V^*(s) &= r(s, \pi(s)) \\ &+ \gamma \mathbb{E}_{s' \sim P(\cdot|s, \pi(s))} V^*(s') \end{aligned}$$

Bellman Eq

1. For  $V^*$ , we have  $V^*(s) = \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s') \right], \forall s$

2. For  $V$  that satisfies  $V(s) = \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s') \right], \forall s$ ,  
we have  $V(s) = V^*(s), \forall s$

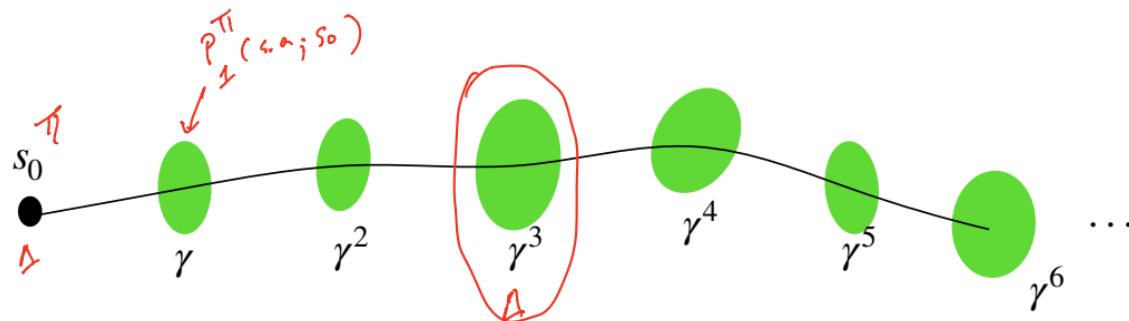
In HW0, we will study Bellman Optimality for  $Q^*/Q$

## Recap: State-action distribution

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{\substack{\text{a red arrow points to this part} \\ a_0, s_1, a_1, \dots, s_{h-1}, a_{h-1}}} \mathbb{P}^\pi(s_0, a_0, \dots, s_{h-1}, a_{h-1} | s_h = s, a_h = a)$$

# Recap: State-action distribution

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \dots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \dots, s_{h-1}, a_{h-1}; s_h = s, a_h = a)$$



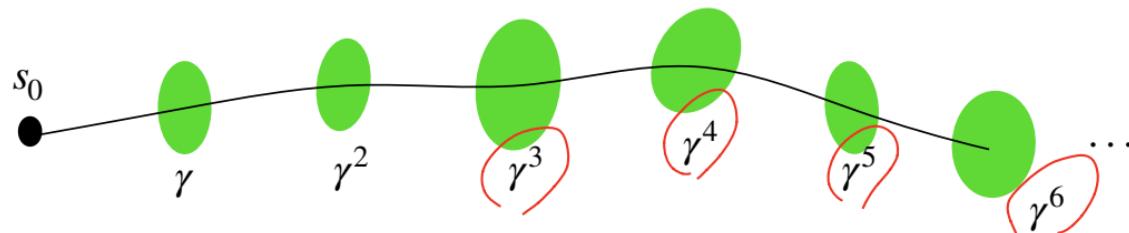
## Recap: State-action distribution

$$\frac{1-\gamma}{1-\gamma} \Rightarrow 1 + \gamma + \gamma^2 + \gamma^3 -$$
$$\gamma = \frac{1}{1-\gamma}$$

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \dots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \dots, s_{h-1}, a_{h-1}; s_h = s, a_h = a)$$

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

using  $V^\pi(s_0)$  and  $d_{s_0}^\pi(s, a)$



# Today: Policy Evaluation

**Key Question:**

**Given MDP  $\mathcal{M} = (S, A, r, P, \gamma)$  & a  $\pi : S \mapsto A$ ,  
how good is  $\pi$ ?**

i.e., how to compute  $\underbrace{V^\pi(s), \forall s}$ ?

# Motivation for Policy Evaluation



We want to **evaluate** our strategy against some opponent  
(we can abstract our strategy as policy  $\pi$ )

# Motivation for Policy Evaluation



We want to **evaluate** our strategy  
against some opponent  
(we can abstract our strategy as  
policy  $\pi$ )



We want to **evaluate** our  
recommendation strategy before  
we release it to users

## A more fundamental motivation...

Recall that we have  $A^S$  many policies.

To select the optimal policy, we need to do evaluation

# Outline:

1. **Exact** Policy Evaluation  $v^\pi$

2. **Approximate** Policy Evaluation via an Iterative Algorithm

$$v \approx v^\pi$$

$\uparrow_{BE}$

# Exact Policy Evaluation

Setup: we have MDP  $\mathcal{M} = (S, A, P, \gamma, r)$ , and policy  $\pi$ , we want to compute  $V^\pi$

# Exact Policy Evaluation

Setup: we have MDP  $\mathcal{M} = (S, A, P, \gamma, r)$ , and policy  $\pi$ , we want to compute  $V^\pi$

We know that for  $V^\pi$ , we have **Bellman equation:**

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

# Exact Policy Evaluation

Setup: we have MDP  $\mathcal{M} = (S, A, P, \gamma, r)$ , and policy  $\pi$ , we want to compute  $V^\pi$

We know that for  $V^\pi$ , we have **Bellman equation**:

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

This gives us  $S$  many linear constraints

# Exact Policy Evaluation

Let's form linear constraints. Denote  $V(s)$  as our estimator for  $s \in S$

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

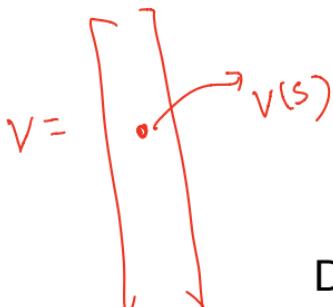
$$V(s_1) = r(s_1, \pi(s_1)) + \sum_{s' \in S} P(s' | s_1, \pi(s_1))V(s')$$

$$V(s_2) = r(s_2, \pi(s_2)) + \sum_{s' \in S} P(s' | s_2, \pi(s_2))V(s')$$

# Exact Policy Evaluation

Let's form linear constraints. Denote  $V(s)$  as our estimator for  $s \in S$

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



Denote  $V \in \mathbb{R}^{|S|}$ ,  $R \in \mathbb{R}^{|S|}$ , where  $R_s = r(s, \pi(s))$ , and  
 $P \in \mathbb{R}^{|S| \times |S|}$ , where  $P_{s',s} = P(s'|s, \pi(s))$ ,

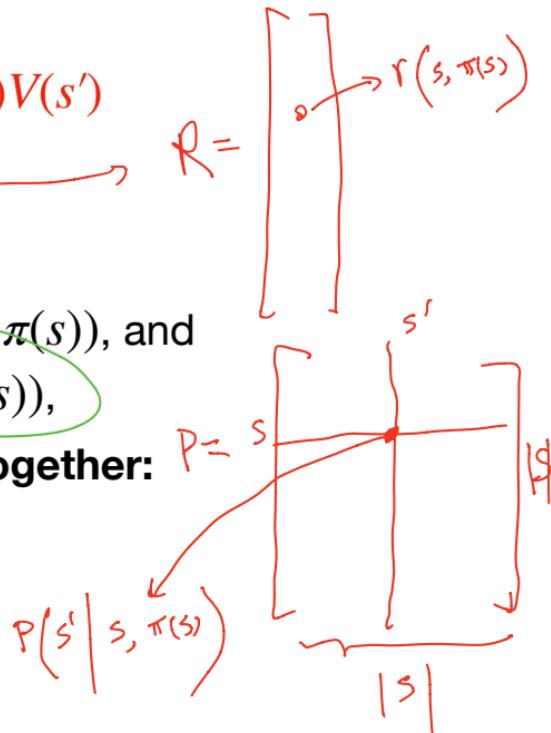
we can **combine all  $S$  many constraints together:**

\* (correction)

$$V = R + \gamma PV$$

$$P_{s,s'} = P(s'|s, \pi(s))$$

i.e., row is indexed by  $s$ , column is indexed by  $s'$



$\pi: S \mapsto A$

## Exact Policy Evaluation

$V \in \mathbb{R}^{|S|}$ ,  $R \in \mathbb{R}^{|S|}$ , where  $R_s = r(s, \pi(s))$ , and  $P \in \mathbb{R}^{|S| \times |S|}$ , where  $P_{s',s} = P(s' | s, \pi(s))$ , we can combine all constraints together:

$$V = R + \gamma PV$$
$$\sum_{s'} P(s' | s, \pi(s)) \cdot V(s')$$
$$= E_{s' \sim P(\cdot | s, \pi(s))} [V(s')]$$

# Exact Policy Evaluation

$\cancel{R} \rightarrow V(I - \gamma P) = R$   
Since  $V = \cancel{R} + \gamma PV$ , we can obtain  $V$  as:

$$V = \underbrace{(I - \gamma P)^{-1}}_{\text{underlined}} R$$

## Exact Policy Evaluation

$$V - \gamma P V = R \Rightarrow (I - \gamma P) V = R$$

Since  $V = r + \gamma PV$ , we can obtain  $V$  as:

$$V = (I - \gamma P)^{-1} R$$

In HW0, we will show that  $(I - \gamma P)$  is full rank (thus invertible)

$A$  is full rank;

$$\forall x, \text{ if } x \neq \vec{0} \quad Ax \neq \vec{0}$$

## Summary so far:

$$V = r(s, \pi(s)) + \gamma P(\cdot | s, \pi(s))$$

The diagram illustrates the components of the Bellman equation. On the left, a vertical rectangle is divided into three horizontal sections. The top section is labeled  $V(s)$  in red. The middle section is empty. The bottom section is labeled  $V$  in black. An equals sign follows. To the right of the equals sign is another vertical rectangle divided into three horizontal sections. The top section is empty. The middle section is labeled  $r(s, \pi(s))$  in red. The bottom section is empty. A plus sign follows. To the right of the plus sign is the Greek letter  $\gamma$ . To the right of  $\gamma$  is a large square divided into four quadrants. The top-right quadrant contains the text  $P(\cdot | s, \pi(s))$  in red. The other three quadrants are empty. Below each of these four components is a large letter:  $R$  below the first,  $P$  below the second, and  $V$  below the third and fourth.

$$V = (I - \gamma P)^{-1} R$$

## Summary so far:

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

$$\underline{Q^\pi} \in \mathbb{R}^{S \times |A|}$$

$$V = (I - \gamma P)^{-1} R$$

Downside: computation expensive: matrix inverse is  $O(S^3)$



## Outline:

- 1. Exact Policy Evaluation  $\leftarrow O(s^3)$

2. Approximate Policy Evaluation via an Iterative Algorithm

(An approximation solution could be enough, i.e., trade accuracy for computation)

## **Detour: fix-point solution**

Consider  $x^* = f(x^*)$ ,  $f: [a, b] \mapsto [a, b]$

## Detour: fix-point solution

Consider  $x^* = f(x^*)$ ,  $f: [a, b] \mapsto [a, b]$

Common approach to find  $x^*$ :

Initialize  $x^0 \in [a, b]$ , repeat:  $\underbrace{x^{t+1} = f(x^t)}_{\nearrow x^*}$

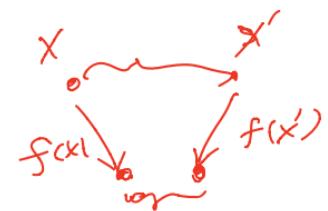
$f f f f \dots (x^*)$

## Detour: fix-point solution

Consider  $x^* = f(x^*)$ ,  $f: [a, b] \mapsto [a, b]$

Common approach to find  $x^*$ :

Initialize  $x^0 \in [a, b]$ , repeat:  $x^{t+1} = f(x^t)$



If  $f$  is a contraction mapping,

i.e.,  $\forall x, x', |f(x) - f(x')| \leq \gamma |x - x'|$ , for some  $\gamma \in [0, 1)$ , then:  
 $x^t \rightarrow x^*$ , as  $t \rightarrow \infty$

$$\left| x^{t+1} - x^* \right| = \left| f(x_t) - f(x^*) \right| \leq \gamma \left| x_t - x^* \right| \stackrel{?}{=} \gamma^2 \left| x_{t-1} - x^* \right|$$

**$V^\pi$  is a fix-point solution:**

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

$V^\pi$  is a fix-point solution:

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

$$V^\pi \in \mathbb{R}^{|S|}$$
$$V^\pi = R + \gamma P V^\pi$$

The diagram illustrates the iterative update of the value function  $V^\pi$ . It starts with the expression  $V^\pi \in \mathbb{R}^{|S|}$ , where  $|S|$  represents the number of states. This is followed by the Bellman equation  $V^\pi = R + \gamma P V^\pi$ . A green arrow labeled  $f$  points from  $V^\pi$  to  $f(V^\pi)$ , representing the function mapping the current value function to the next iteration. Another green arrow points from  $f(V^\pi)$  back to  $V^\pi$ . Red arrows point from the terms  $R$  and  $\gamma P V^\pi$  to their respective positions in the equation, indicating the components of the update rule.

## Iterative Policy Evaluation:

$$\begin{aligned} r(s,a) &\in [0, 1] \\ 1 + \gamma + \gamma^2 + \gamma^3 + \dots &= \frac{1}{1-\gamma} \end{aligned}$$

Algorithm (Iterative PE):

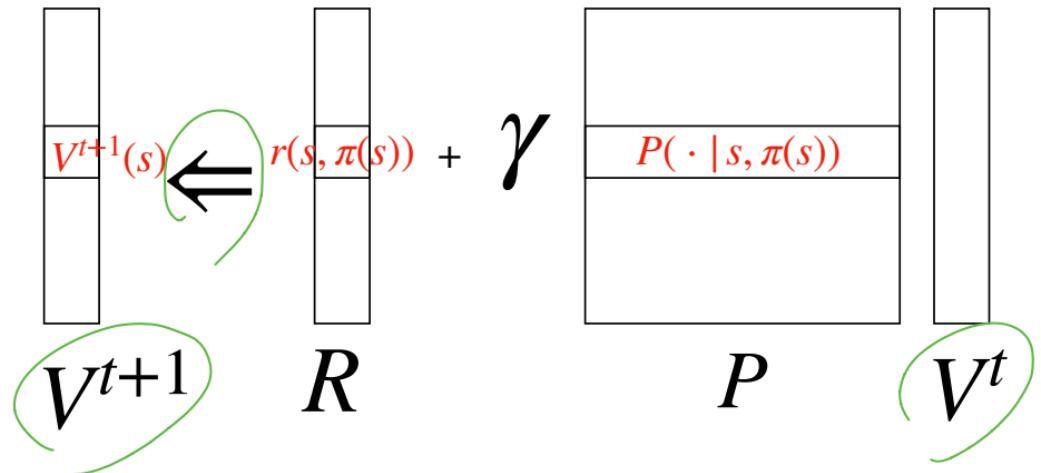
Start with some initialization  $V^0 \in [0, 1/(1-\gamma)]^{|S|}$ , repeat for  $t = 0 \dots$ :

$$V^{t+1} \leftarrow R + \gamma P V^t$$

# Iterative Policy Evaluation:

Algorithm (Iterative PE):

Start with some initialization  $V^0 \in [0, 1/(1 - \gamma)]^{|S|}$ , repeat for  $t = 0 \dots$ :

$$V^{t+1} \leftarrow R + \gamma P V^t$$


$$V^\pi \in \mathbb{R}^{|S|}$$

## Iterative Policy Evaluation:

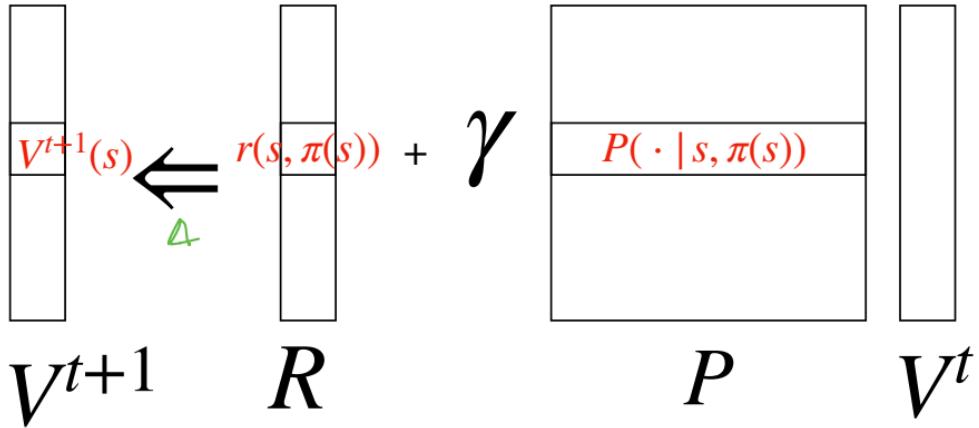
$$x \in [a, b]^d$$

$$x_i \in [a_i, b_i], \forall i \in [d]$$

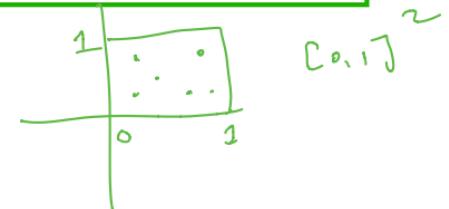
**Algorithm (Iterative PE):**

Start with some initialization  $V^0 \in [0, 1/(1 - \gamma)]^{|S|}$ , repeat for  $t = 0 \dots$ :

$$V^{t+1} \leftarrow R + \gamma P V^t$$



Q: What's computation complexity per iteration?



$O(|S|^2)$   
per-Iteration

# Iterative Policy Evaluation:

$$V^{t+1} \leftarrow R + \gamma P V^t$$

$$V^{t+1}(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^t(s')$$
$$V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^\pi(s')$$

$\gamma \varepsilon$

$\geq 0$

$\gamma \varepsilon$

$\varepsilon$

# Convergence of Iterative PE

**Theorem:**

Recall  $\gamma \in [0,1)$ . After  $t$  iterations, we have:

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

$$\begin{aligned} & \|V\|_\infty \\ &= \max_{i \in \text{cd}} |V_i| \end{aligned}$$

# Convergence of Iterative PE

## Theorem:

Recall  $\gamma \in [0,1)$ . After  $t$  iterations, we have:

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

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

# Convergence of Iterative PE

## Theorem:

Recall  $\gamma \in [0,1)$ . After  $t$  iterations, we have:

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

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

By Alg      ↴ BE on  $V^\pi$

# Convergence of Iterative PE

## Theorem:

Recall  $\gamma \in [0,1)$ . After  $t$  iterations, we have:

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

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

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

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

# Convergence of Iterative PE

## Theorem:

Recall  $\gamma \in [0,1)$ . After  $t$  iterations, we have:

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

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

# Convergence of Iterative PE

## Theorem:

Recall  $\gamma \in [0,1)$ . After  $t$  iterations, we have:

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

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

# Convergence of Iterative PE

## Theorem:

Recall  $\gamma \in [0,1)$ . After  $t$  iterations, we have:

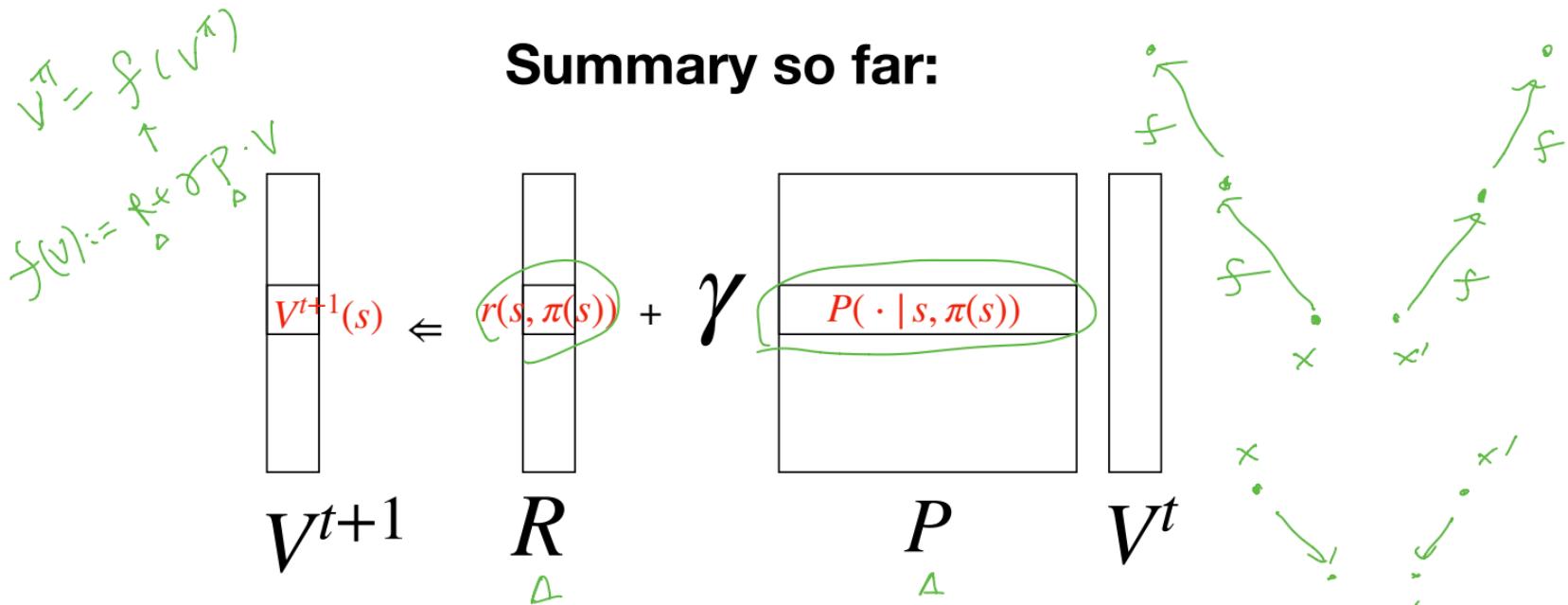
$$\forall s, |V^t(s) - V^\pi(s)| \leq \gamma^t \|V^0 - V^\pi\|_\infty$$

$$\begin{aligned}
 & \forall s, |V^{t+1}(s) - V^\pi(s)| \\
 & \quad \stackrel{\text{def}}{=} |V^{t+1}(s) - V^\pi(s)| \leq \|V^{t+1} - V^\pi\|_\infty \\
 & \quad \Rightarrow \|V^{t+1} - V^\pi\|_\infty \leq \gamma \|V^t - V^\pi\|_\infty \\
 & = \left| r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^t(s') - \left( r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^\pi(s') \right) \right| \\
 & = \gamma \left| \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^t(s') - \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^\pi(s') \right| \\
 & \leq \gamma \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} |V^t(s') - V^\pi(s')| \\
 & \leq \gamma \|V^t - V^\pi\|_\infty \quad \Rightarrow \|V^{t+1} - V^\pi\|_\infty \leq \gamma \|V^t - V^\pi\|_\infty \in \dots \gamma^{t+1} \|V_0 - V^\pi\|_\infty
 \end{aligned}$$

$\Downarrow$

$\gamma < 1$

## Summary so far:



Convergence:

$$\| V^{t+1} - V^\pi \|_\infty \leq \gamma \| V^t - V^\pi \|_\infty \leq \gamma^{t+1} \| V^0 - V^\pi \|_\infty$$

## **Outline:**

 1. Exact Policy Evaluation

 2. Approximate Policy Evaluation via an Iterative Algorithm

# Summary

Deterministic

Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?

# Summary

**Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?**

1. The exact algorithm  $V = (I - \gamma P)^{-1}R$  requires matrix inverse  $O(S^3)$

# Summary

**Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?**

1. The exact algorithm  $V = (I - \gamma P)^{-1}R$  requires matrix inverse  $O(S^3)$
1. For iterative PE algorithm, to find a  $\epsilon$  accurate value function, we need # of iterations:

# Summary

**Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?**

1. The exact algorithm  $V = (I - \gamma P)^{-1}R$  requires matrix inverse  $O(S^3)$
1. For iterative PE algorithm, to find a  $\epsilon$  accurate value function, we need # of iterations:

$$\ln\left(\frac{\|V^0 - V^*\|_\infty}{\epsilon}\right) / \ln(1/\gamma)$$

$\|V^t - V^*\|_\infty \leq \gamma^t \|V^0 - V^*\|_\infty$   
 $\leq \epsilon$   
 $\Rightarrow \gamma^t \leq \frac{\|V^0 - V^*\|_\infty}{\epsilon}$

# Summary

Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?

1. The exact algorithm  $V = (I - \gamma P)^{-1}R$  requires matrix inverse  $O(S^3)$
1. For iterative PE algorithm, to find a  $\epsilon$  accurate value function, we need # of iterations:

$$\ln\left(\frac{\|V^0 - V^*\|_\infty}{\epsilon}\right) / \ln(1/\gamma)$$

versus  $S^3$

Computation wise, we need  $O\left(S^2 \ln\left(\frac{1}{\epsilon}\right)\right)$

$\epsilon = 0.01$

# Summary

Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?

Bellman Equation



A fix-point equation:

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

# Summary

Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?

Bellman Equation



A fix-point equation:

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

Fix-point iteration  
framework

# Summary

Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?

Bellman Equation

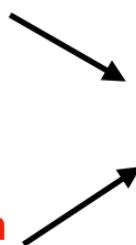


A fix-point equation:

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

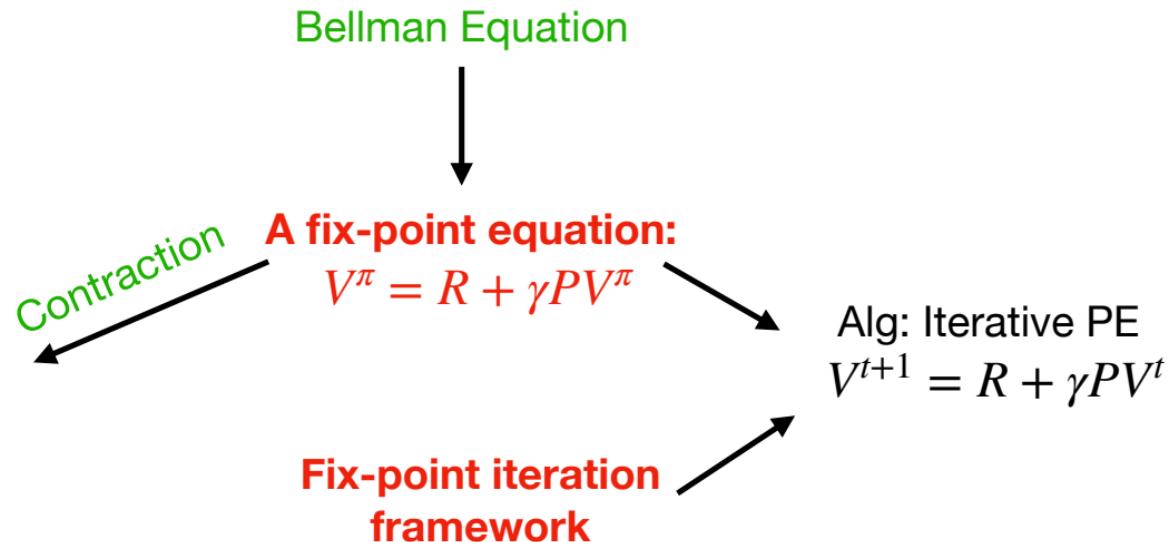
Alg: Iterative PE  
 $V^{t+1} = R + \gamma PV^t$

Fix-point iteration  
framework



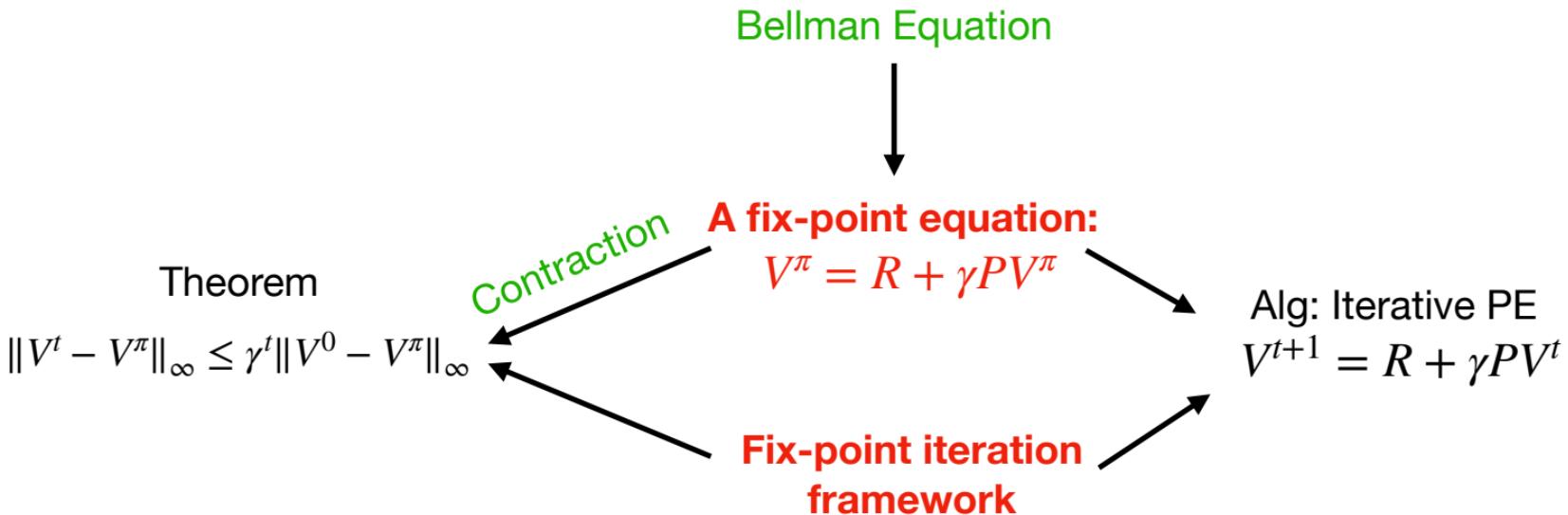
# Summary

Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?



# Summary

Key Question today: Given MDP  $\mathcal{M}$ , and a policy  $\pi$ , How to compute  $V^\pi(s), \forall s$ ?



$$\pi^* \\ V^* = r(s, \underline{\pi^*(s)}) + \gamma \mathbb{E}_{\substack{s' \sim P(\cdot | s, \underline{\pi^*(s)})}} V^*(s')$$

$$f: |f(x) - f(x')| \leq \gamma |x' - x|$$

$$V^* = \| R + \gamma PV - (R + \gamma PV') \|_\infty \\ \leq \gamma \| V - V' \|_\infty$$

## Next two lectures:

Given MDP  $\mathcal{M}$ , how to compute the optimal policy  $\pi^*$ , and  $V^*$

$$V^t \rightarrow V^* \\ \pi^* \Leftarrow \arg \max_{\pi} \left( \underset{\Delta}{\mathbb{E}} \left[ r(s, a) + \gamma \mathbb{E}_{\substack{s' \sim P(s'|s, a)}} V^t(s') \right] \right)$$