Policy Evaluation
Announcements

HW0 is out today (due March 2rd, 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

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

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

Policy \( \pi : S \mapsto A \)
Recap: Definitions

\[ M = \{ S, A, P, r, \gamma \} \]

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

Policy \( \pi : S \mapsto A \)

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

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

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

Policy \( \pi : S \mapsto A \)

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

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

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

\[ \mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}(s_0, a_0, \ldots, s_{h-1}, a_{h-1}; s_h = s, a_h = a) \]
Recap: State-action distribution

\[ P_h^\pi(s, a; s_0) = \sum_{a_0, a_1, \ldots, a_{h-1}, s_{h-1}} P_\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a) \]
Recap: State-action distribution

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

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

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

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

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

This gives us $S$ many linear constraints
Exact Policy Evaluation

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

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

\[
V(s_1) = r(s_1, \pi(s_1)) + \sum_{s' \notin S} P(s' | s_1, \pi(s_1)) V(s')
\]

\[
V(s_2) = r(s_2, \pi(s_2)) + \sum_{s' \notin S} P(s' | s_2, \pi(s_2)) V(s')
\]
Exact Policy Evaluation

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

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

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:

$$
V = R + \gamma PV
$$

*Correction:*

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

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

\[ V = R + \gamma PV \]
Exact Policy Evaluation

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

$$V = (I - \gamma P)^{-1}R$$
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)
Summary so far:

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

\[
V = (I - \gamma P)^{-1}R
\]
Summary so far:

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

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

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

1. Exact Policy Evaluation

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 \( x^* = f(x^*) \), \( f : [a, b] \rightarrow [a, b] \)

Common approach to find \( x^* \):

Initialize \( x^0 \in [a, b] \), repeat: \( x^{t+1} = f(x^t) \)
Detour: fix-point solution

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

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:

\( x^t \to x^* \), as \( t \to \infty \)
$V^\pi$ is a fix-point solution:

\[
\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^\pi(s')
\]
$V^\pi$ is a fix-point solution:

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

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

$V^{t+1} \leftarrow R + \gamma PV^t$
Iterative Policy Evaluation:

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

$$V^{t+1} \leftarrow R + \gamma PV^t$$
Iterative Policy Evaluation:

**Algorithm (Iterative PE):**

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

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

$V^t$ is the value function at iteration $t$.

**Q:** What’s computation complexity per iteration?

$O(S^2)$ per iteration.
Iterative Policy Evaluation:

\[ V^{t+1} \leftarrow R + \gamma PV^t \]

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

\[ V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^\pi(s') \]
Theorem:

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

$$\forall s, \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty$$
Convergence of Iterative PE

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

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

$$\forall s, \left| V^{t+1}(s) - V^\pi(s) \right|$$
Convergence of Iterative PE

**Theorem:**

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

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

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

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

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

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

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

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

$$= \gamma \left| \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right|$$
Theorem:
Recall $\gamma \in [0,1)$. After $t$ iterations, we have:
$$\forall s, \quad \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_{\infty}$$
Convergence of Iterative PE

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

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

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

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

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

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

$$\leq \gamma \left\| V^t - V^\pi \right\|_\infty$$
Convergence of Iterative PE

**Theorem:**

Recall \( \gamma \in [0,1) \). After \( t \) iterations, we have:
\[
\forall s, \quad \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty
\]

\[
\forall s, \quad \left| V^{t+1}(s) - V^\pi(s) \right| = \gamma \left[ r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(.|s, \pi(s))} V^t(s') - \left( r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(.|s, \pi(s))} V^\pi(s') \right) \right]
\]

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

\[
\leq \gamma \left| V^t(s') - V^\pi(s') \right| \leq \gamma \left\| V^t - V^\pi \right\|_\infty
\]

\[
\Rightarrow \left\| V^{t+1} - V^\pi \right\|_\infty \leq \gamma \left\| V^t - V^\pi \right\|_\infty
\]

\[
\Rightarrow \quad \forall s, \quad \left| V^{t+1}(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty
\]

\[
\Rightarrow \quad \exists T \text{ such that } \gamma^T \left\| V^0 - V^\pi \right\|_\infty = 0
\]

\[
\Rightarrow \quad V^t(s) \rightarrow V^\pi(s) \quad \text{as} \quad t \rightarrow \infty
\]
Summary so far:

\[
V^{t+1}(s) \leftarrow r(s, \pi(s)) + \gamma P(\cdot | s, \pi(s))
\]

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
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:

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

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

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

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

Bellman Equation

A fix-point equation:

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

Contraction

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 $V^\pi(s), \forall s$?

Bellman Equation

A fix-point equation:

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

Theorem

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

Contraction

Alg: Iterative PE

$$V^{t+1} = R + \gamma PV^t$$

Fix-point iteration framework
Next two lectures:

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

$$V^* = \arg \max_{\pi} \left( r(s, a) + \gamma \mathbb{E}_{s' \sim P(s|s, a)} V^*(s') \right)$$