

Lecture 15: Batch RL

Emma Brunskill

CS234 Reinforcement Learning.

Winter 2019

Slides drawn from Philip Thomas with modifications

Class Structure

- Last time: Meta Reinforcement Learning
- **This time: Batch RL**
- Next time: Quiz

A Scientific Experiment

A Group

$$\frac{3}{6} + \frac{2}{8} = \frac{10}{24} = \frac{5}{12}$$

2 Finally, reduce the sum to lowest terms:

$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$



1 Compare these fractions using the cross-multiplication strategy.

$$\frac{4}{5} \quad \frac{9}{10}$$

$4 \times 10 = 40$ $9 \times 5 = 45$



40 < 45

Avg Score: 95

A Scientific Experiment

how large is
NA NB

who is in
these groups

A Group	<p>1 $\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}$</p> <p>2 Finally, reduce the sum to lowest terms:</p> $\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$	<p>1 Compare these fractions using the cross-multiplication strategy.</p>  <p>4 × 10 = 40 9 × 5 = 45 40 < 45</p>	Avg Score: 95
B Group	<p>1 Compare these fractions using the cross-multiplication strategy.</p>  <p>4 × 10 = 40 9 × 5 = 45 40 < 45</p>	<p>2 Finally, reduce the sum to lowest terms:</p> $\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$	Avg Score: 92

Var

Var
statistics

What to do for a new student?

What Should We Do For a New Student?

A Group

$$\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}$$

2 Finally, reduce the sum to lowest terms:

$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$



1 Compare these fractions using the cross-multiplication strategy.

$\frac{4}{5}$ $\frac{9}{10}$

$4 \times 10 = 40$ $9 \times 5 = 45$

$40 < 45$

Avg Score: 95

B Group

1 Compare these fractions using the cross-multiplication strategy.

$\frac{4}{5}$ $\frac{9}{10}$

$4 \times 10 = 40$ $9 \times 5 = 45$

$40 < 45$



$$\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}$$

2 Finally, reduce the sum to lowest terms:

$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$

Avg Score: 92

Involves Counterfactual Reasoning

A Group

$$\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}$$

2 Finally, reduce the sum to lowest terms:

$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$


1 Compare these fractions using the cross-multiplication strategy:

$$\frac{4}{5} \text{ vs } \frac{9}{10}$$

$4 \times 10 = 40$ $9 \times 5 = 45$

40 < 45

Avg Score: 95

B Group

1 Compare these fractions using the cross-multiplication strategy:

$$\frac{4}{5} \text{ vs } \frac{9}{10}$$

$4 \times 10 = 40$ $9 \times 5 = 45$

40 < 45



$$\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}$$

2 Finally, reduce the sum to lowest terms:

$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$

Avg Score: 92

B Group

$$\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}$$

2 Finally, reduce the sum to lowest terms:

$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$


1 Compare these fractions using the cross-multiplication strategy:

$$\frac{4}{5} \text{ vs } \frac{9}{10}$$

$4 \times 10 = 40$ $9 \times 5 = 45$

40 < 45

???

Involves Generalization

2 actions

$r=0$

$r=0$

reward
 $r = \text{test score}$

A Group

1

$$\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}$$

2 Finally, reduce the sum to lowest terms:

$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$


1 Compare these fractions using the cross-multiplication strategy.

$$\frac{4}{5} \text{ vs } \frac{9}{10}$$

$4 \times 10 = 40$ $9 \times 5 = 45$

$40 < 45$

Avg Score: 95

B Group

1 Compare these fractions using the cross-multiplication strategy.

$$\frac{4}{5} \text{ vs } \frac{9}{10}$$

$4 \times 10 = 40$ $9 \times 5 = 45$

$40 < 45$



2 Finally, reduce the sum to lowest terms:

$$\frac{2}{6} + \frac{3}{8} = \frac{18}{24} = \frac{3}{4}$$
$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$

Avg Score: 92

B Group

2 Finally, reduce the sum to lowest terms:

$$\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$$


1 Compare these fractions using the cross-multiplication strategy.

$$\frac{4}{5} \text{ vs } \frac{9}{10}$$

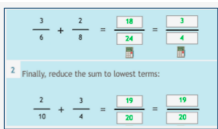
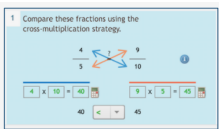
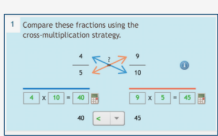
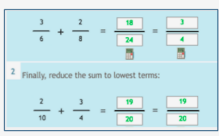
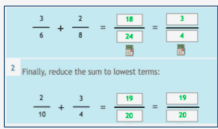
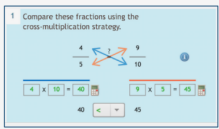
$4 \times 10 = 40$ $9 \times 5 = 45$

$40 < 45$

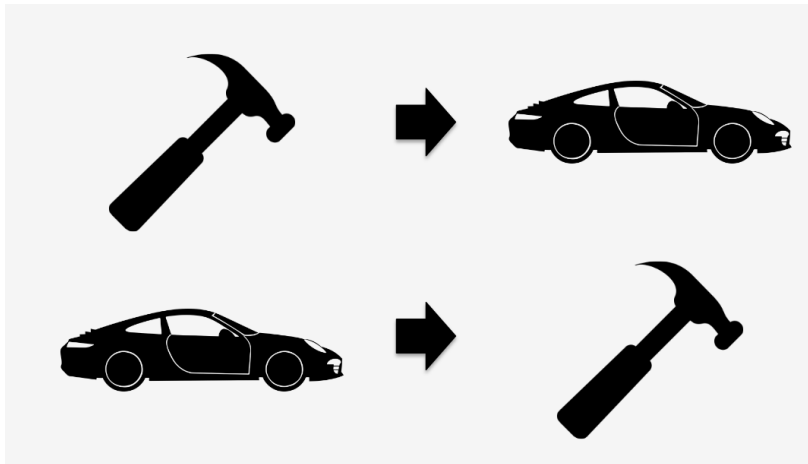
???

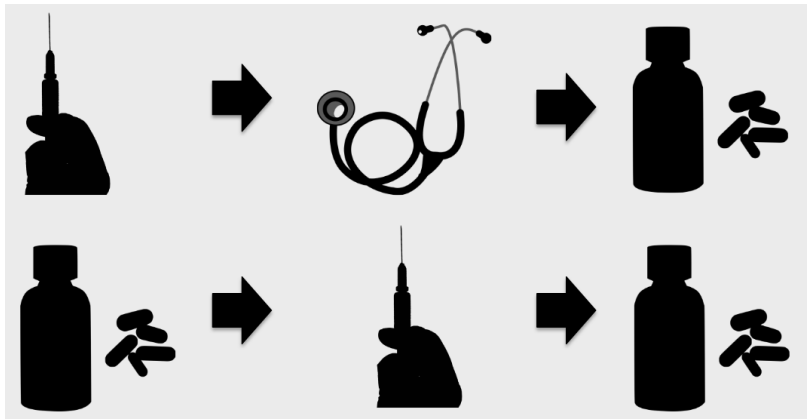
Batch Reinforcement Learning

Off Policy Offline Batch RL

A Group			Avg Score: 95
B Group			Avg Score: 92
B Group			???

Batch RL





The Problem

- If you apply an existing method, do you have confidence that it will work?

A property of many real applications

- Deploying "bad" policies can be costly or dangerous

What property should a safe batch reinforcement learning algorithm have?

- Given past experience from current policy/policies, produce a new policy
 - “Guarantee that with probability at least $1 - \delta$, will not change your policy to one that is worse than the current policy.”
 - You get to choose δ
 - Guarantee not contingent on the tuning of any hyperparameters

1 Notation

2 Create a safe batch reinforcement learning algorithm

- Off-policy policy evaluation (OPE)
- High-confidence off-policy policy evaluation (HCOPE)
- Safe policy improvement (SPI)

- Policy π : $\pi(a) = P(a_t = a \mid s_t = s)$
- Trajectory: $T = (s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_L, a_L, r_L)$
- Historical data: $D = \{T_1, T_2, \dots, T_n\}$
- Historical data from behavior policy, π_b $\begin{matrix} s \rightarrow a \\ h \rightarrow a \end{matrix}$
- Objective:

*generated
by doctors
 π_b ?*

$$V^\pi = \mathbb{E}\left[\sum_{t=1}^L \gamma^t R_t \mid \pi\right]$$

Safe batch reinforcement learning algorithm



- Reinforcement learning algorithm, \mathcal{A}
- Historical data, D , which is a random variable
- Policy produced by the algorithm, $\mathcal{A}(D)$, which is a random variable
- a safe batch reinforcement learning algorithm, \mathcal{A} , satisfies:

estimate $V^{\pi_b} = \frac{1}{n} \sum_{i=1}^n G_i$ } return

$$\Pr(\underbrace{V^{\mathcal{A}(D)}}_{\substack{\text{value of policy} \\ \text{used to generate data}}} \geq \underbrace{V^{\pi_b}}_{\substack{\text{value of policy} \\ \text{used to generate data}}}) \geq 1 - \delta$$

$\mathcal{A}(D) \rightarrow \pi$

or, in general

$$\Pr(\underbrace{V^{\mathcal{A}(D)}}_{\substack{\text{value of policy} \\ \text{used to generate data}}} \geq \underbrace{V_{\min}}_{(0,1)}) \geq 1 - \delta$$

1 Notation

2 Create a safe batch reinforcement learning algorithm

- Off-policy policy evaluation (OPE)
- High-confidence off-policy policy evaluation (HCOPE)
- Safe policy improvement (SPI)

Create a safe batch reinforcement learning algorithm

- Off-policy policy evaluation (OPE) $V^{A(D)}$ V^{π_e}
 - For any evaluation policy, π_e , Convert historical data, D , into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
 - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1 - \delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
 - Use HCOPE method to create a safe batch reinforcement learning algorithm,

$$\operatorname{argmax}_{\pi} V^{\pi_e} \quad \text{w/ some confidence bounds}$$

Off-policy policy evaluation (OPE)



Importance Sampling

$$IS(D) = \frac{1}{n} \sum_{i=1}^n \left(\prod_{t=1}^L \frac{\pi_e(a_t | s_t)}{\pi_b(a_t | s_t)} \right) \left(\sum_{t=1}^L \gamma^t R_t^i \right)$$

$$\mathbb{E}[IS(D)] = V^{\pi_e}$$

Create a safe batch reinforcement learning algorithm

- Off-policy policy evaluation (OPE)
 - For any evaluation policy, π_e , Convert historical data, D , into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
 - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1 - \delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
 - Use HCOPE method to create a safe batch reinforcement learning algorithm

High-confidence off-policy policy evaluation (HCOPE)

think back to exploration



Hoeffding's inequality

- Let X_1, \dots, X_n be n independent identically distributed random variables such that $X_i \in [0, b]$
- Then with probability at least $1 - \delta$:

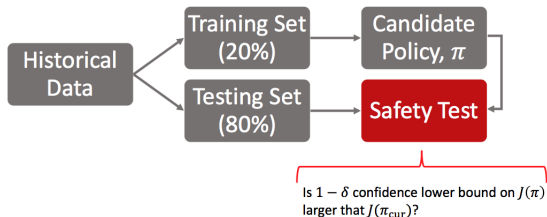
$$\mathbb{E}[X_i] \geq \frac{1}{n} \sum_{i=1}^n X_i - b \sqrt{\frac{\ln(1/\delta)}{2n}},$$

where $X_i = \frac{1}{n} \sum_{i=1}^n (w_i \sum_{t=1}^L \gamma^t R_t^i)$ in our case.

Safe policy improvement (SPI)



Safe policy improvement (SPI)



Create a safe batch reinforcement learning algorithm

- Off-policy policy evaluation (OPE)
 - For any evaluation policy, π_e , Convert historical data, D , into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
 - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1 - \delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
 - Use HCOPE method to create a safe batch reinforcement learning algorithm

Monte Carlo (MC) Off Policy Evaluation

Q learning off policy
samples & bootstraps
can be biased

- Aim: estimate value of policy π_1 , $V^{\pi_1}(s)$, given episodes generated under behavior policy π_2

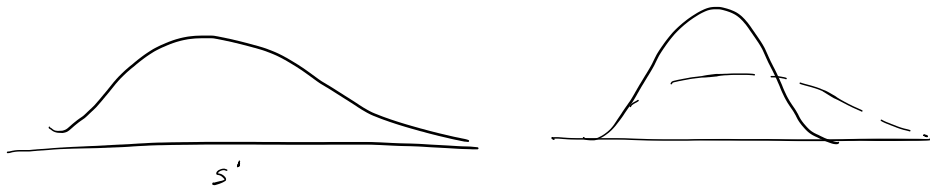
- D_{return}*
- $s_1, a_1, r_1, s_2, a_2, r_2, \dots$ where the actions are sampled from π_2
 - $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots$ in MDP M under policy π
 - $V^\pi(s) = \mathbb{E}_\pi[G_t | s_t = s]$
 - Have data from a different policy, behavior policy π_2
 - If π_2 is stochastic, can often use it to estimate the value of an alternate policy (formal conditions to follow)
 - Again, no requirement that have a model nor that state is Markov

Monte Carlo (MC) Off Policy Evaluation: Distribution Mismatch

$T = s \ a \ r \ s' \ a' \ r' \ s'' \ \dots$

- Distribution of episodes & resulting returns differs between policies

distrib γ $\pi_b \neq$ distrib γ from π_c



Importance Sampling

- Goal: estimate the expected value of a function $f(x)$ under some probability distribution $p(x)$, $\mathbb{E}_{x \sim p}[f(x)]$
- Have data x_1, x_2, \dots, x_n sampled from distribution $q(s)$
- Under a few assumptions, we can use samples to obtain an unbiased estimate of $\mathbb{E}_{x \sim q}[f(x)]$

$$\begin{aligned}\mathbb{E}_{x \sim q}[f(x)] &= \int_x q(x) f(x) dx \\ &= \int_x \frac{p(x)}{p(x)} q(x) f(x) dx \\ &= \int_x p(x) \left[\frac{q(x)}{p(x)} f(x) \right] dx \\ &\approx \frac{1}{n} \sum_i \frac{q(x_i)}{p(x_i)} f(x_i)\end{aligned}$$

when $x_i \sim p(x)$

Importance Sampling (IS) for Policy Evaluation

$$\frac{p(h_j | \pi_e)}{p(h_j | \pi_b)}$$

↑ h_1, G_1

traj, Σ rewards

↓ h_2, G_2

⋮

Markov
could condition
full history

✓

- Let h_j be episode j (history) of states, actions and rewards

$$h_j = (s_{j,1}, a_{j,1}, r_{j,1}, s_{j,2}, a_{j,2}, r_{j,2}, \dots, s_{j,L_j}(\text{terminal}))$$

$$p(h_j | \pi_e) = p(s_{j,1}) \prod_{j=1}^{L_j-1} \underbrace{p(a_{j+1} | s_j)}_{\text{policy}} \underbrace{p(s_{j+1} | s_j, a_j)}_{\text{trans model}} \underbrace{p(r_{j+1} | a_j, s_j)}_{\text{reward model}}$$

$$\begin{aligned} \frac{p(h_j | \pi_e)}{p(h_j | \pi_b)} &= \frac{p(s_{j,1})}{p(s_{j,1})} \prod \frac{p(a_j | s_j)^{\pi_e}}{p(a_j | s_j)^{\pi_b}} \frac{p(s_{j+1} | s_j, a_j)}{p(s_{j+1} | s_j, a_j)} \frac{p(r_{j+1})}{p(r_{j+1})} \\ &= \prod \frac{\pi_e(a_j | s_j)}{\pi_b(a_j | s_j)} \end{aligned}$$

Importance Sampling (IS) for Policy Evaluation

- Let h_j be episode j (history) of states, actions and rewards

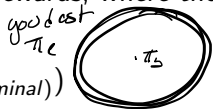
$$h_j = (s_{j,1}, a_{j,1}, r_{j,1}, s_{j,2}, a_{j,2}, r_{j,2}, \dots, s_{j,L_j}(\text{terminal}))$$

$$\begin{aligned} p(h_j | \pi, s = s_{j,1}) &= p(a_{j,1} | s_{j,1}) p(r_{j,1} | s_{j,1}, a_{j,1}) p(s_{j,2} | s_{j,1}, a_{j,1}) \\ &\quad p(a_{j,2} | s_{j,2}) p(r_{j,2} | s_{j,2}, a_{j,2}) p(s_{j,3} | s_{j,2}, a_{j,2}) \dots \\ &= \prod_{t=1}^{L_j-1} p(a_{j,t} | s_{j,t}) p(r_{j,t} | s_{j,t}, a_{j,t}) p(s_{j,t+1} | s_{j,t}, a_{j,t}) \\ &= \prod_{t=1}^{L_j-1} \pi(a_{j,t} | s_{j,t}) p(r_{j,t} | s_{j,t}, a_{j,t}) p(s_{j,t+1} | s_{j,t}, a_{j,t}) \end{aligned}$$

Importance Sampling (IS) for Policy Evaluation

- Let h_j be episode j (history) of states, actions and rewards, where the actions are sampled from π_2

$$h_j = (s_{j,1}, a_{j,1}, r_{j,1}, s_{j,2}, a_{j,2}, r_{j,2}, \dots, s_{j,L_j}(\text{terminal}))$$



$$V^{\pi_1}(s) \approx \sum_{j=1}^n \frac{p(h_j | \pi_1, s)}{p(h_j | \pi_2, s)} G(h_j)$$

$(r_1 + \gamma r_2 + \gamma^2 r_3 \dots)$

$$= \sum_{j=1}^n \prod_{i=1}^n \left(\frac{\pi_e(a_{ji} | s_{ji})}{\pi_b(a_{ji} | s_{ji})} \right) G(h_j)$$

no dynamics
" reward
" need to be Markov

$$\pi_b(a|s) = 0 \quad \text{but} \quad \pi_e(a|s) > 0$$

coverage or support

$$\pi_b(a|s) = 0 \quad \text{if} \quad \pi_e(a|s) = 0$$

$$\pi_b(a|s) > 0$$

$\forall a, s \text{ s.t. } \pi_e(a|s) > 0$

Importance Sampling for Policy Evaluation

1st used for RL Precup 2000?

- Aim: estimate $V^{\pi_1}(s)$ given episodes generated under policy π_2
 - $s_1, a_1, r_1, s_2, a_2, r_2, \dots$ where the actions are sampled from π_2
- Have access to $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots$ in MDP M under policy π_2
- Want $V^{\pi_1}(s) = \mathbb{E}_{\pi_1}[G_t | s_t = s]$
- IS = Monte Carlo estimate given off policy data
- Model-free method
- Does not require Markov assumption
- Under some assumptions, unbiased & consistent estimator of V^{π_1}
- Can be used when agent is interacting with environment to estimate value of policies different than agent's control policy

under a few assumptions
as $n \rightarrow \infty \overset{\pi_2}{\downarrow} \rightarrow V^{\pi_1}$

Leveraging Future Can't Influence Past Rewards

recall π gradient

- Importance sampling (IS):

$$G_T = r + r' + r'' + \dots$$

$$IS(D) = \frac{1}{n} \sum_{i=1}^n \left(\prod_{t=1}^L \frac{\pi_e(a_t | s_t)}{\pi_b(a_t | s_t)} \right) \left(\sum_{t=1}^L \gamma^t R_t^i \right)$$

- Per-decision importance sampling (PDIS)

$$PSID(D) = \sum_{t=1}^L \gamma^t \frac{1}{n} \sum_{i=1}^n \left(\prod_{\tau=1}^t \frac{\pi_e(a_\tau | s_\tau)}{\pi_b(a_\tau | s_\tau)} \right) R_t^i$$

only up to point got reward

Off-policy policy evaluation (revisited)

- Importance sampling (IS):

$$IS(D) = \frac{1}{n} \sum_{i=1}^n w_i \left(\sum_{t=1}^L \gamma^t R_t^i \right)$$

$\pi \frac{\pi_{\theta}(a|s)}{\pi_b(a|s)} \leftarrow$ might be super small

- Weighted importance sampling (WIS)

$$WIS(D) = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \left(\sum_{t=1}^L \gamma^t R_t^i \right)$$

biased
still consistent
lower variance

Off-policy policy evaluation (revisited)

- Weighted importance sampling (WIS)

$$WIS(D) = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \left(\sum_{t=1}^L \gamma^t R_t^i \right)$$

- Biased. When $n = 1$, $\mathbb{E}[WIS] = V(\pi_b)$
- Strongly consistent estimator of V^{π_e}
 - i.e. $\Pr(\lim_{n \rightarrow \infty} WIS(D) = V^{\pi_e}) = 1$
 - If
 - Finite horizon
 - One behavior policy, or bounded rewards

Thomas 3 Brunskill
16 ML 2016

Control variates

- Given: X
- Estimate: $\mu = \mathbb{E}[X]$
- $\hat{\mu} = X$
- Unbiased: $\mathbb{E}[\hat{\mu}] = \mathbb{E}[X] = \mu$
- Variance: $\text{Var}(\hat{\mu}) = \text{Var}(X)$

Control variates

- Given: $X, Y, \mathbb{E}[Y]$
- Estimate: $\mu = \mathbb{E}[X]$
- $\hat{\mu} = X - Y + \mathbb{E}[Y]$

Y Q func
 $\mathbb{E}[Y]$ V

- Unbiased:

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[X - Y + \mathbb{E}[Y]] = \mathbb{E}[X] - \mathbb{E}[Y] + \mathbb{E}[Y] = \mathbb{E}[X]$$

- Variance:

$$\text{Var}(\hat{\mu}) = \text{Var}(X - Y + \mathbb{E}[Y]) = \text{Var}(X - Y)$$

n

Control variates

- Given: $X, Y, \mathbb{E}[Y]$
- Estimate: $\mu = \mathbb{E}[X]$
- $\hat{\mu} = X - Y + \mathbb{E}[Y]$

- Unbiased:

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[X - Y + \mathbb{E}[Y]] = \mathbb{E}[X] - \mathbb{E}[Y] + \mathbb{E}[Y] = \mathbb{E}[X] = \mu$$

- Variance:

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \text{Var}(X - Y + \mathbb{E}[Y]) = \text{Var}(X - Y) \\ &= \underline{\text{Var}(X)} + \underline{\text{Var}(Y)} - \underline{2\text{Cov}(X, Y)} \end{aligned}$$

- Lower variance if $2\text{Cov}(X, Y) > \text{Var}(Y)$ if true $\text{Var}(\hat{\mu}) < \text{Var}(X)$
- We call Y a control variate
- We saw this idea before: baseline term in policy gradient estimation

Off-policy policy evaluation (revisited)

- Idea: add a control variate to importance sampling estimators
 - X is the importance sampling estimator
 - Y is a control variate build from an approximate model of the MDP
 Q some estimate of state action value

Off-policy policy evaluation (revisited)

- Idea: add a control variate to importance sampling estimators
 - X is the importance sampling estimator
 - Y is a control variate build from an approximate model of the MDP

2011 multi armed bandit (Dudik et al.)

- Called the doubly robust estimator (Jiang and Li, 2015)
 - Robust to (1) poor approximate model, and (2) error in estimates of π_b
 - control variate*
 - If the model is poor, the estimates are still unbiased
 - If the sampling policy is unknown, but the model is good, MSE will still be low
- Non-recursive and weighted forms, as well as control variate view provided by Thomas and Brunskill (ICML 2016)

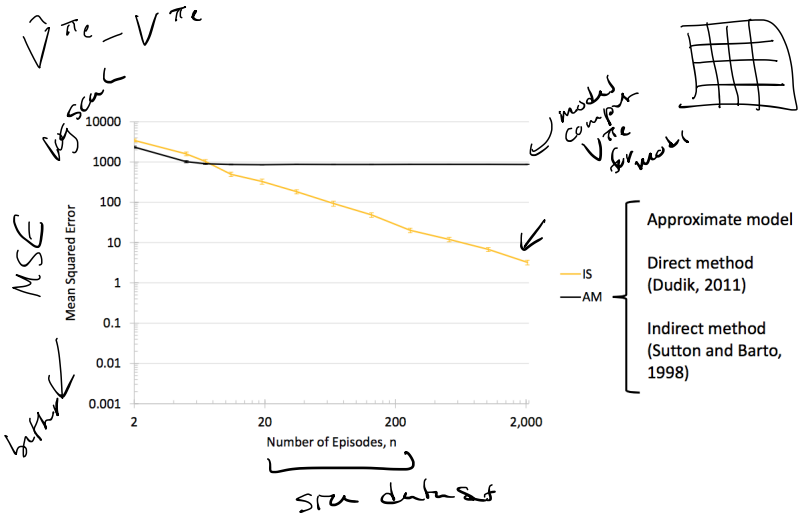
Off-policy policy evaluation (revisited)

$$DR(\pi_e | D) = \frac{1}{n} \sum_{i=1}^n \sum_{t=0}^{\infty} \gamma^t w_t^i (R_t^i - \hat{q}^{\pi_e}(S_t^i, A_t^i)) + \gamma^t \rho_{t-1}^i \hat{v}^{\pi_e}(S_t^i),$$

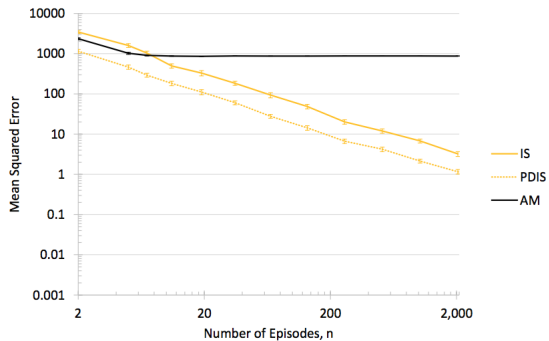
Handwritten annotations:
- An arrow points from "raw returns" to R_t^i .
- An arrow points from γ to the discount factor in the second term.
- $E[\gamma]$ is written to the right of the equation.
- An arrow points from "IS weights" to w_t^i .

$$\text{where } w_t^i = \prod_{\tau=1}^t \frac{\pi_e(a_\tau | s_\tau)}{\pi_b(a_\tau | s_\tau)}$$

Empirical Results (Gridworld)

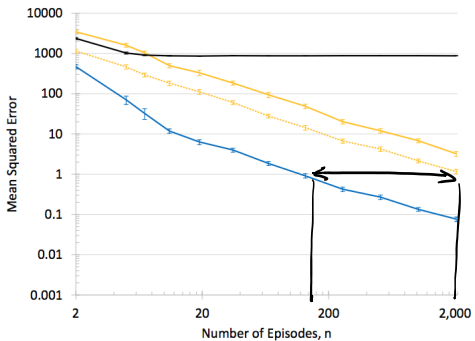


Empirical Results (Gridworld)



Empirical Results (Gridworld)

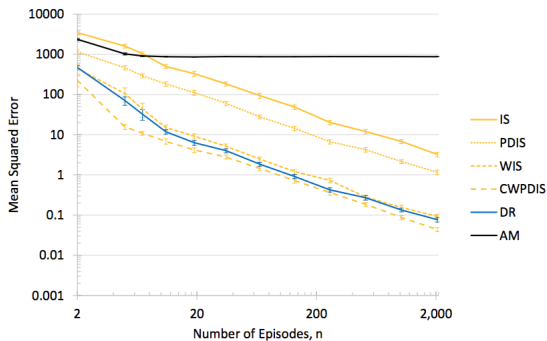
$$\hat{V}^{\pi_e} - V^{\pi_e}$$



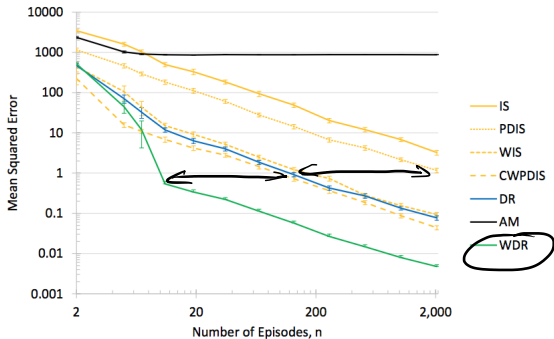
- IS
- PDIS
- DR
- AM

AM + IS

Empirical Results (Gridworld)



Empirical Results (Gridworld)

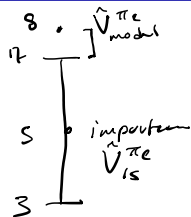


Off-policy policy evaluation (revisited): Blending

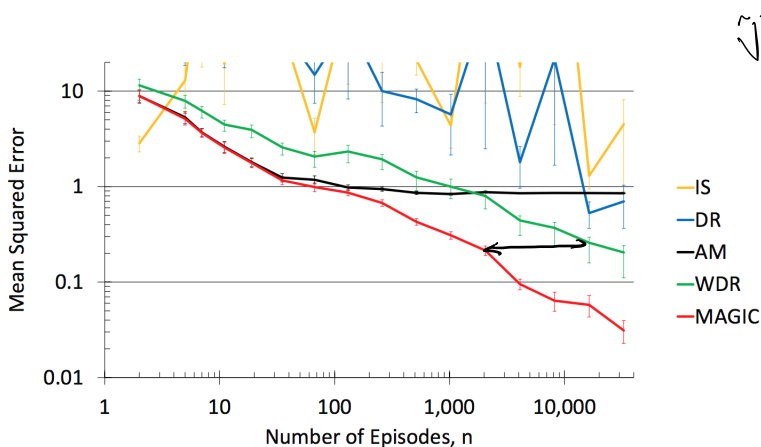
$$MSE = f(\text{bias}, \text{var})$$

$$\text{bias} : \hat{V}^{\pi_e} - V^{\pi_e}$$

- Importance sampling is unbiased but high variance
- Model based estimate is biased but low variance
- Doubly robust is one way to combine the two
- Can also trade between importance sampling and model based estimate within a trajectory
- MAGIC estimator (Thomas and Brunskill ICML 2016) *directly min MSE*
- Can be particularly useful when part of the world is non-Markovian in the given model, and other parts of the world are Markov

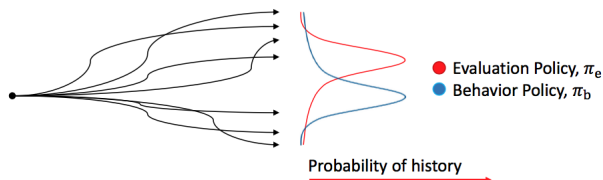


Can Need an Order of Magnitude Less Data To Get Good Estimates



Off-policy policy evaluation (revisited)

- What if $\text{supp}(\pi_e \subset \text{supp}(\pi_b))$
- There is a state-action pair, (s, a) , such that $\pi_e(a | s) = 0$, but $\pi_b(a | s) \neq 0$.
- If we see a history where (s, a) occurs, what weight should we give it?
- $$IS(D) = \frac{1}{n} \sum_{i=1}^n \left(\prod_{t=1}^L \frac{\pi_e(a_t | s_t)}{\pi_b(a_t | s_t)} \right) \left(\sum_{t=1}^L \gamma^t R_t^i \right)$$

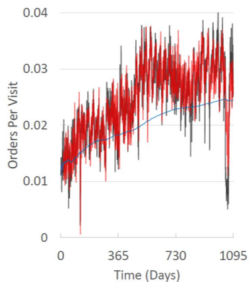
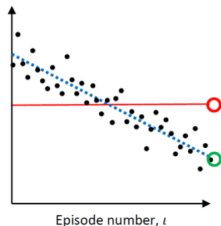


Off-policy policy evaluation (revisited)

- What if there are zero samples ($n = 0$)?
 - The importance sampling estimate is undefined
- What if no samples are in $\text{supp}(\pi_e)$ (or $\text{supp}(p)$ in general)?
 - Importance sampling says: the estimate is zero
 - Alternate approach: undefined
- Importance sampling estimator is unbiased if $n > 0$
- Alternate approach will be unbiased given that at least one sample is in the support of p
- Alternate approach detailed in Importance Sampling with Unequal Support (Thomas and Brunskill, AAI 2017)

Off-policy policy evaluation (revisited)

- Thomas et. al. Predictive Off-Policy Policy Evaluation for Nonstationary Decision Problems, with Applications to Digital Marketing (AAAI 2017)



Create a safe batch reinforcement learning algorithm

$$P(V^{A(D)} - V^{\pi_e} > 0) \geq 1 - \delta$$

- Off-policy policy evaluation (OPE)
 - For any evaluation policy, π_e , Convert historical data, D , into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
 - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1 - \delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
 - Use HCOPE method to create a safe batch reinforcement learning algorithm,

High-confidence off-policy policy evaluation (revisited)

- Consider using IS + Hoeffding's inequality for HCOPE on mountain car *exploration*

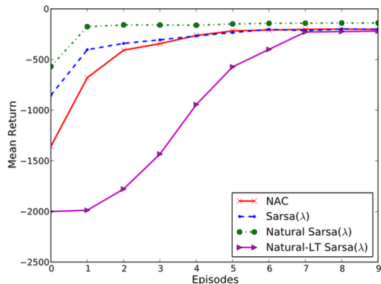
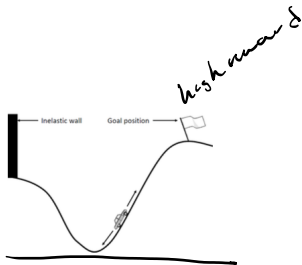


Figure 3: Mountain Car (Sarsa(λ))
Natural Temporal Difference Learning, Dabney and Thomas, 2014

Hoeffding's inequality

empirical mean - true mean

- Let X_1, \dots, X_n be n independent identically distributed random variables such that $X_i \in [0, b]$
- Then with probability at least $1 - \delta$:

$$b = \max \left(G, \pi_{i=c}^T \frac{\pi_c(a|s)}{\pi_b(a|s)} \right)$$

↑
(0,1)

$$\mathbb{E}[X_i] \geq \frac{1}{n} \sum_{i=1}^n X_i - b \sqrt{\frac{\ln(1/\delta)}{2n}}$$

where $X_i = \frac{1}{n} \sum_{i=1}^n (w_i \sum_{t=1}^L \gamma^t R_t^i)$ in our case.

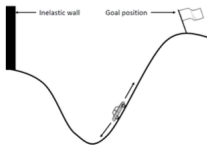
$$L = 200$$
$$\left(\frac{1}{.1}\right)^{200} \cdot \sqrt{\frac{1}{n}}$$

High-confidence off-policy policy evaluation (revisited)

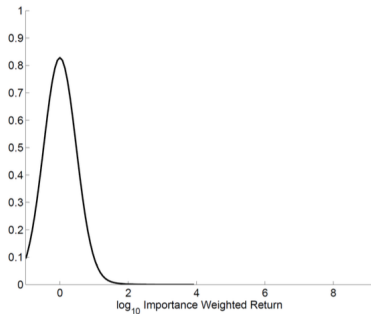
- Using 100,000 trajectories
- Evaluation policy's true performance is $0.19 \in [0, 1]$
- We get a 95% confidence lower bound of: $-5,8310,000$

✓ true V is between 0 & 1

$.19 > -5$ million



What went wrong



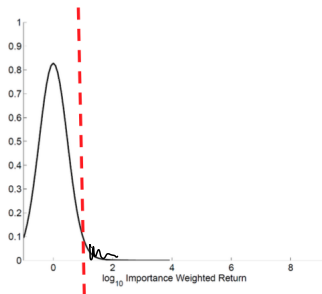
$$w_i = \prod_{t=1}^L \frac{\pi_e(a_t | s_t)}{\pi_b(a_t | s_t)}$$

pretty small

$$10^2 \left(\frac{1}{-1}\right)^L$$

High-confidence off-policy policy evaluation (revisited)

- Removing the upper tail only decreases the expected value.



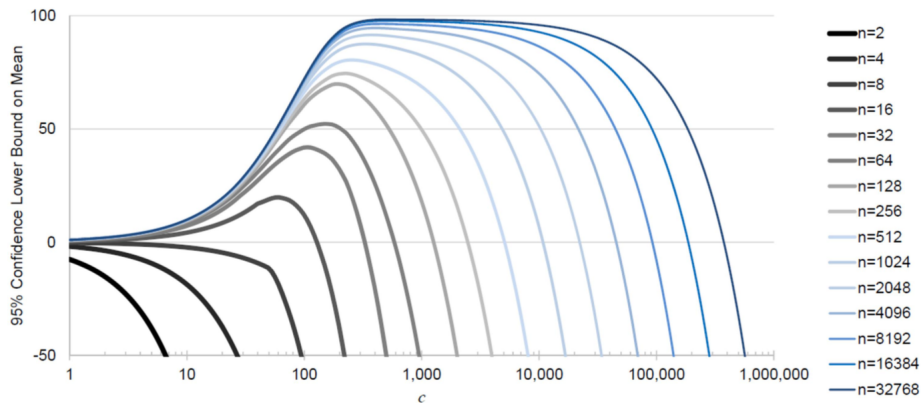
High-confidence off-policy policy evaluation (revisited)

- Thomas et. al, High confidence off-policy evaluation, AAAI 2015

Theorem 1. Let X_1, \dots, X_n be n independent real-valued random variables such that for each $i \in \{1, \dots, n\}$, we have $\mathbb{P}[0 \leq X_i] = 1$, $\mathbb{E}[X_i] \leq \mu$, and some threshold value $c_i > 0$. Let $\delta > 0$ and $Y_i := \min\{X_i, c_i\}$. Then with probability at least $1 - \delta$, we have

$$\mu \geq \underbrace{\left(\sum_{i=1}^n \frac{1}{c_i}\right)^{-1} \sum_{i=1}^n \frac{Y_i}{c_i}}_{\text{empirical mean}} - \underbrace{\left(\sum_{i=1}^n \frac{1}{c_i}\right)^{-1} \frac{7n \ln(2/\delta)}{3(n-1)}}_{\text{term that goes to zero as } 1/n \text{ as } n \rightarrow \infty} - \underbrace{\left(\sum_{i=1}^n \frac{1}{c_i}\right)^{-1} \sqrt{\frac{\ln(2/\delta)}{n-1} \sum_{i,j=1}^n \left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)^2}}_{\text{term that goes to zero as } 1/\sqrt{n} \text{ as } n \rightarrow \infty}. \quad (3)$$

High-confidence off-policy policy evaluation (revisited)



High-confidence off-policy policy evaluation (revisited)

- Use 20% of the data to optimize c (cutoff)
- Use 80% to compute lower bound with optimized c
- Mountain car results: 100k traj

confidence interval

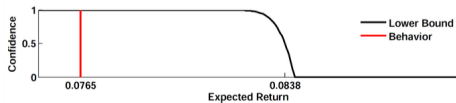
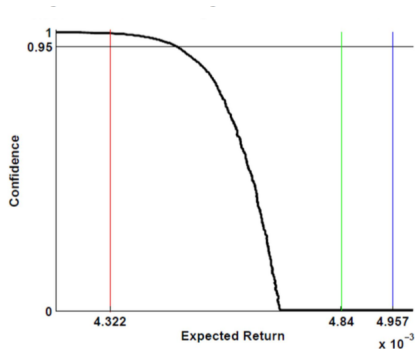
	CUT	Chernoff-Hoeffding	Maurer	Anderson	Bubeck et al.
95% Confidence lower bound on the mean	0.145	-5,831,000	-129,703	0.055	-.046

$\sqrt{\pi \epsilon}$

True
.19

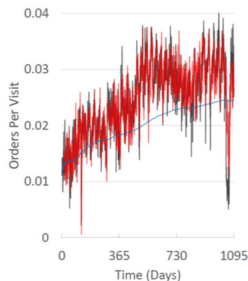
High-confidence off-policy policy evaluation (revisited)

Digital marketing:



High-confidence off-policy policy evaluation (revisited)

Cognitive dissonance:



$$\mathbb{E}[X_i] \geq \frac{1}{n} \sum_{i=1}^n X_i - b \sqrt{\frac{\ln(1/\delta)}{2n}}$$

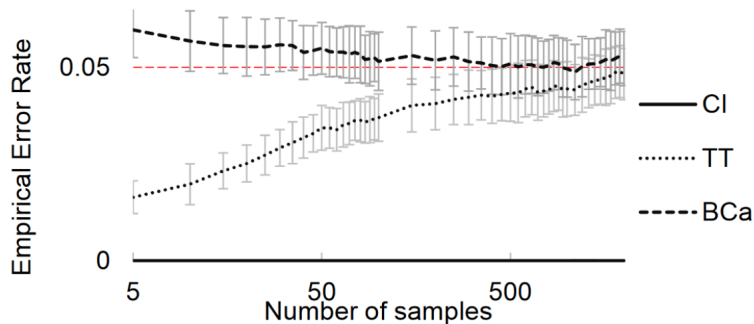
High-confidence off-policy policy evaluation (revisited)

- Student's t-test
 - Assumes that $IS(D)$ is normally distributed
 - By the central limit theorem, it (is as $n \rightarrow \infty$)

$$\Pr \left(\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] \geq \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}}{\sqrt{n}} t_{1-\delta, n-1}$$
$$\geq 1 - \delta$$

- Efron's Bootstrap methods (e.g., BCa)
 - Also, without importance sampling: Hanna, Stone, and Niekum, AAMAS 2017

High-confidence off-policy policy evaluation (revisited)



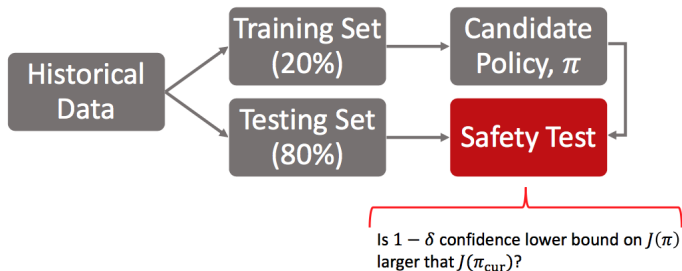
P. S. Thomas. Safe reinforcement learning (PhD Thesis, 2015)

Create a safe batch reinforcement learning algorithm

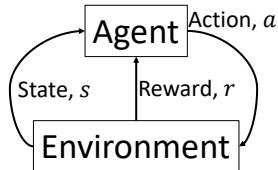
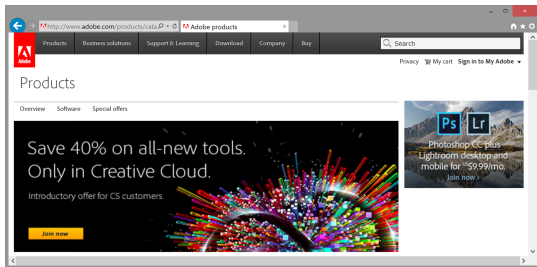
- Off-policy policy evaluation (OPE)
 - For any evaluation policy, π_e , Convert historical data, D , into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
 - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1 - \delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
 - Use HCOPE method to create a safe batch reinforcement learning algorithm

Safe policy improvement (revisited)

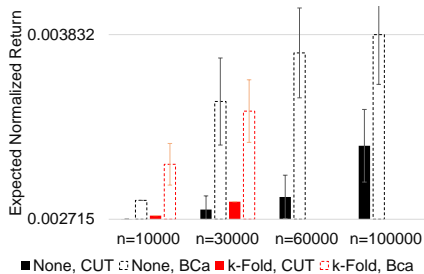
Thomas et. al, ICML 2015



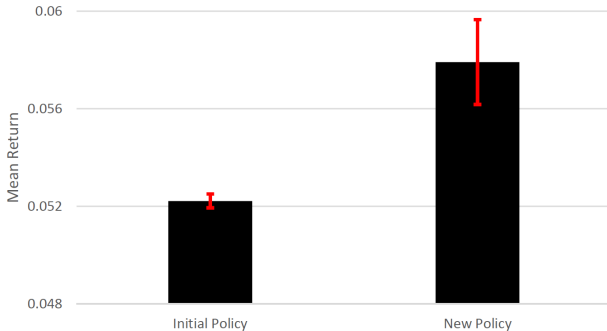
Empirical Results: Digital Marketing



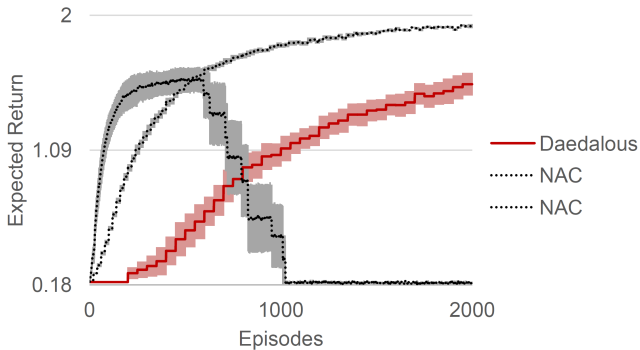
Empirical Results: Digital Marketing



Empirical Results: Digital Marketing



Empirical Results: Digital Marketing



Other Relevant Work

- How to deal with long horizons? (Guo, Thomas, Brunskill NIPS 2017)
- How to deal with importance sampling being “unfair”? (Doroudi, Thomas and Brunskill, best paper UAI 2017)
- What to do when the behavior policy is not known? (Liu, Gottesman, Raghu, Komorowski, Faisal, Doshi-Velez, Brunskill NeurIPS 2018)
- What to do when the behavior policy is deterministic?
- What to do when care about safe exploration?
- What to do when care about performance on a single trajectory
- For last two, see great work by Marco Pavone’s group, Pieter Abbeel’s group, Shie Mannor’s group and Claire Tomlin’s group, amongst others

Off Policy Policy Evaluation and Selection

- Very important topic: healthcare, education, marketing, ...
- Insights are relevant to on policy learning
- Big focus of my lab
- A number of others on campus also working in this area (e.g. Stefan Wager, Susan Athey...)
- Very interesting area at the intersection of causality and control

What You Should Know: Off Policy Policy Evaluation and Selection

$D \rightarrow \text{good } V^{\pi_e}$

- Be able to define and apply importance sampling for off policy policy evaluation
- Define some limitations of IS (variance)
- List a couple alternatives (weighted IS, doubly robust)
- Define why we might want safe reinforcement learning
- Define the scope of the guarantees implied by safe policy improvement as defined in this lecture

Class Structure

- Last time: Meta Reinforcement Learning
- **This time: Batch RL**
- Next time: Quiz

Off-policy policy evaluation (revisited)

- Weighted per-decision importance sampling
 - Also called consistent weighted per-decision importance sampling
 - A fun exercise!