

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

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

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

A Scientific Experiment

A Group

1 $\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}$

2 Finally, reduce the sum to lowest terms:
 $\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$



Avg Score: 95

B Group

1 Compare these fractions using the cross-multiplication strategy.
 $\frac{4}{5} > \frac{9}{10}$

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

2 Finally, reduce the sum to lowest terms:
 $\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}$



Avg Score: 92

What Should We Do For a New Student?

A Group

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

Finally, reduce the sum to lowest terms:

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



Avg Score: 95

B Group

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

Compare these fractions using the cross-multiplication strategy.

$$\frac{4}{5} \times \frac{10}{9} = \frac{40}{45} < \frac{9}{10}$$



Avg Score: 92

Involves Counterfactual Reasoning

A Group

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

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



1 Compare these fractions using the cross-multiplication strategy.

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

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.

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

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

40 < 45

???

Involves Generalization

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.

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

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

40 < 45

Avg Score: 95

B Group

1

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

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

40 < 45



1

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

2 Finally, reduce the sum to lowest terms:

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

Avg Score: 92

B 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

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

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

40 < 45

???

Batch Reinforcement Learning

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.

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

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

40 < 45

Avg Score: 95

B Group

1

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

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

40 < 45



1

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

2 Finally, reduce the sum to lowest terms:

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

Avg Score: 92

B 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

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

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

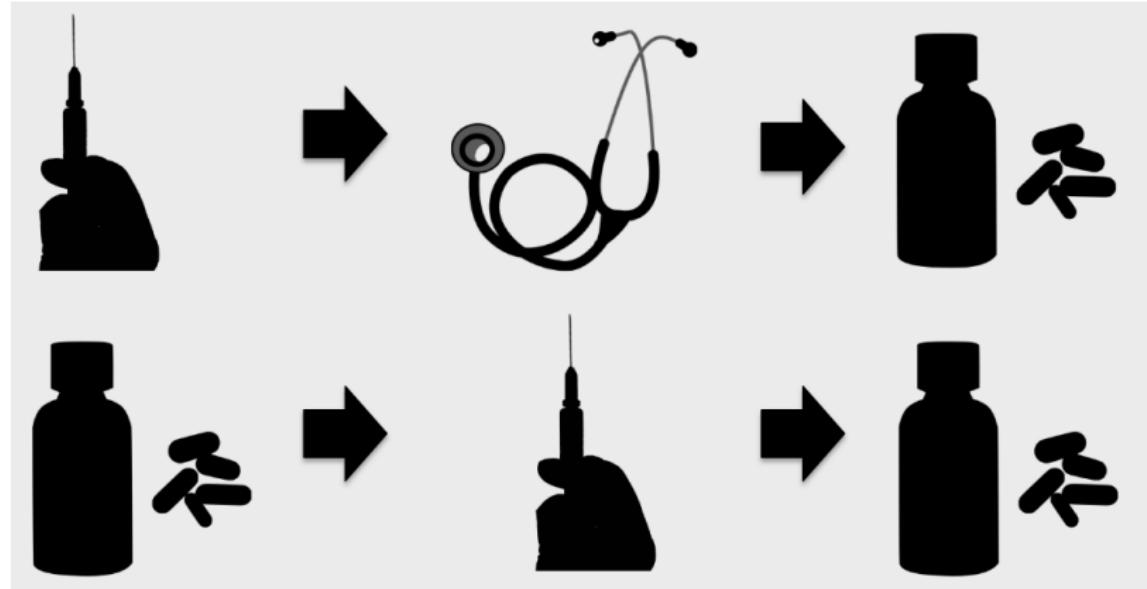
40 < 45

???

Batch RL



Batch RL



The Problem

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

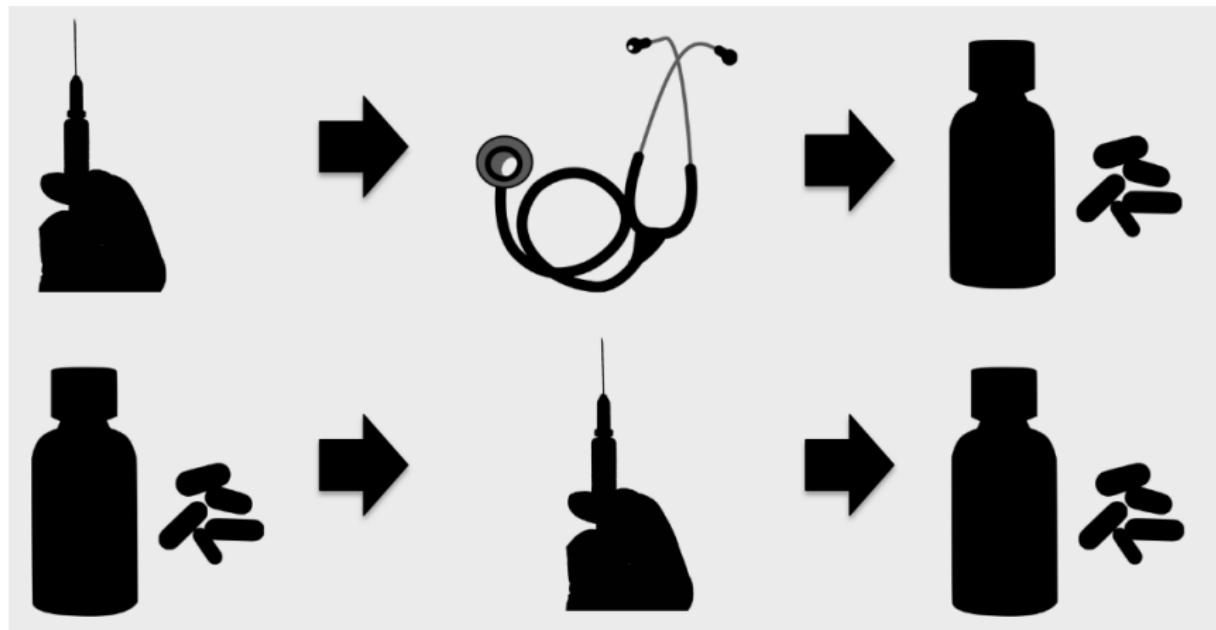
A property of many real applications

- Deploying "bad" policies can be costly or dangerous

Deploying bad policies can be costly



Deploying bad policies can be 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

Table of Contents

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)

Notation

- 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
- Objective:

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

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:

$$\Pr(V^{\mathcal{A}(\mathcal{D})} \geq V^{\pi_b}) \geq 1 - \delta$$

or, in general

$$\Pr(V^{\mathcal{A}(\mathcal{D})} \geq V_{min}) \geq 1 - \delta$$

Table of Contents

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)
 - 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,

Off-policy policy evaluation (OPE)

Off-policy policy evaluation (OPE)



Importance Sampling

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)



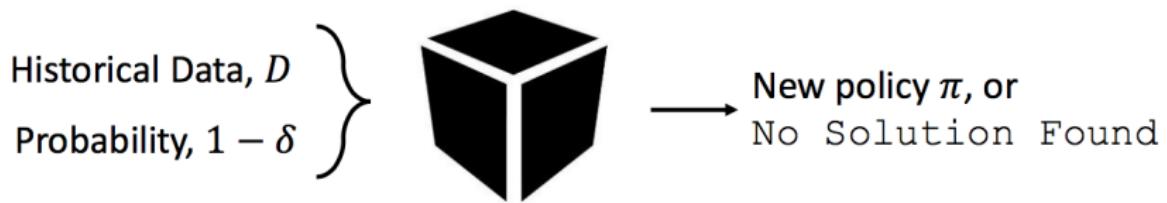
Hoeffding's inequality

- 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 $1 - \delta$:

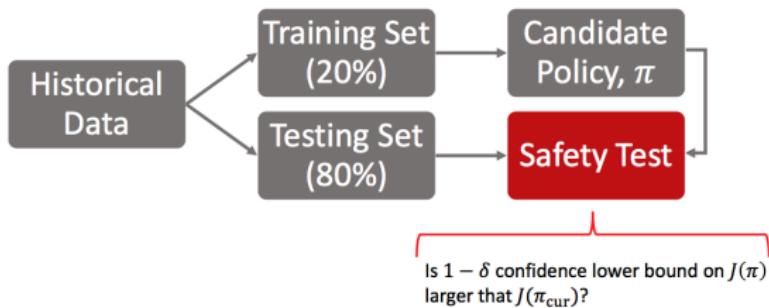
$$\mathbb{E}[X_i] \geq \frac{1}{n} \sum_{i=1}^n X_i - b \sqrt{\frac{\ln(1/\delta)}{2n}},$$

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)



Off-policy policy evaluation

- Importance sampling (IS):

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

- Per-decision importance sampling (PDIS)

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

Off-policy policy evaluation (revisited)

- Importance sampling (IS):

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

- Weighted importance sampling (WIS)

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

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

Off-policy policy evaluation (revisited)

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

Control variates

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

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:

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

- Lower variance if $2\text{Cov}(X, Y) > \text{Var}(Y)$
- 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
 - $\mathbb{E}[Y] = 0$ in this case
 - $PDIS_{CV}(D) = PDIS(D) - CV(D)$
- Called the doubly robust estimator (Jiang and Li, 2015)
 - Robust to (1) poor approximate model, and (2) error in estimates of π_b
 - 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
 - $DR(D) = PDIS_{CV}(D)$
- Non-recursive and weighted forms, as well as control variate view provided by Thomas and Brunskill (ICML 2016)

Off-policy policy evaluation (revisited)

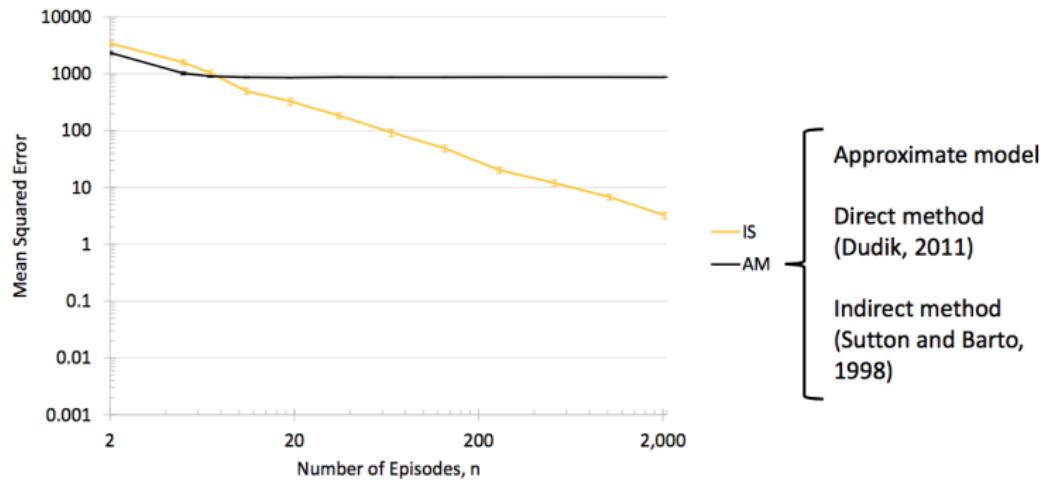
$$DR(\pi_e \mid D) = \frac{1}{n} \sum_{i=1}^n \sum_{t=0}^{\infty} \gamma^t w_t^i (R_t^i - \hat{q}^{\pi_e}(S_t^i, A_t^i)) + \gamma^t \rho_{t-1}^i \hat{v}^{\pi_e}(S_t^i),$$

where $w_t^i = \prod_{\tau_1}^t \frac{\pi_e(a_\tau \mid s_\tau)}{\pi_b(a_\tau \mid s_\tau)}$

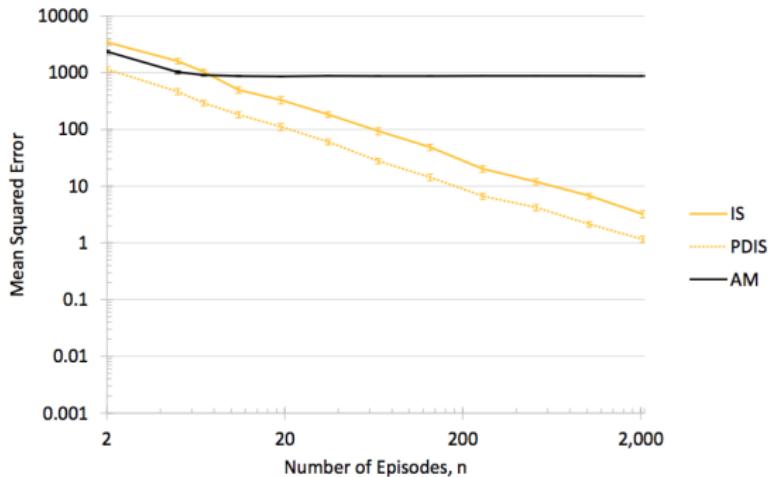
- Recall: we want the control variate Y to cancel with X :

$$R - q(S, A) + \gamma v(S')$$

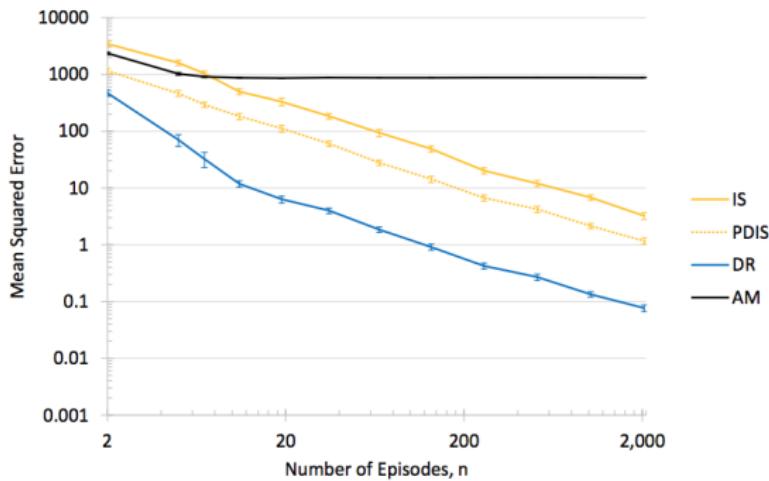
Empirical Results (Gridworld)



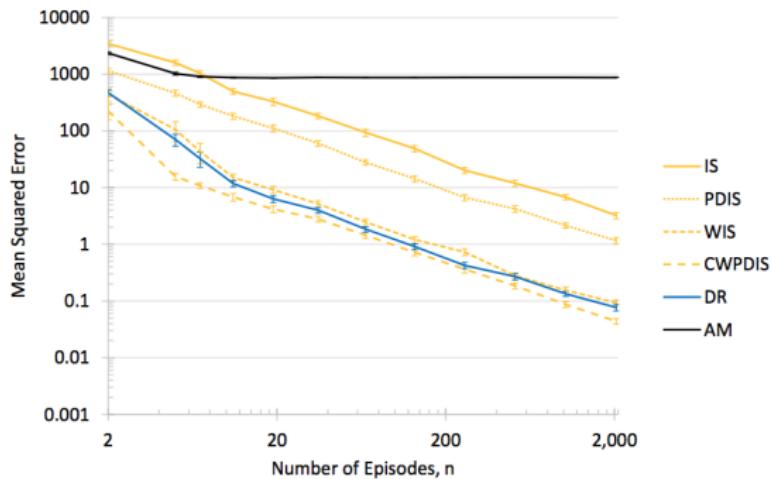
Empirical Results (Gridworld)



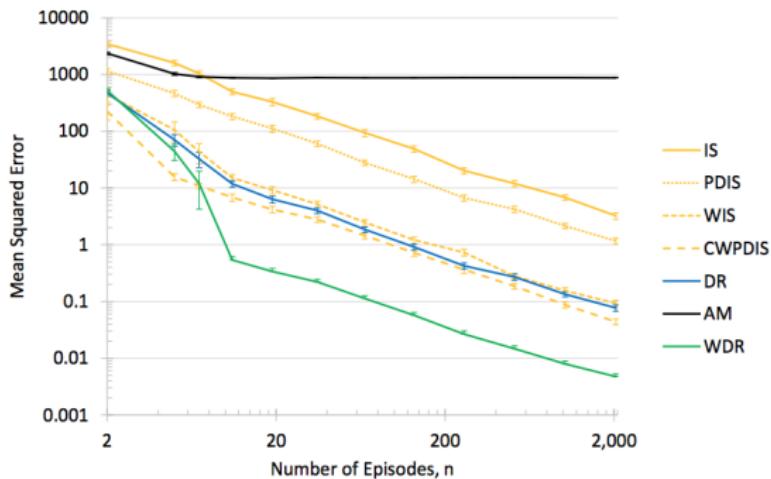
Empirical Results (Gridworld)



Empirical Results (Gridworld)



Empirical Results (Gridworld)

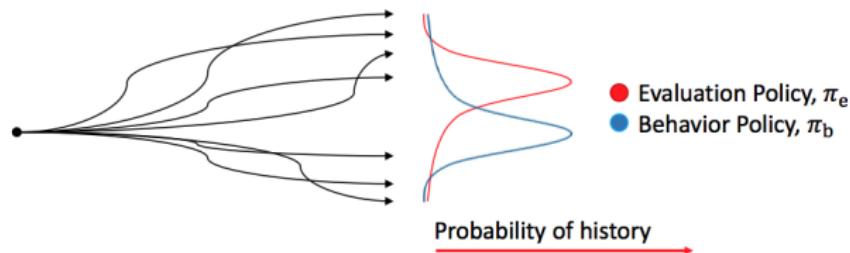


Off-policy policy evaluation (revisited): Blending

- 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 2016)
- Can be particularly useful when part of the world is non-Markovian in the given model, and other parts of the world are Markov

Off-policy policy evaluation (revisited)

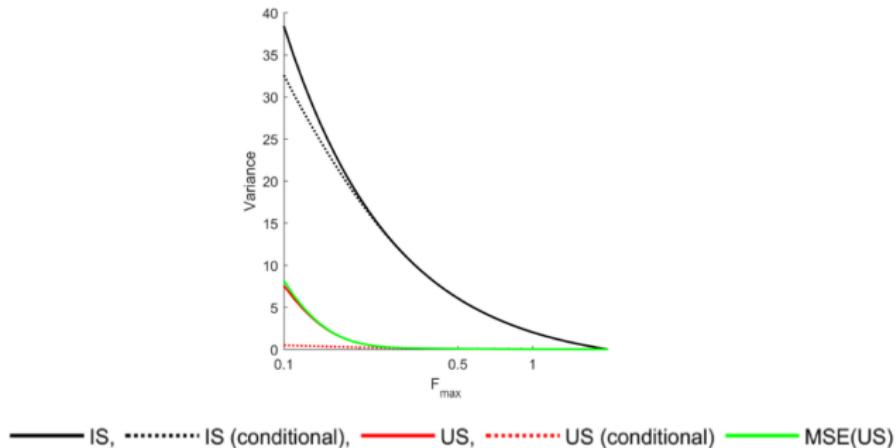
- What if $\text{supp}(\pi_e) \subset \text{supp}(\pi_b)$
- There is a state-action pair, (s, a) , such that $\pi_e(a | s) = 0$, but $\pi_b(a | s) \neq 0$.
- If we see a history where (s, a) occurs, what weight should we give it?
- $IS(D) = \frac{1}{n} \sum_{i=1}^n \left(\prod_{t=1}^L \frac{\pi_e(a_t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



Off-policy policy evaluation (revisited)

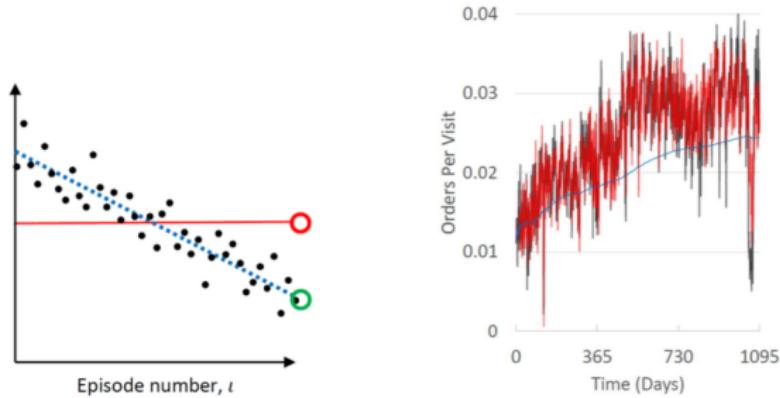
- What if there are zero samples ($n = 0$)?
 - The importance sampling estimate is undefined
- What if no samples are in $\text{supp}(\pi_e)$ (or $\text{supp}(p)$ in general)?
 - Importance sampling says: the estimate is zero
 - Alternate approach: undefined
- Importance sampling estimator is unbiased 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, AAAI 2017)

Can Need An Order of Magnitude Less Data

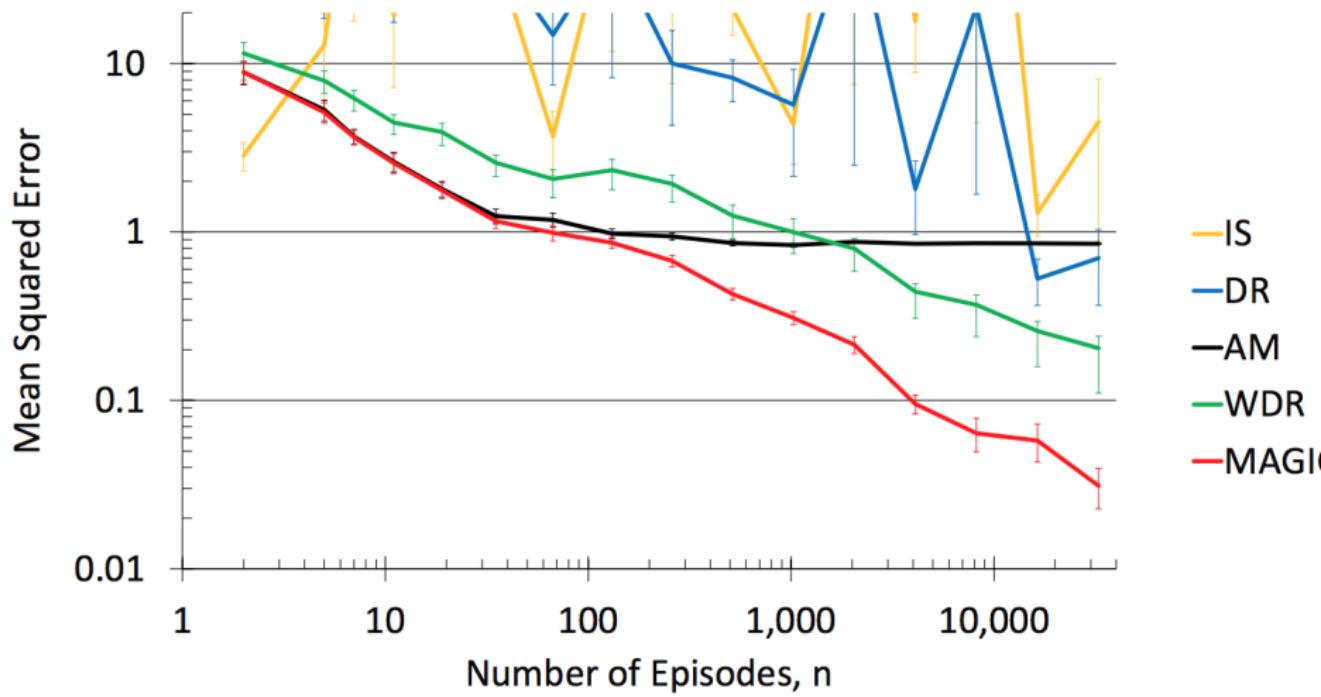


Off-policy policy evaluation (revisited)

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



Off-policy policy evaluation (revisited)



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 (revisited)

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

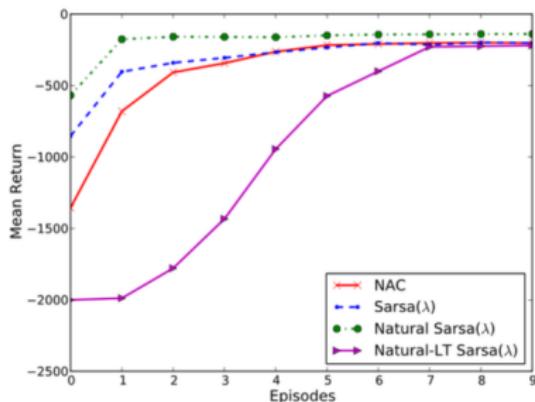
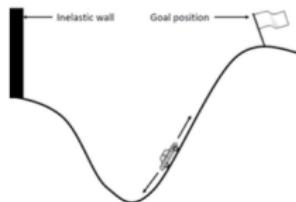
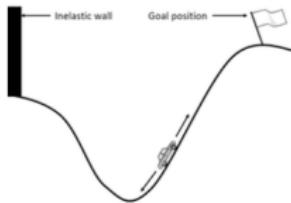


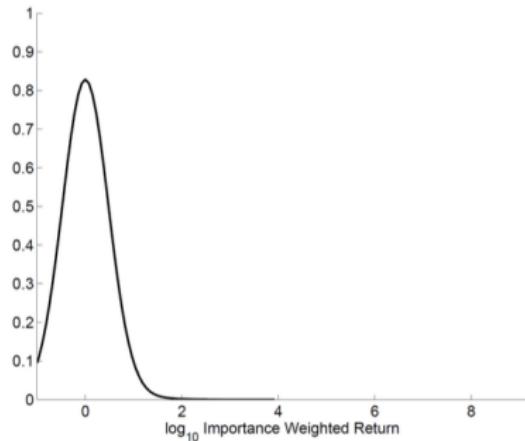
Figure 3: Mountain Car (Sarsa(λ))
Natural Temporal Difference Learning, Dabney and Thomas, 2014

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



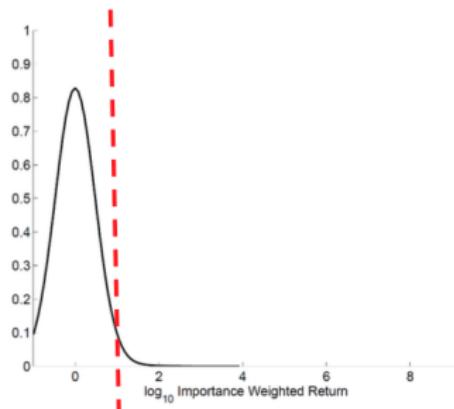
What went wrong



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

High-confidence off-policy policy evaluation (revisited)

- Removing the upper tail only decreases the expected value.



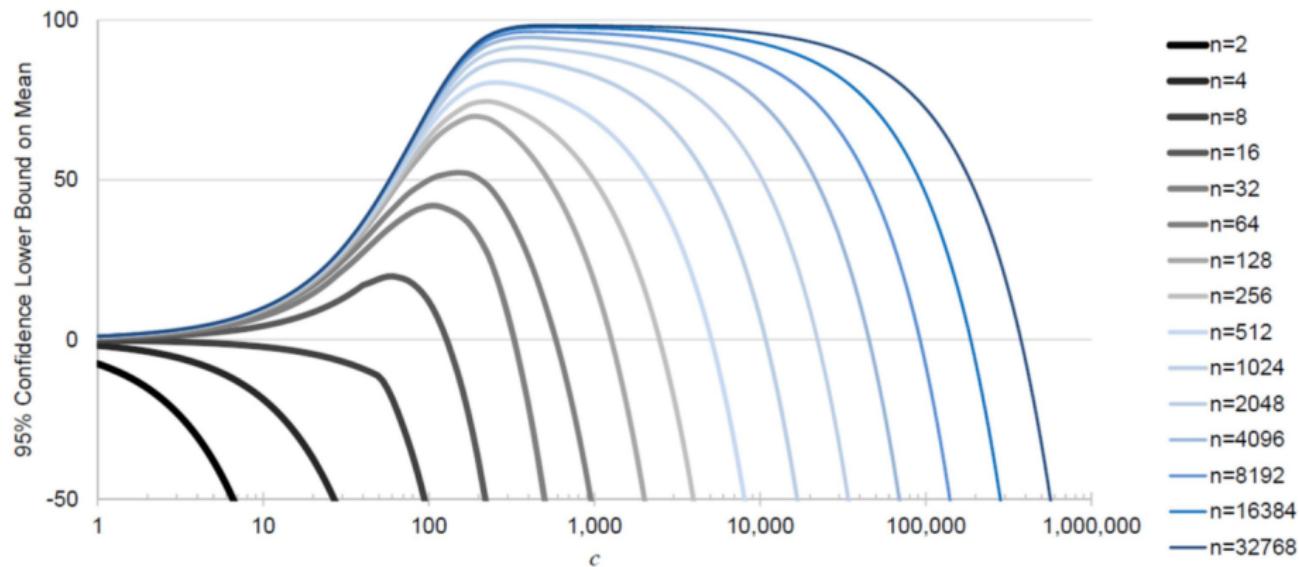
High-confidence off-policy policy evaluation (revisited)

- 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)



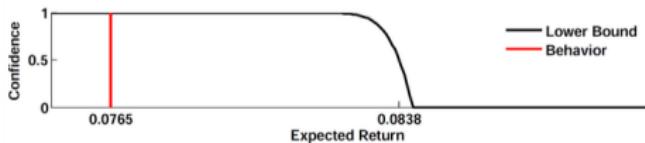
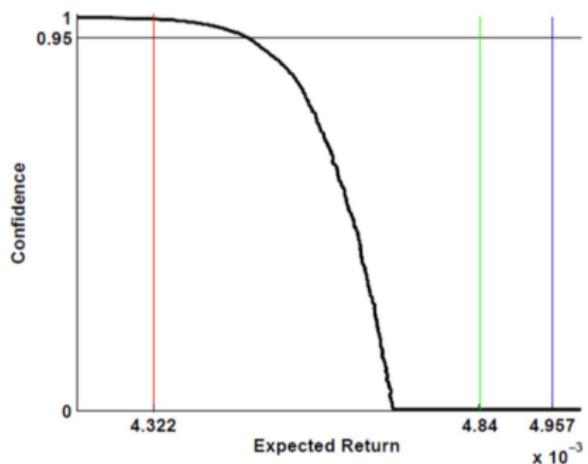
High-confidence off-policy policy evaluation (revisited)

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

	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

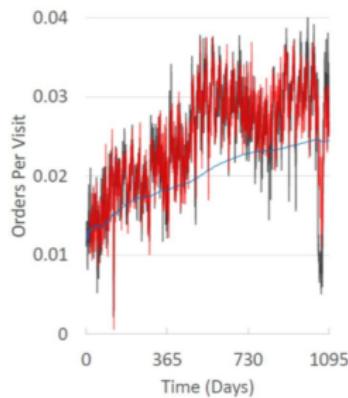
High-confidence off-policy policy evaluation (revisited)

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

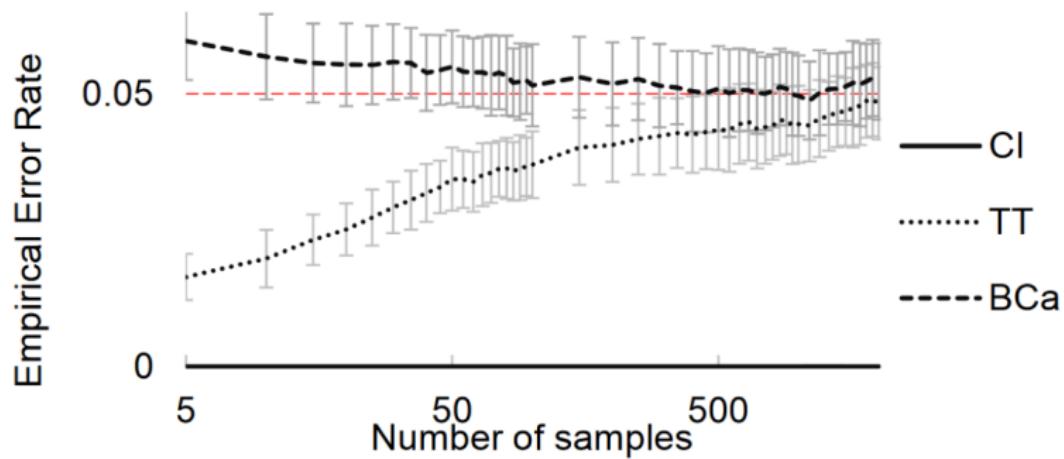
- Assumes that $IS(D)$ is normally distributed
- By the central limit theorem, it (is as $n \rightarrow \infty$)

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

- Efron's Bootstrap methods (e.g., BCa)

- Also, without importance sampling: Hanna, Stone, and Niekum, AAMAS 2017

High-confidence off-policy policy evaluation (revisited)



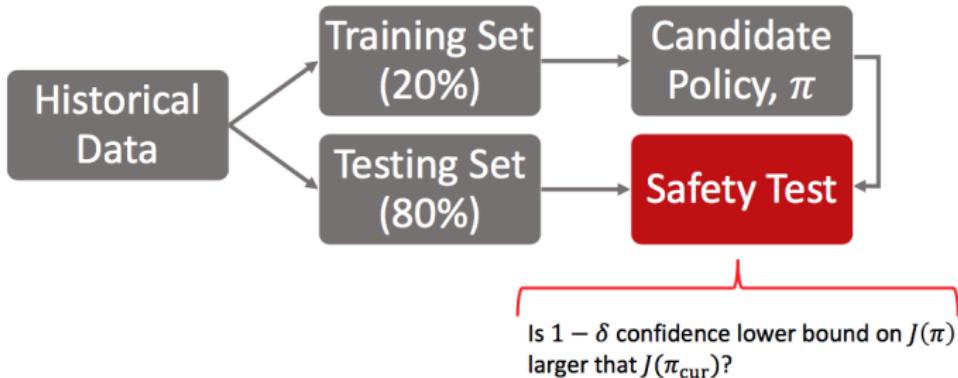
P. S. Thomas. Safe reinforcement learning (PhD Thesis, 2015)

Create a safe batch reinforcement learning algorithm

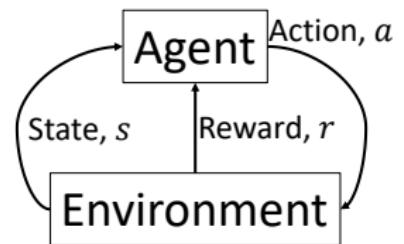
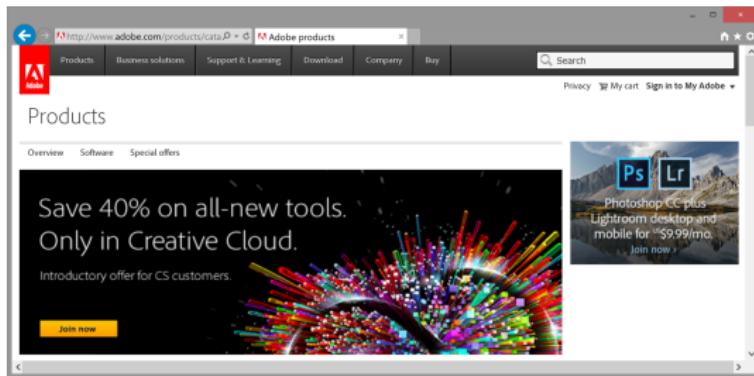
- Off-policy policy evaluation (OPE)
 - For any evaluation policy, π_e , Convert historical data, D , into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
 - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} 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)

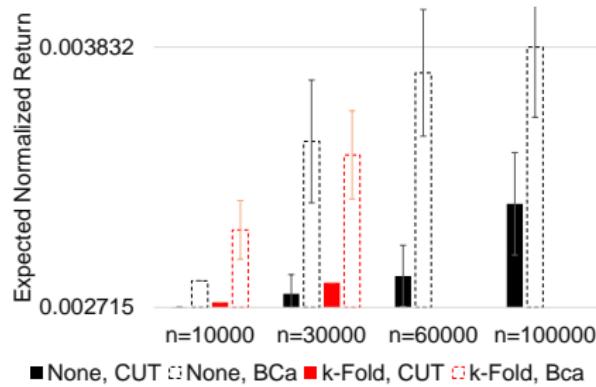
Thomas et. al, ICML 2015



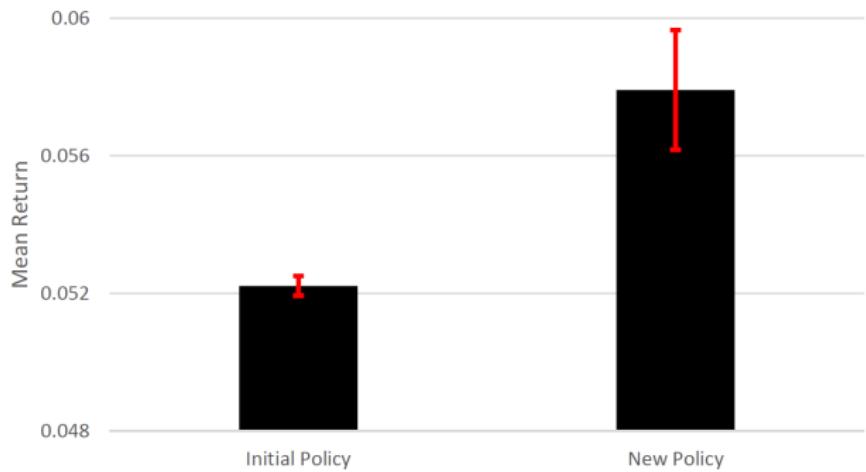
Empirical Results: Digital Marketing



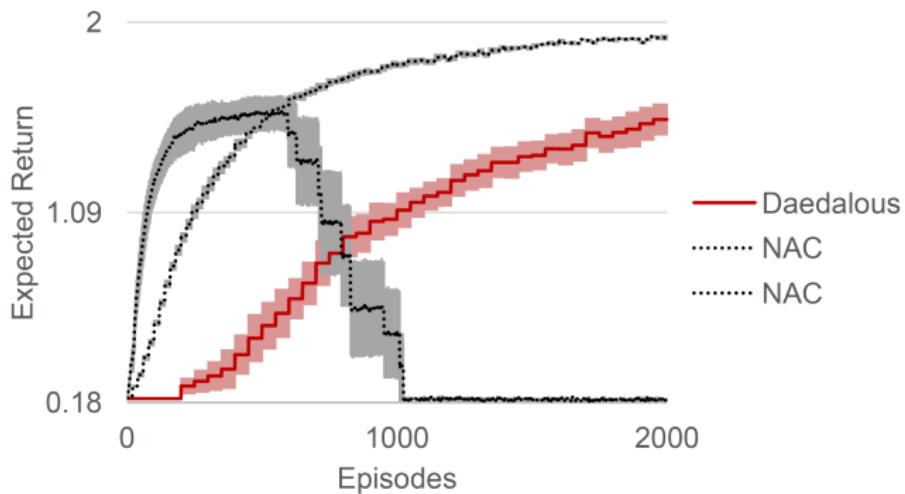
Empirical Results: Digital Marketing



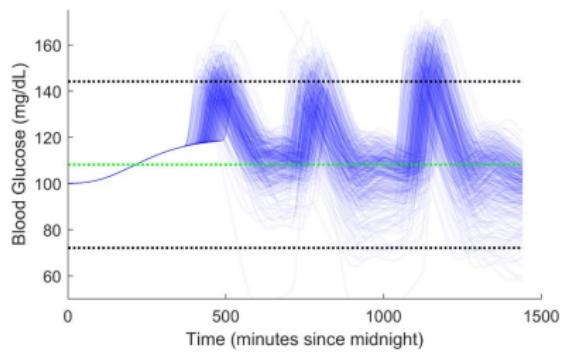
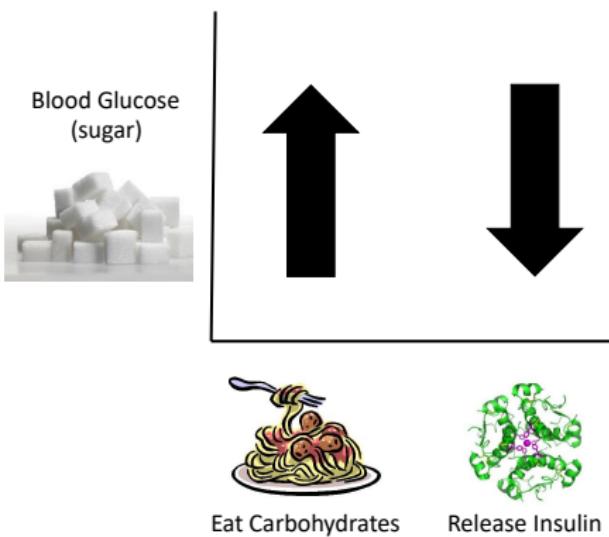
Empirical Results: Digital Marketing



Empirical Results: Digital Marketing



Example Results : Diabetes Treatment



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

- 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