Lecture 14: Batch RL

Emma Brunskill

CS234 Reinforcement Learning.

Winter 2020
Slides drawn from Philip Thomas with modifications

*Note: we only went carefully through slides before slide 34. The remaining slides are kept for those interested but will not be material required for the quiz. See the last slide for a summary of what you should know

Refresh Your Understanding: Fast RL III

Select all that are true:

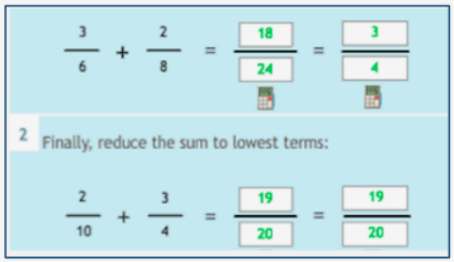
- Thompson sampling for MDPs the posterior over the dynamics can be updated after each transition
- When using a Beta prior for a Bernoulli reward parameter for an (s,a) pair, the posterior after N samples of that pair time steps can be the same as after N+2 samples
- The optimism bonuses discussed for MBIE-EB depend on the maximum reward but not on the maximum value function
- In class we discussed adding a bonus term to the policy gradient update for a (s,a,r,s') tuple using Q-learning with function approximation. Adding this bonus term will ensure all Q estimates used to make decisions online using DQN are optimistic with respect to Q*
- Not sure

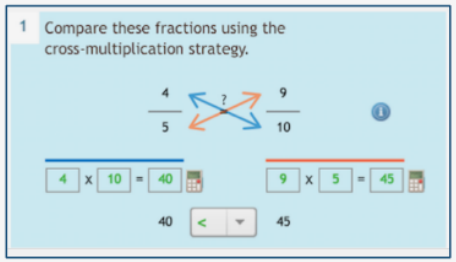
Class Structure

- Last time: Fast Reinforcement Learning
- **This time: Batch RL**
- Next time: Guest Lecture

A Scientific Experiment

A Group

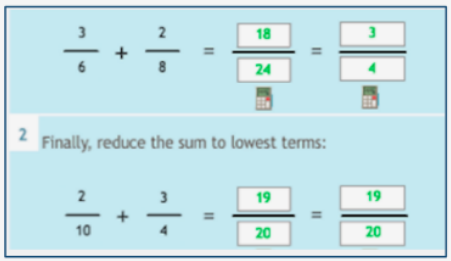


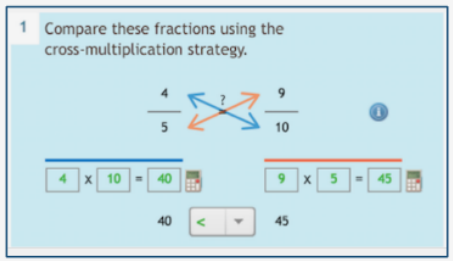


Avg Score: 95

A Scientific Experiment

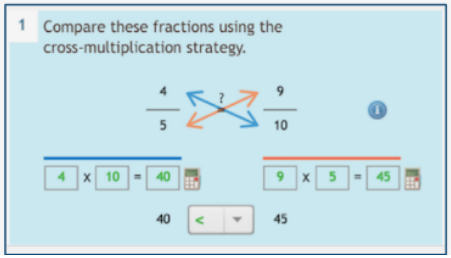
A Group

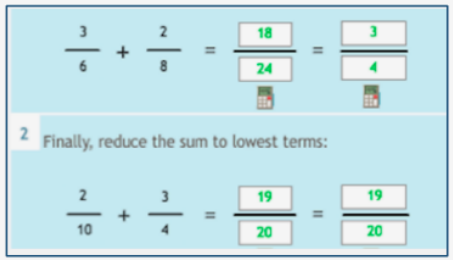




Avg Score: 95

B Group





Avg Score: 92

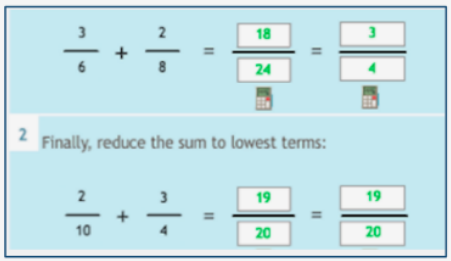
What Should We Do For a New Student?

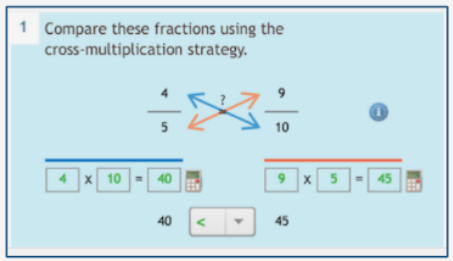
kp new group

sample size

distrib of scores

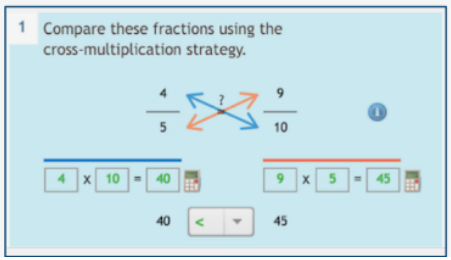
A Group

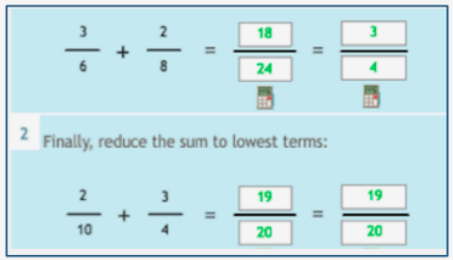




Avg Score: 95

B Group

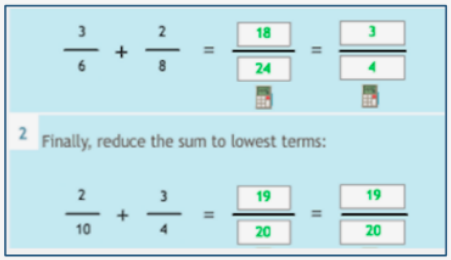


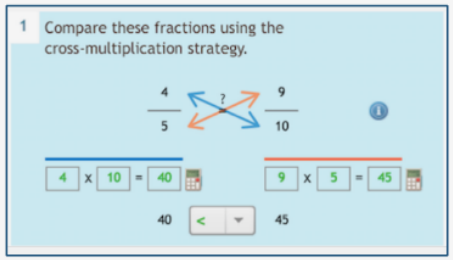


Avg Score: 92

Involves Counterfactual Reasoning

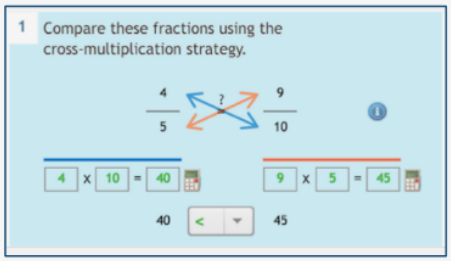
A Group

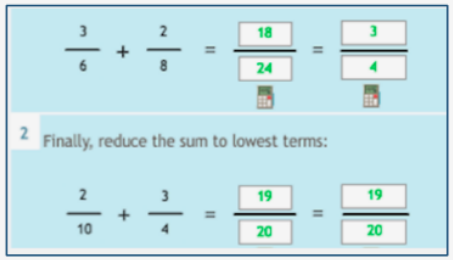




Avg Score: 95

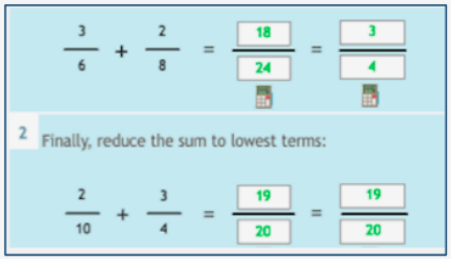
B Group

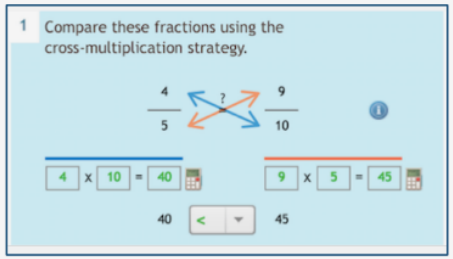




Avg Score: 92

B Group

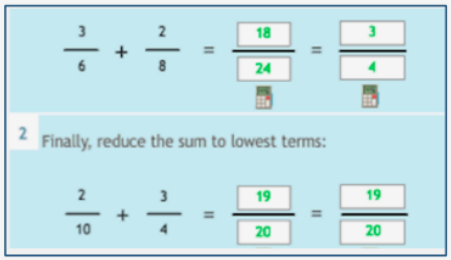


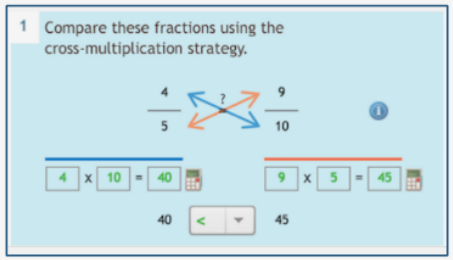


???

Involves Generalization

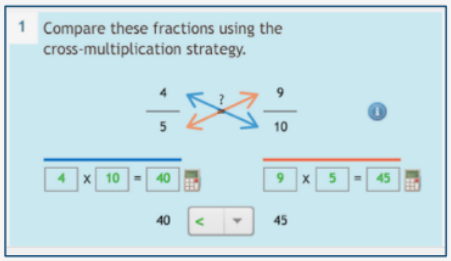
A Group

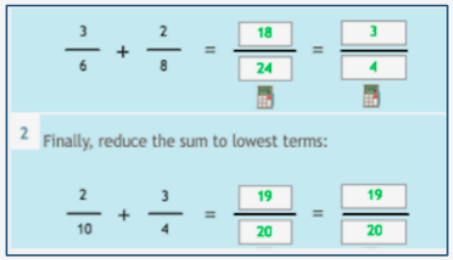




Avg Score: 95

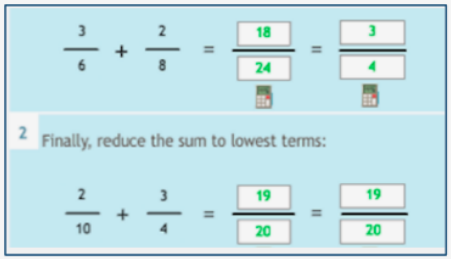
B Group

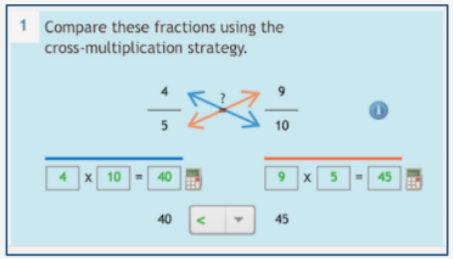




Avg Score: 92

B Group

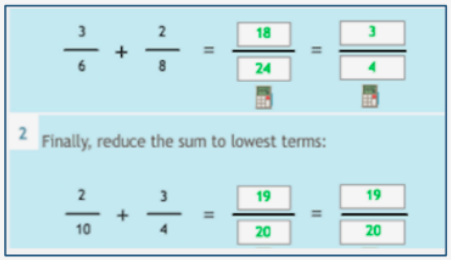


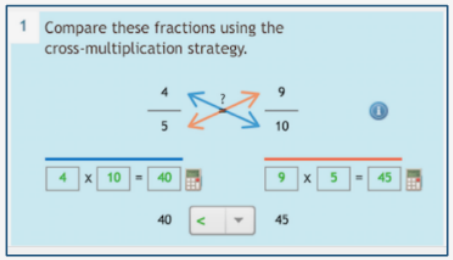


???

Batch Reinforcement Learning

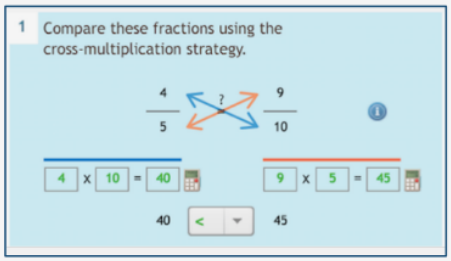
A Group

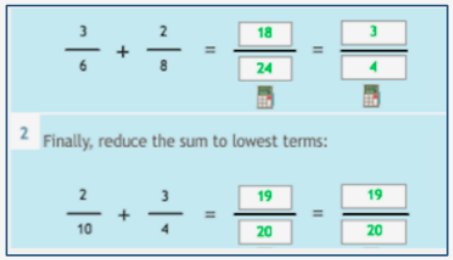




Avg Score: 95

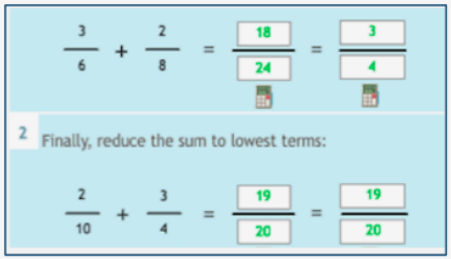
B Group

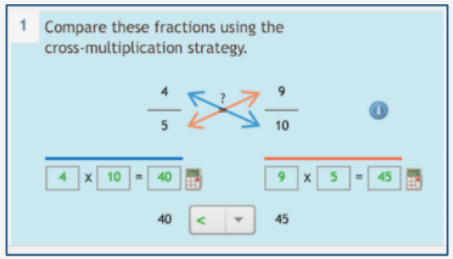




Avg Score: 92

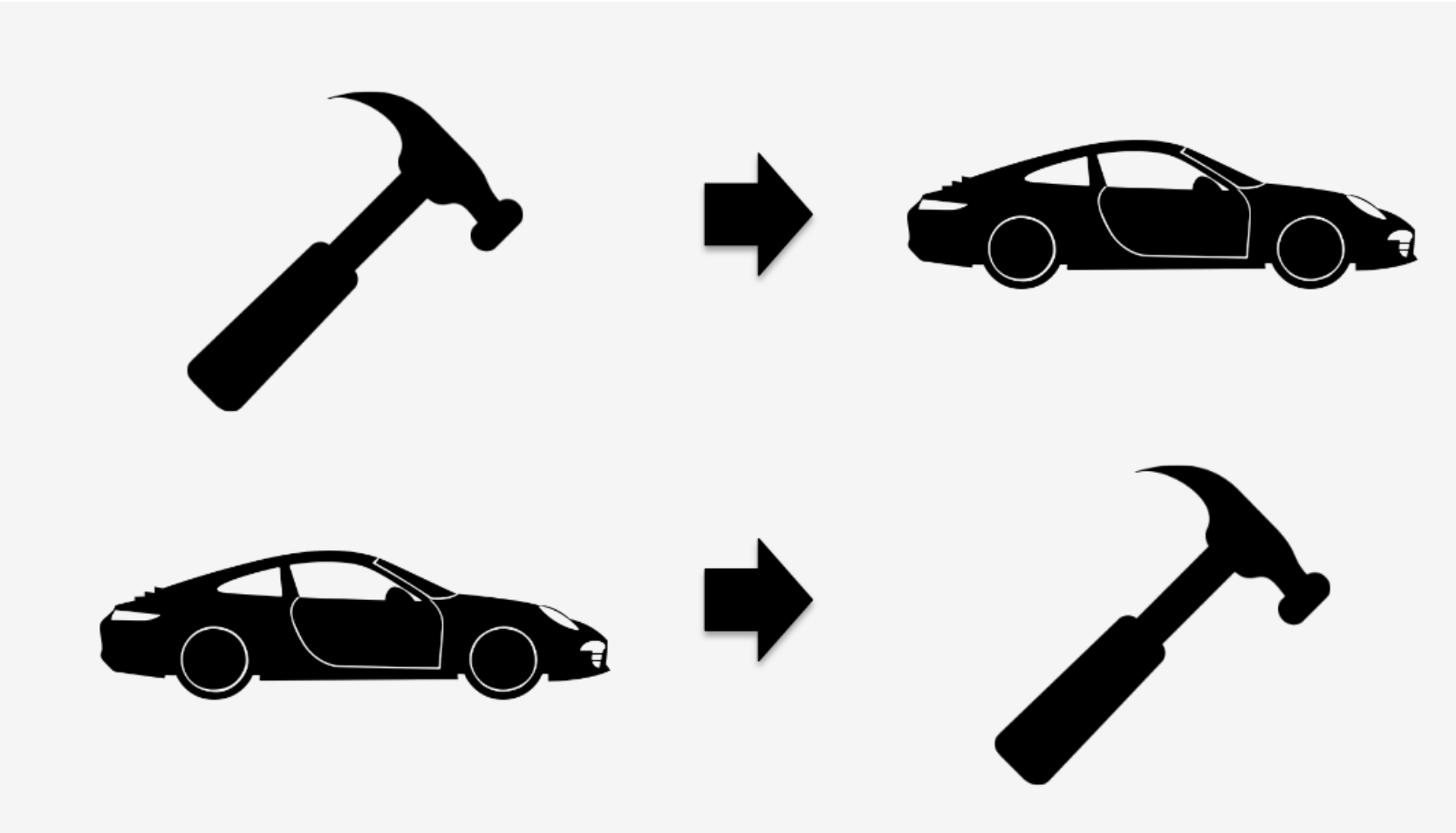
B Group



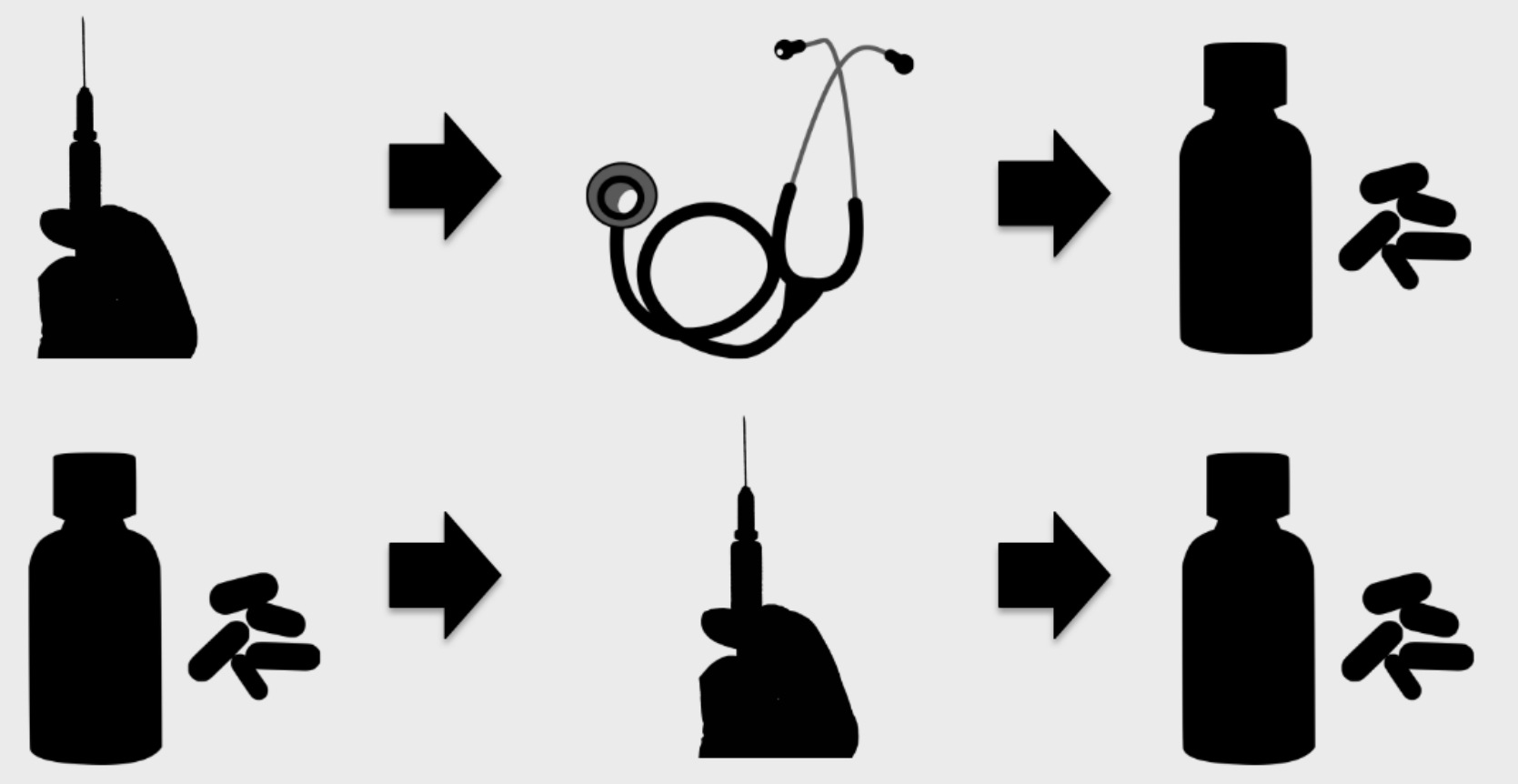


???

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

What property should a safe batch reinforcement learning algorithm have?

TRPO
policy gradient methods

- 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)
   - Safe policy improvement (SPI)

Notation

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

$$V^{\pi} = \mathbb{E}[\sum_{t=1}^{L} \gamma^t R_t \,|\, \pi]$$

Safe batch reinforcement learning algorithm

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

- 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:

value of current policy used in dataset

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

value of policy output by alg

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)
- 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,
- Methods today focused on work by Philip Thomas UAI and ICML 2015 papers.

Off-policy policy evaluation (OPE)

Historical Data, D
Proposed Policy, π_e

$\longrightarrow$ Estimate of V^{π_e}

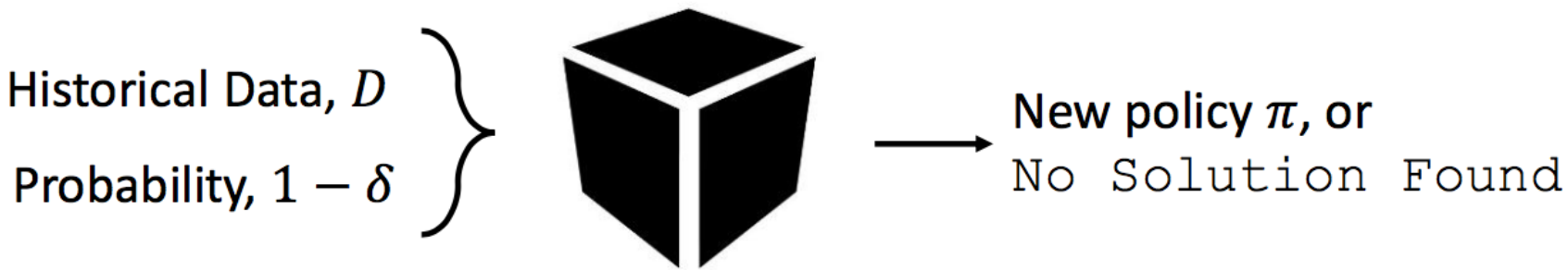
evaluation

High-confidence off-policy policy evaluation (HCOPE)

Historical Data, D
Proposed Policy, π_e
Probability, $1-\delta$

$1-\delta$ confidence lower bound on V^{π_e}

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

- Aim: estimate value of policy π_1, $V^{\pi_1}(s)$, given episodes generated under behavior policy π_2
  - $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ where the actions are sampled from π_2
- $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$ in MDP M under policy π
- $V^{\pi}(s) = \mathbb{E}_{\pi}[G_t | s_t = s]$

unbiased
high var

- 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

coverage

- Distribution of episodes & resulting returns differs between policies

imitation learning

IRL

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 $x_1, x_2, \ldots, x_n$ sampled from distribution $q(s)$
- Under a few assumptions, we can use samples to obtain an unbiased estimate of $\mathbb{E}_{x\sim q}[f(x)] \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i)$ $x_i \sim q$

Importance Sampling

~~$\mathbb{E}_{x \sim q}[f(x)] = \int_x q(x) f(x)$~~

"Overlap"

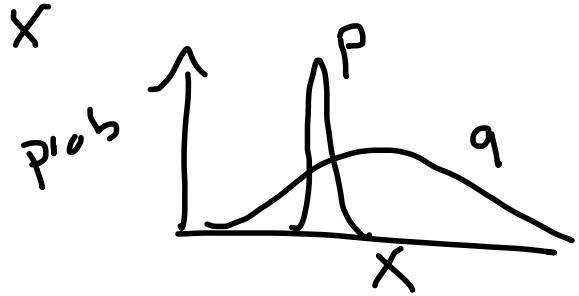
$$\mathbb{E}_{x \sim p}[f(x)] = \int_x p(x) f(x) dx$$

$$= \int_x \frac{q(x)}{q(x)} p(x) f(x) dx$$

$$= \int_x q(x) \left[\frac{p(x)}{q(x)} f(x) \right] dx$$

$$= \mathbb{E}_{x \sim q}\left[\frac{p}{q} f \right]$$

$$\approx \frac{1}{N} \sum_{\substack{i=1 \\ x_i \sim q}}^{N} \frac{p(x_i)}{q(x_i)} f(x_i)$$



unbiased

$\forall x$ s.t. $p(x) > 0$ then $q(x) > 0$

Checking Your Understanding: Importance Sampling

We can use importance sampling to do batch bandit policy evaluation. Consider we have a dataset for pulls from 3 arms. Consider that arm 1 is a Bernoulli where with probability .98 we get 0 and with probability 0.02 we get 100. Arm 2 is a Bernoulli where with probability 0.55 the reward is 2 else the reward is 0. Arm 3 has a probability of yielding a reward of 1 with probability 0.5 else it gets 0. Select all that are true.

- Data is sampled from pi1 where with probability 0.8 it pulls arm 3 else it pulls arm 2. The policy we wish to evaluate, pi2, pulls arm 2 with probability 0.5 else it pulls arm 1. pi2 has higher true reward than pi1.
- We cannot use pi1 to get an unbiased estimate of the average reward pi2 using importance sampling. pi1 never pulls arm1
- We can use pi1 to get a lower bound on the average reward of pi2 using importance sampling. if all rewards are pos
- If rewards can be positive or negative, we can still get a lower bound on pi2 using data from pi1 using importance sampling
- Now assume pi1 selects arm1 with probability 0.2 and arm2 with probability 0.8. We can use importance sampling to get an unbiased estimate of pi2 using data from pi1.
- Still with the same pi1, it is likely with N=20 pulls that the estimate using IS for pi2 will be higher than the empirical value of pi1.
- Not sure

Importance Sampling (IS) for RL Policy Evaluation

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

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

Importance Sampling (IS) for Policy Evaluation

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

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

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

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

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

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

policy gradient

$$= \prod_{i=1}^{L-1} \frac{p(a_i|\pi_1, s_i)}{p(a_i|\pi_2, s_i)}$$

Importance Sampling for Policy Evaluation

- Aim: estimate $V^{\pi_1}(s)$ given episodes generated under policy π_2
  - $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ where the actions are sampled from π_2
- Have access to $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$ 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

Leveraging Future Can't Influence Past Rewards

- Importance sampling (IS):

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

policy grad

don't need dynamics model

- Per-decision importance sampling (PDIS)

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

Off-policy policy evaluation

*Note: we only went carefully through slides before this point. The remaining slides are kept for those interested but will not be material required for the quiz. See the last slide for a summary of what you should know

- 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

- 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 or unbiased?

Off-policy policy evaluation

- 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. $\Pr(\lim_{n \to \infty} WIS(D) = V^{\pi_e}) = 1$
  - If
    - Finite horizon
    - One behavior policy, or bounded rewards

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:
  $\mathbb{E}[\hat{\mu}] =$
- Variance:

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

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:

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

- Lower variance if $2Cov(X, Y) > Var(Y)$
- We call Y a control variate
- We saw this idea before: baseline term in policy gradient estimation

Off-policy policy evaluation

- 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

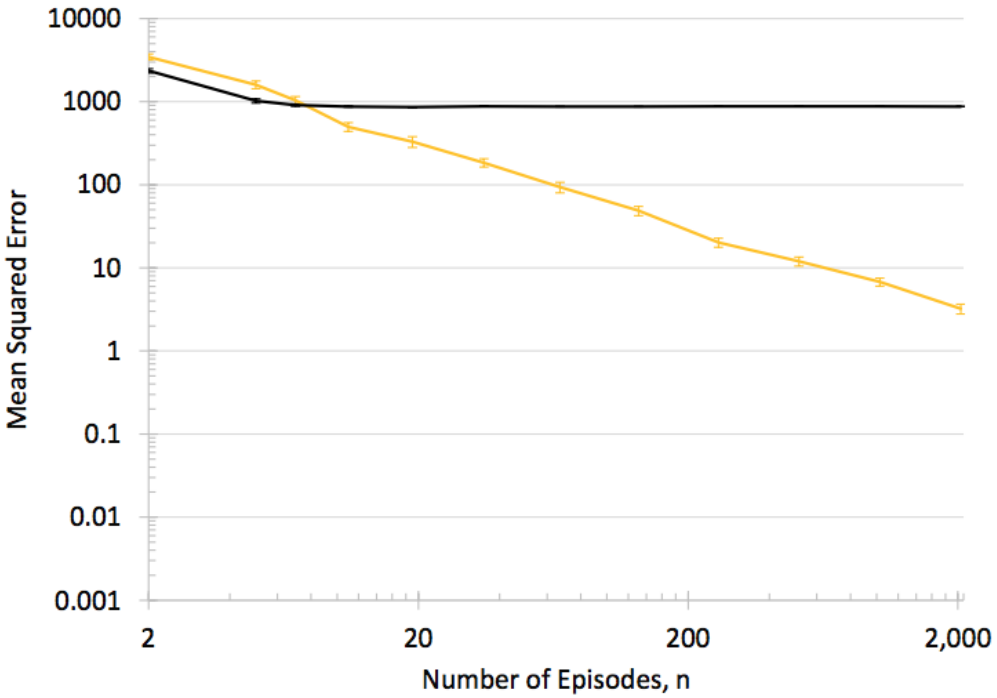
- Called the doubly robust estimator (Jiang and Li, 2015)
  - Robust to (1) poor approximate model, and (2) error in estimates of π_b
    - If the model is poor,the estimates are still unbiased
    - If the sampling policy is unknown, but the model is good, MSE will still be low
- Non-recursive and weighted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

Off-policy policy evaluation

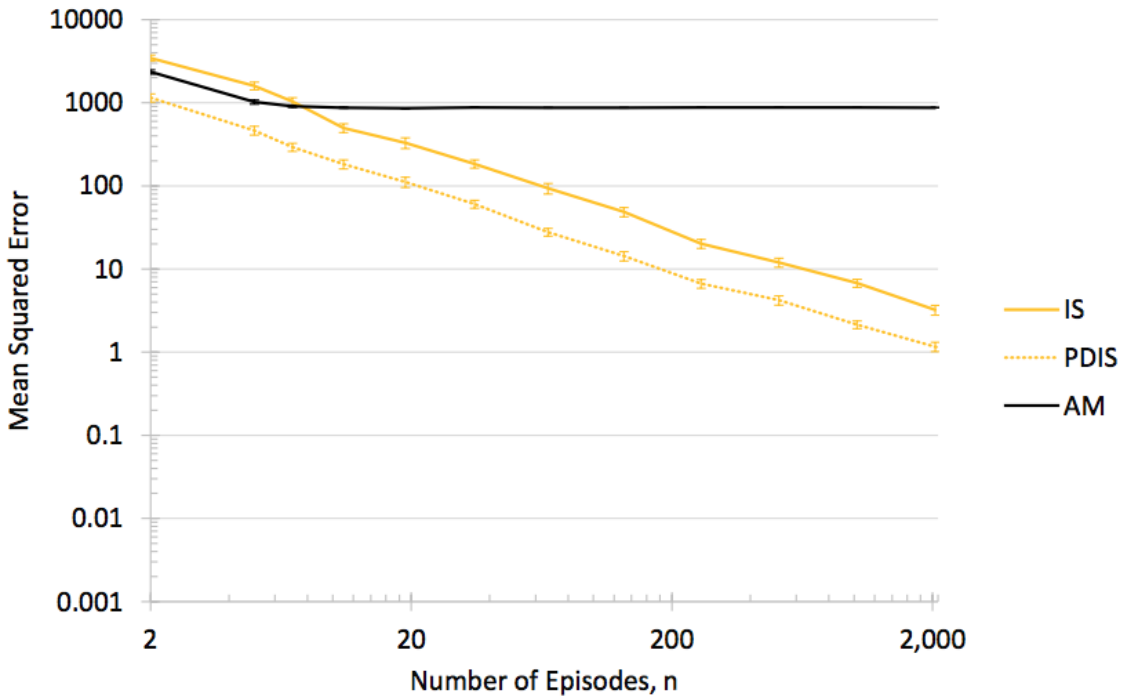
$$DR(\pi_e \mid D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i(R_t^i - \hat{q}^{\pi_e}(S_t^i, A_t^i)) + \gamma^t \rho_{t-1}^i \hat{v}^{\pi_e}(S_t^i),$$

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

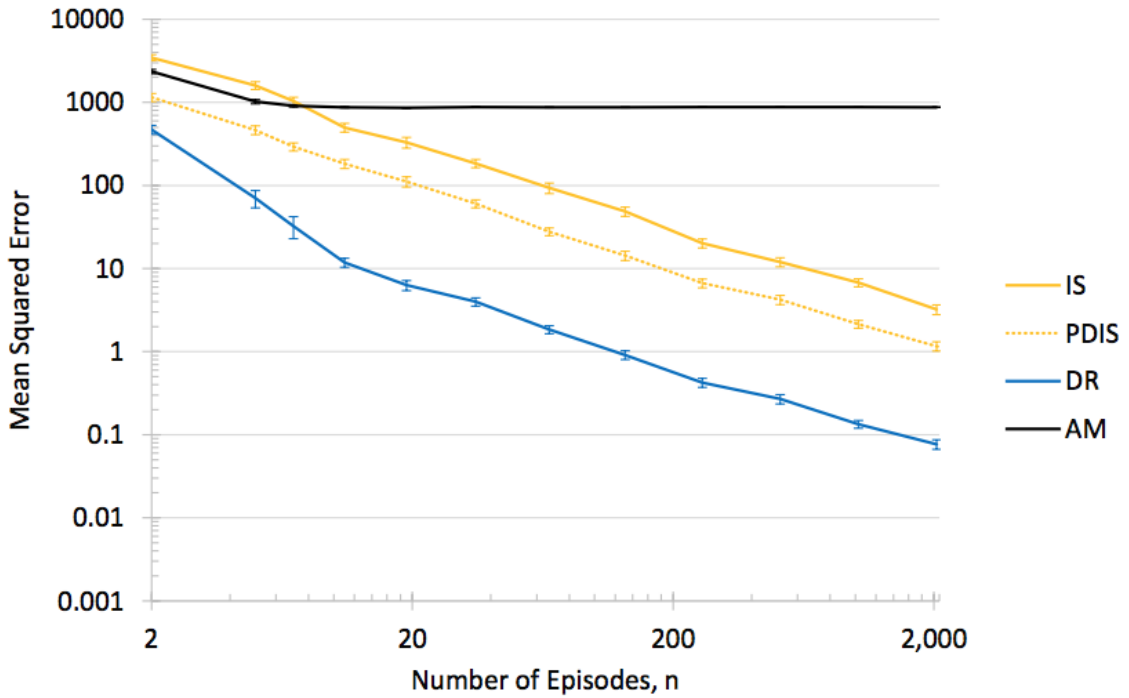
Empirical Results (Gridworld)



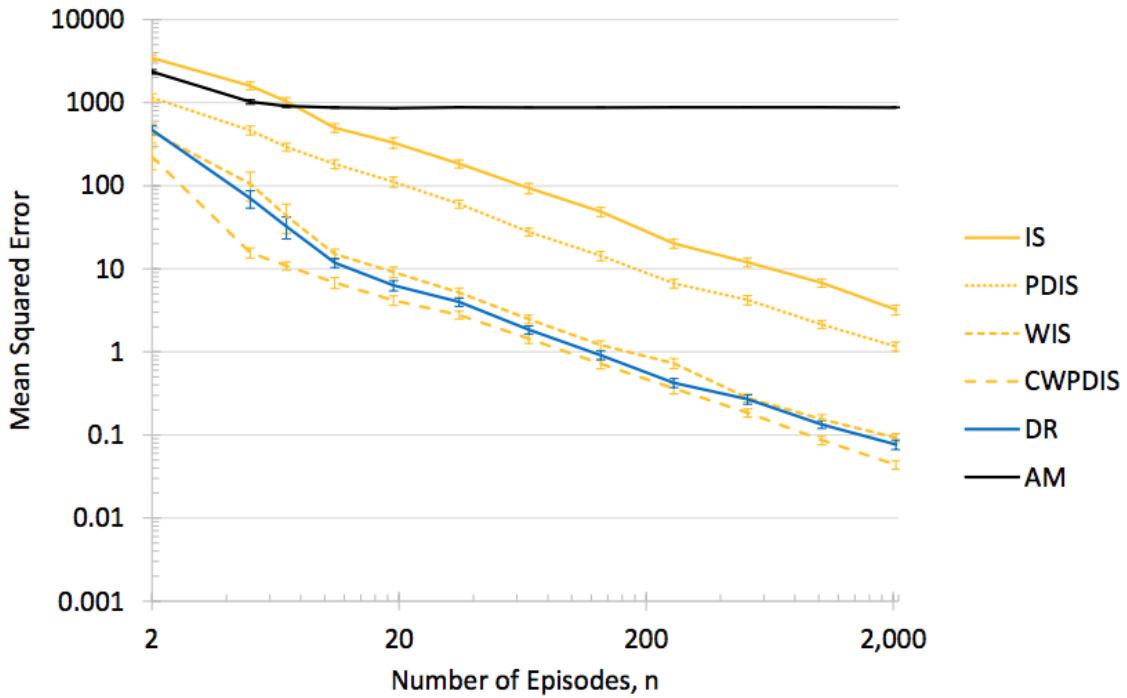
Empirical Results (Gridworld)



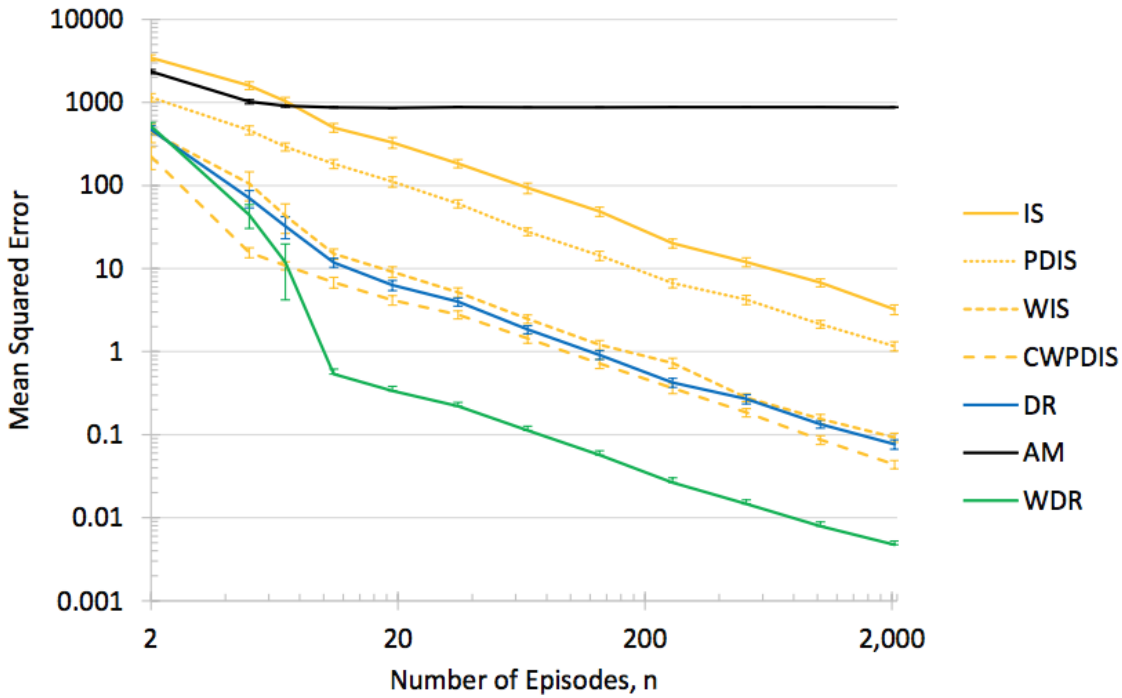
Empirical Results (Gridworld)



Empirical Results (Gridworld)



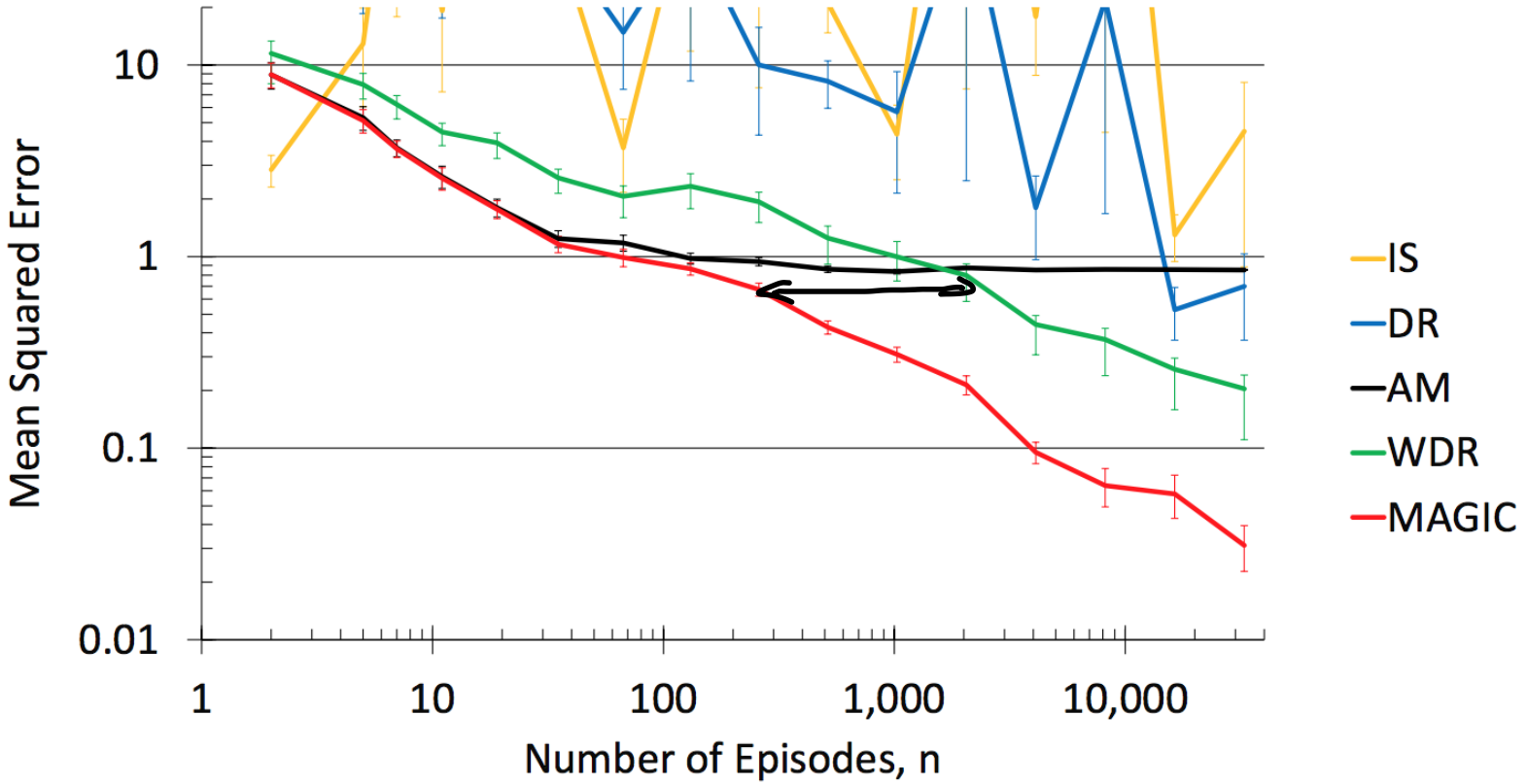
Empirical Results (Gridworld)



Off-policy policy evaluation : Blending

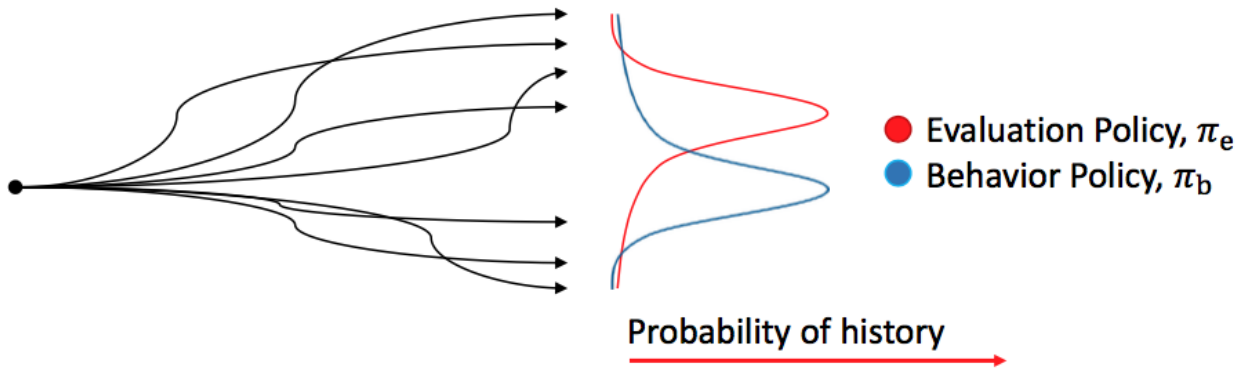
- Importance sampling is unbiased but high variance
- Model based estimate is biased but low variance
- Doubly robust is one way to combine the two
- Can also trade between importance sampling and model based estimate within a trajectory
- MAGIC estimator (Thomas and Brunskill ICML 2016)
- Can be particularly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

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



Off-policy policy evaluation

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

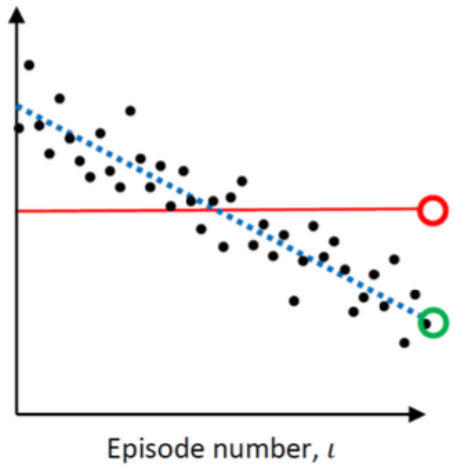


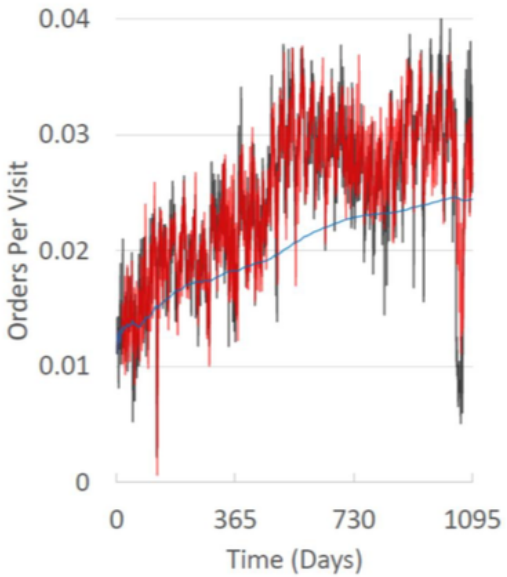
Off-policy policy evaluation

- What if there are zero samples ($n = 0$)?
  - The importance sampling estimate is undefined
- What if no samples are in supp(π_e) (or 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)

Off-policy policy evaluation

- 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

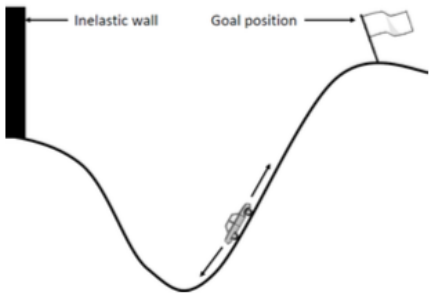
- Off-policy policy evaluation (OPE)
  - For any evaluation policy, π_e, Convert historical data, D, into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e}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

Hoeffding for UCB

- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforcement learning algorithm,

High-confidence off-policy policy evaluation

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



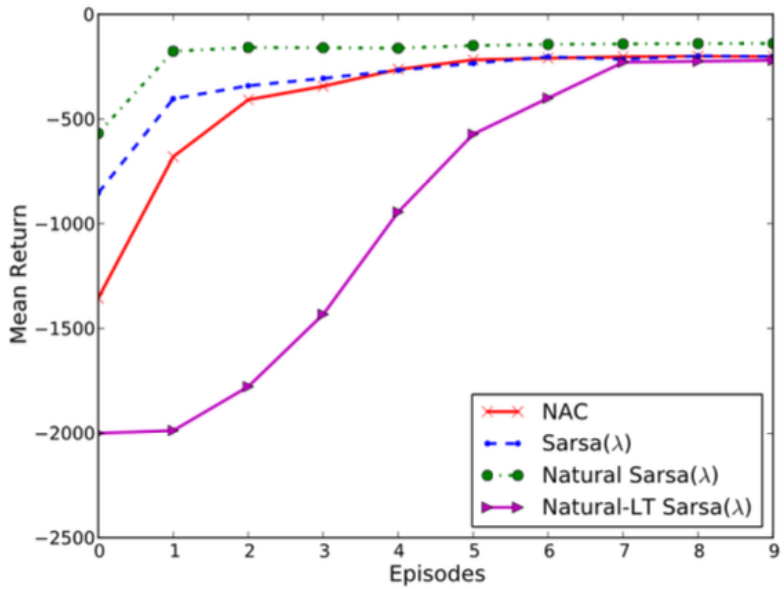


Figure 3: Mountain Car (Sarsa(λ))
Natural Temporal Difference Learning, Dabney and Thomas, 2014

Hoeffding's inequality

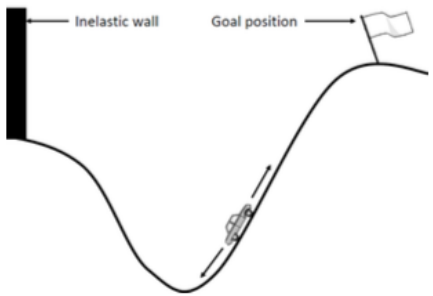
- Let $X_1, \cdots, X_n$ be n independent identically distributed random variables such that $X_i \in [0, b]$
- Then with probability at least $1 - \delta$:

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

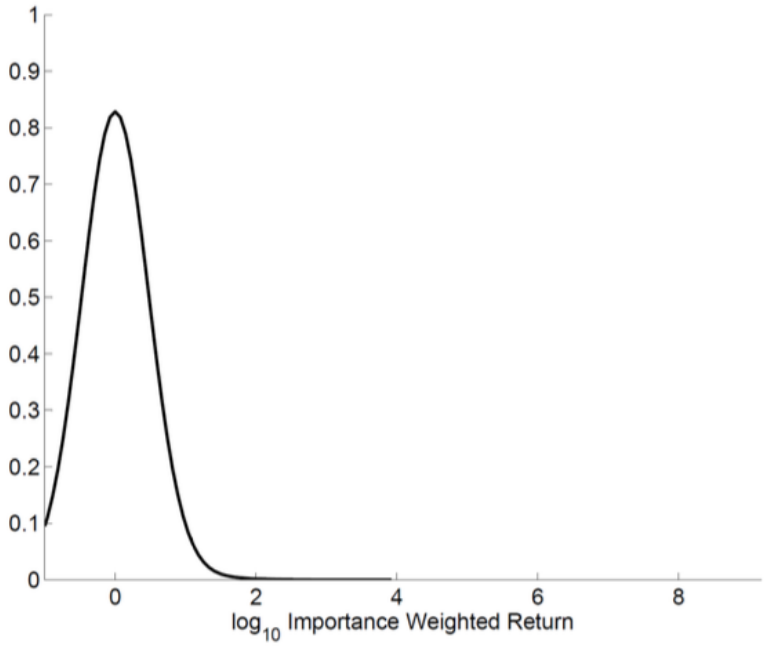
where $X_i = \frac{1}{n} \sum_{i=1}^{n} (w_i \sum_{t=1}^{L} \gamma^t R_t^i)$ in our case.

High-confidence off-policy policy evaluation

- 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



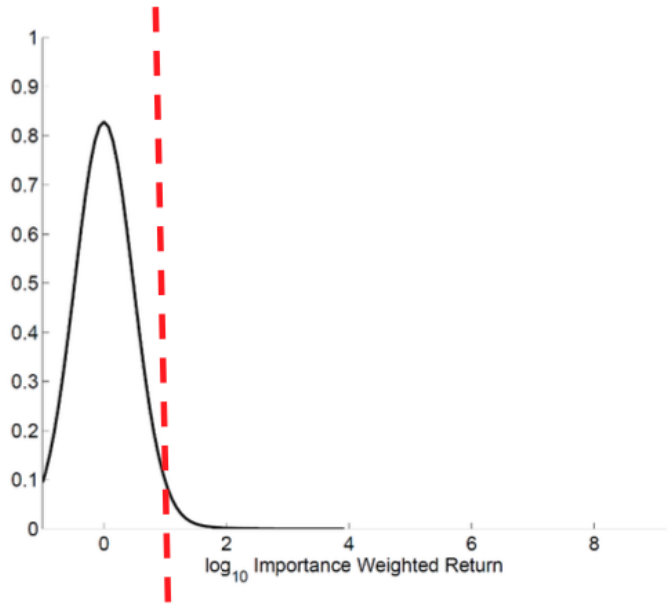
What went wrong



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

High-confidence off-policy policy evaluation

- Removing the upper tail only decreases the expected value.



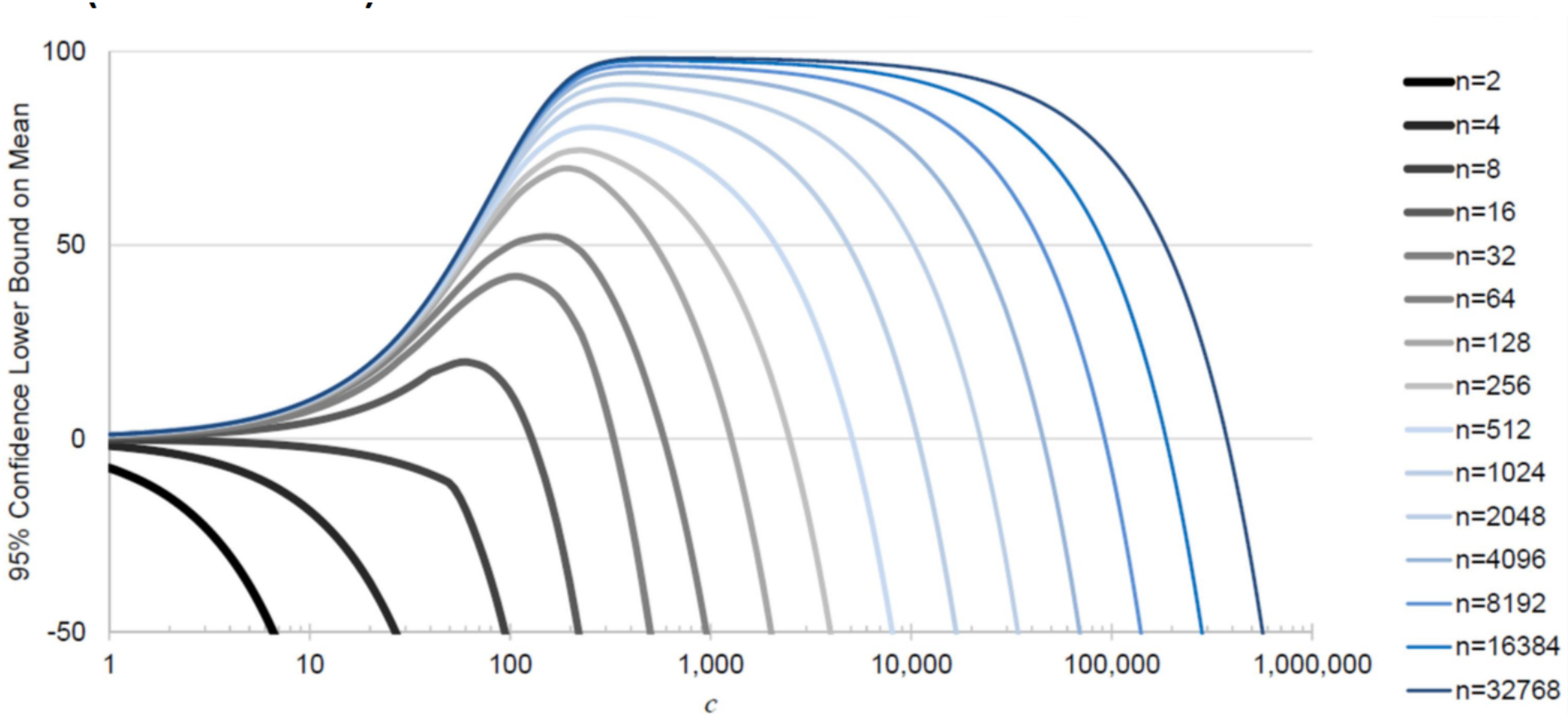
High-confidence off-policy policy evaluation

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

Theorem 1. *Let $X_1, \ldots, X_n$ be n independent real-valued random variables such that for each $i \in \{1, \ldots, n\}$, 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*

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

High-confidence off-policy policy evaluation



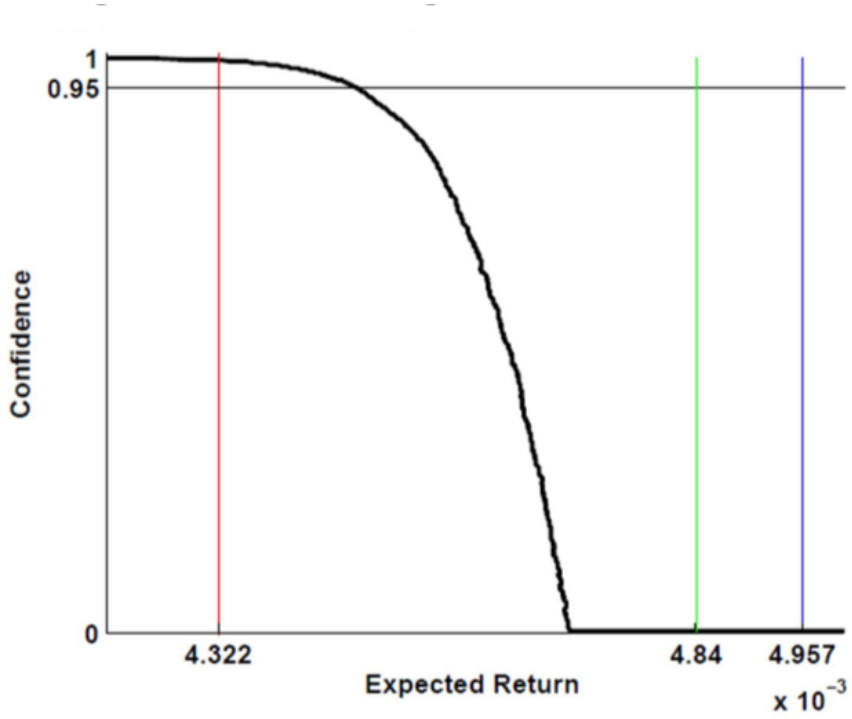
High-confidence off-policy policy evaluation

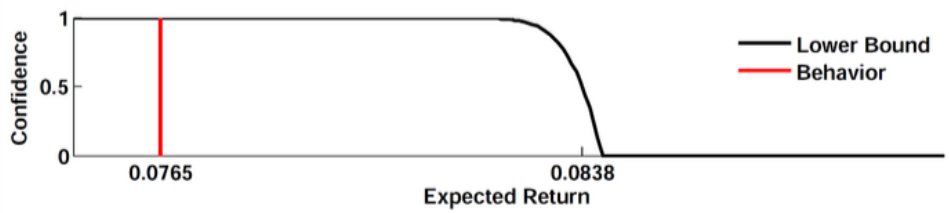
- Use 20% of the data to optimize c (cutoff)
- Use 80% to compute lower bound with optimized c
- Mountain car results:

	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

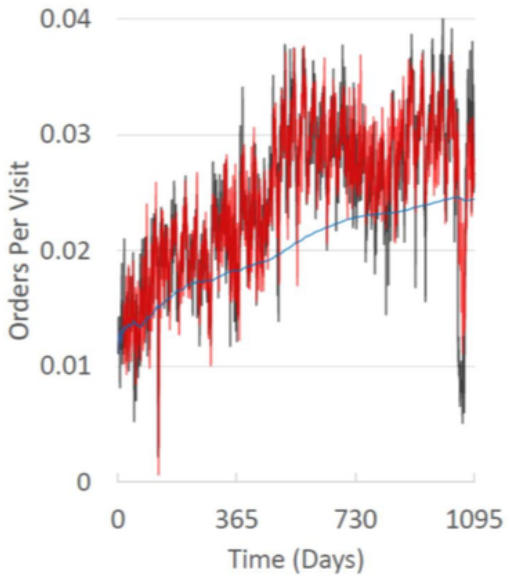
Digital marketing:





High-confidence off-policy policy evaluation

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

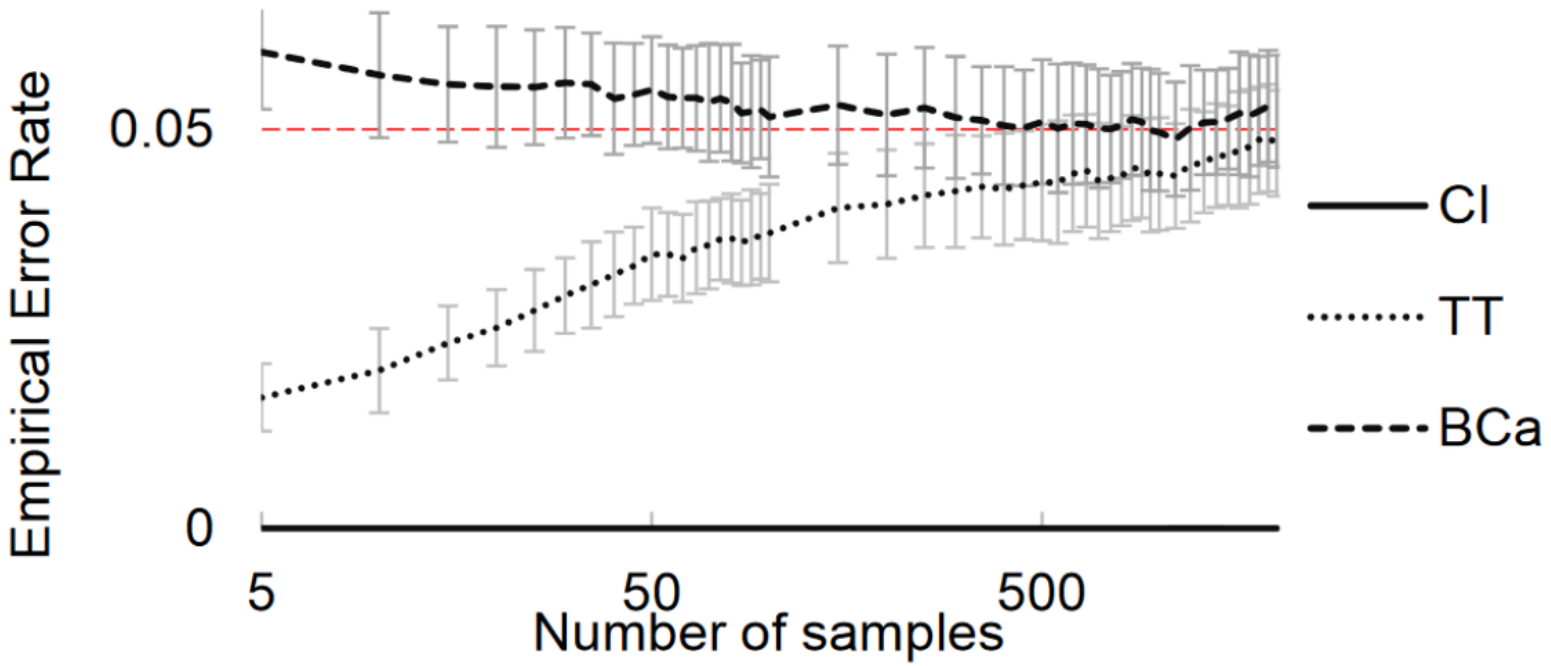
- Student's t-test
  - Assumes that $IS(D)$ is normally distributed
  - By the central limit theorem, it (is as $n \to \infty$)

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

$$\geq 1 - \delta$$

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

High-confidence off-policy policy evaluation



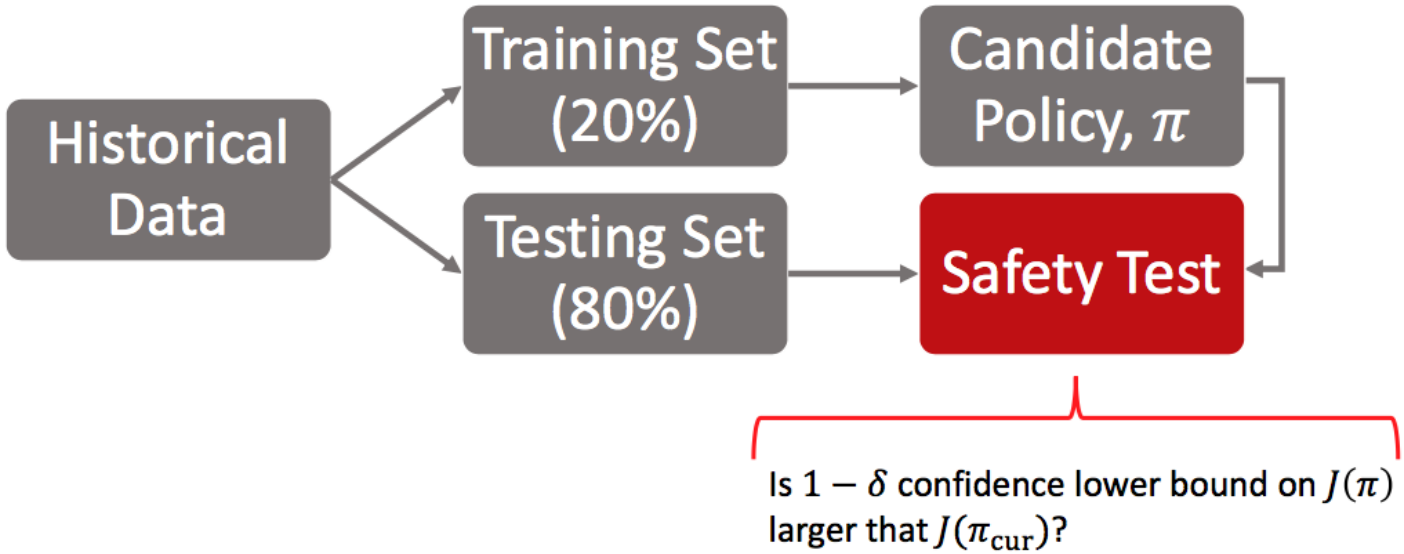
P. S. Thomas. Safe reinforcement learning (PhD Thesis, 2015)

Create a safe batch reinforcement learning algorithm

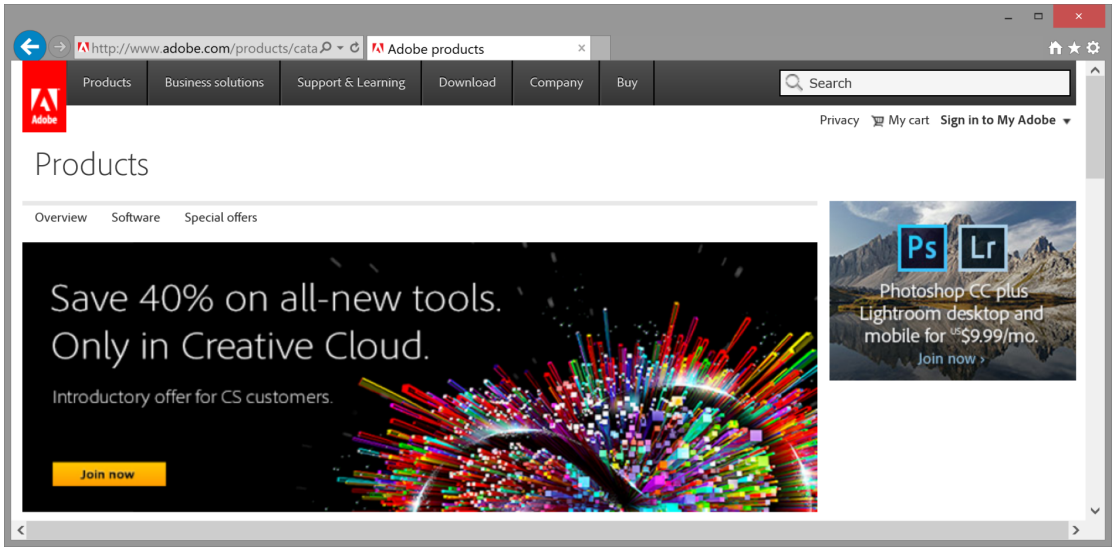
- Off-policy policy evaluation (OPE)
  - For any evaluation policy, π_e, Convert historical data, D, into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e}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

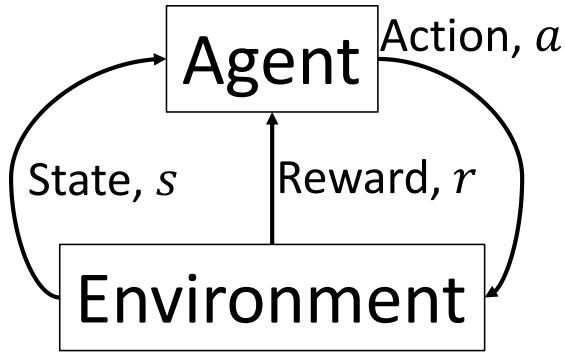
Safe policy improvement

Thomas et. al, ICML 2015

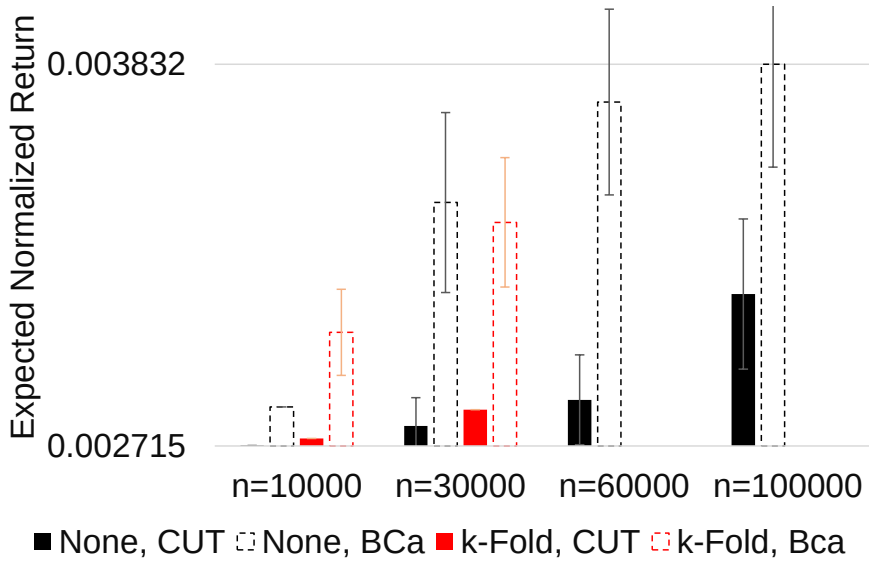


Empirical Results: Digital Marketing

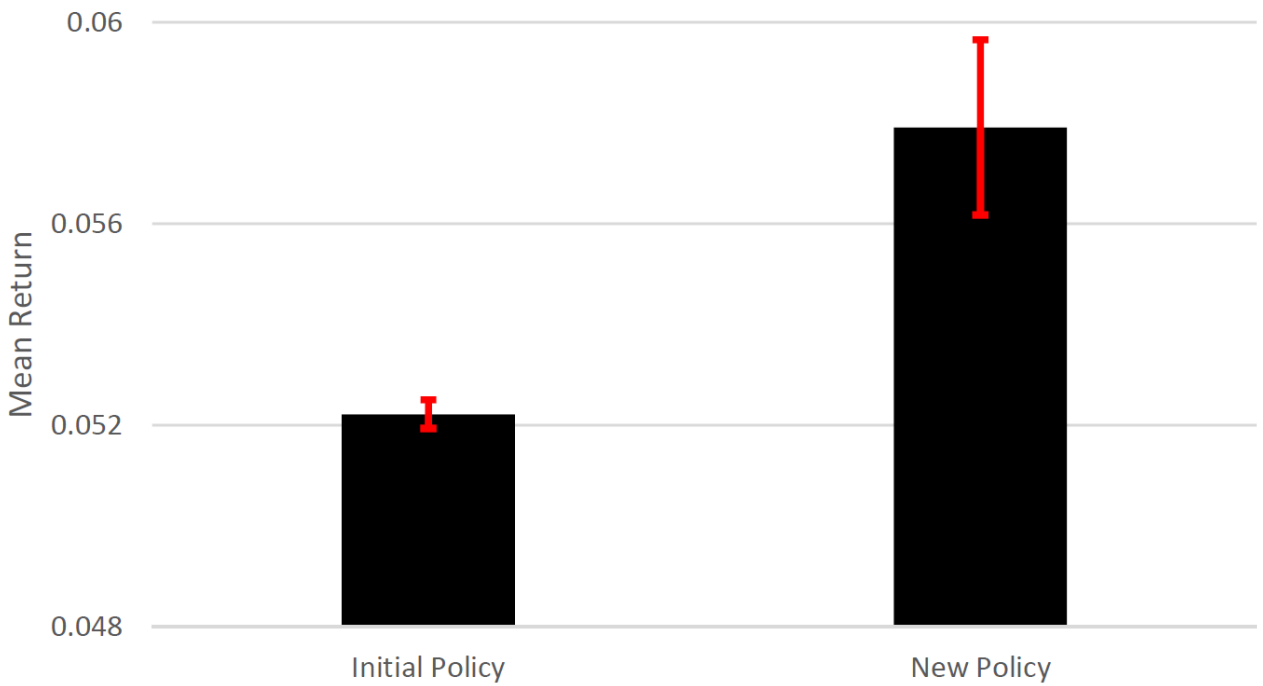




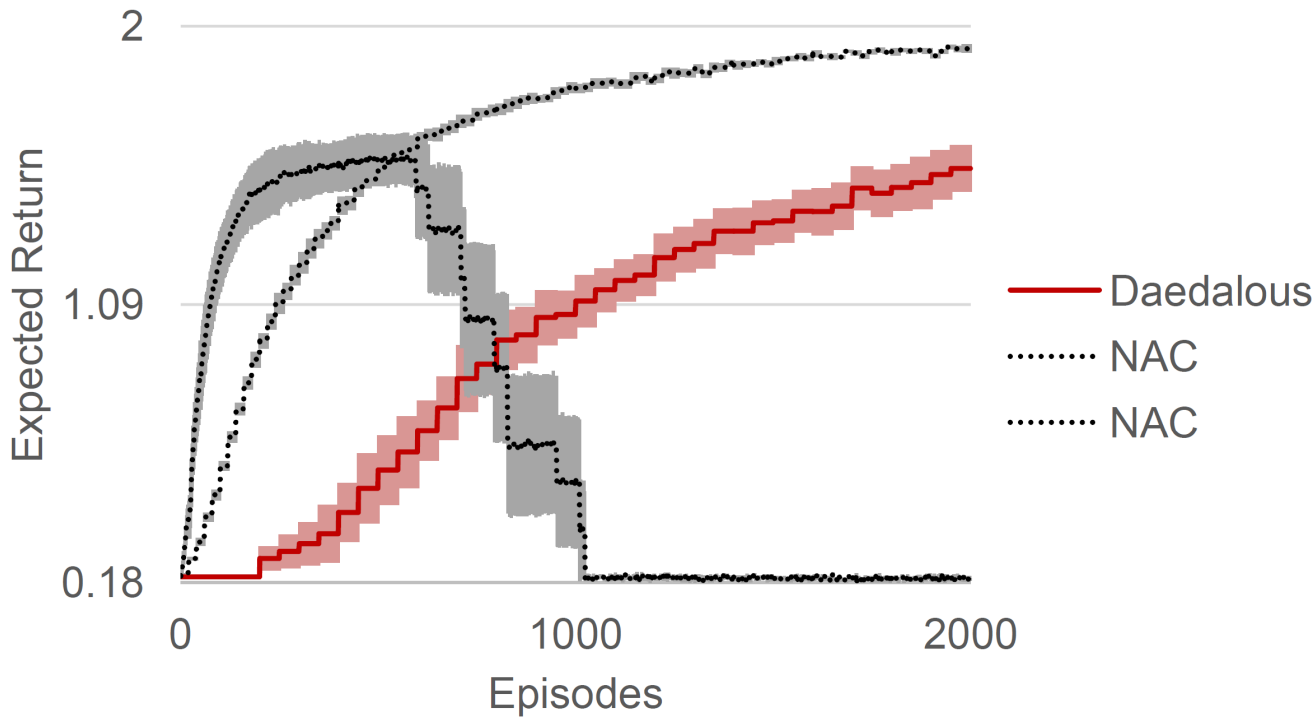
Empirical Results: Digital Marketing



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
- Many others also doing great work in this space, including the groups of Yisong Yue, Susan Murphy, Finale Doshi-Velez, Marco Pavone, Pieter Abbeel, Shie Mannor, Sergey Levine and Claire Tomlin, 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
- Our Science 2019 paper show how to do safe policy improvement for a high fidelity diabetes simulator and discusses the need for ensuring good behavior

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)
- Define why we might want safe reinforcement learning

Class Structure

- Last time: Fast Reinforcement Learning
- **This time: Batch RL**
- Next time: Guest Lecture